Artificial Intelligence in Medical Imaging

Opportunities, Applications and Risks

Erik R. Ranschaert
Sergey Morozov
Paul R. Algra
Editors

Springer

EuSoMII

Artificial Intelligence in Medical Imaging

Erik R. Ranschaert • Sergey Morozov •
Paul R. Algra
Editors

Artificial Intelligence in Medical Imaging

Opportunities, Applications and Risks

Springer

Editors
Erik R. Ranschaert
ETZ Hospital
Tilburg, The Netherlands

Sergey Morozov
Radiology Research and Practical Centre
Moscow, Russia

Paul R. Algra
Department of Radiology
Northwest Hospital Group
Alkmaar, The Netherlands

ISBN 978-3-319-94877-5 ISBN 978-3-319-94878-2 (eBook)
https://doi.org/10.1007/978-3-319-94878-2

Library of Congress Control Number: 2018966513

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

I've Seen the Future . . .

Scientists are people who know more and more about less and less,
until they know everything about nothing.
—(Konrad Lorenz? Web sources vary, so I gave up looking).

More than 50 years ago, in the turbulent spring of the revolutionary year 1968, film director Stanley Kubrick released his radically innovative science-fiction epos *2001: A Space Odyssey*, based on Arthur C. Clarke's novels. Together with a few classmates from school, I saw *2001: A Space Odyssey* at the wide-screen Rubens cinema theater in Antwerp, Belgium, in glorious 70 mm high-resolution projection.

In addition to the movie being a visually breathtaking and dazzling cinematic experience, it was also my very first introduction to the concept of artificial intelligence, and I think that this may hold true for many, if not most, people of my generation. The movie features a spaceship, the Discovery One, controlled by a computer called HAL (*H*euristically programmed *AL*gorithmic computer), a kind of artificial intelligence (AI) system *avant la lettre*, which controls the systems of the spacecraft and interacts with the astronauts on board. Throughout the movie, the presence of HAL is mostly inferred with close-ups of a red camera lens, with a central yellow dot. HAL is smart and scary and rapidly becomes the quintessential film villain, intelligent and treacherous, interacting conversationally with the crewmen in a deceptively soft, calm, and sometimes threatening voice. In this 1968 movie, the computer HAL was able to speak, recognize voices and faces, process natural language, lip-read, interpret human emotions, understand art, have discussions with the astronauts, and even play chess with the humans on board the spaceship.

When I first saw *2001: A Space Odyssey*, I was happily unaware that the term "artificial intelligence" had already been coined during the summer of 1956, when a group of researchers convened at a seminar in Dartmouth College, USA. The adjective "artificial" was meant to designate the cognitive process of "thinking machines," in contradistinction to the reasoning processes of humans. At that time, it was believed that human reasoning was "real," whereas machine thinking was "artificial." The 1960s were a period of optimism and confidence in the increasing computational speed of man-made machines, and many scientists in the field of AI were confident that computers would be capable of doing any work a man can do. Progress was, however, not meant to be linear. The heady decade of the 1960s was followed by what is now sometimes referred to as the "AI winter," a period of disenchantment, stalled progress, and rapidly dwindling research funding. But, like a cat with nine lives, AI came back with a vengeance in the 1980s, 1990s, and especially in the twenty-first century. In 1997, for the first time, a computer chess-playing system called Deep Blue successfully defeated Garry Kasparov, the reigning world chess champion. In 2011, Watson, a question-answering system developed by IBM, beat human champions of the television quiz game *Jeopardy!* in an exhibition match. And in 2016, a computer Go-playing system called AlphaGo became the first nonhuman system to triumph over Lee Sedol, a 9-dan professional master at the game of Go. This proves that machines can be instructed to think like humans, and even exceed their creators, especially since Go is an extremely complex game, more so than chess. But, of course, chess and Go are merely board games, and they are very, very different from real-life situations.

Today, more than 60 years after the "birth" of AI, we have indeed come a long way. The field of AI has continued to grow and to evolve in many different directions. Significant breakthroughs in artificial intelligence have occurred as the result of ongoing advances in data collection and aggregation, processing power, deep learning algorithms, and convolutional neural networks. Some of the most promising applications of AI have been in image processing and image analysis, which brings us to radiology. In just a few short years, AI applications in radiology have "exploded" and AI has become "big business." This is largely due to progress in artificial neural networks, the availability of cloud computing infrastructure, and the increased interest of medical professionals to pursue research in this field. It is not so long ago that image-recognition algorithms could only be used to tackle simple tasks such as differentiating cats from dogs. However, when the potential of machine learning systems is fully exploited, much more complex problems can be tackled, and this has opened up new avenues for radiology. Identifying and characterizing lung nodules on CT scans, computer-aided diagnosis of breast cancer on mammographic films, and automatic calculation of bone age by computer software on plain X-ray films of the hand are among the first such applications. Advanced segmentation techniques have opened up new avenues. Today, in diseases such as multiple sclerosis, Alzheimer's dementia, and traumatic brain injuries, AI is transforming patient care through accurate volume measurements of lesions and brain structures. Deep learning algorithms have been successfully implemented to diagnose different types of brain tumors, on the basis of multiparametric MRI data sets; in one such example, an AI system had an accuracy of 87% in predicting brain tumor neuropathology, outperforming human (neuro-) radiologists who scored only 66% accuracy. We are now pretty confident that AI software can be used to diagnose common neurological diseases with an accuracy rate of close to 90%, comparable to that of an experienced senior doctor. AI systems are proving to be faster, more reliable, and more accurate than human radiologists . . . and, obviously, they are available 24/7, they are never tired or sick, and they continue to "learn" as they analyze more cases.

So, while it would appear that AI and radiology are a match made in heaven, there actually is a lot of hysteria and apprehension around AI and its impact on the future of radiology. It is sad to see that the advent of AI systems has created so much anxiety and self-doubt among radiologists. The AI genie is out of the bottle; we cannot turn the clock back, but we do have the power to determine the future: tomorrow belongs to those who prepare for it today. It seems likely that radiologists who use AI will replace those who don't, since there are many signs indicating that AI will have a profound impact on the world of radiology. As George Bernard Shaw said: "*we are made wise not by the recollection of our past, but by the responsibility for our future*." Nevertheless, I am not afraid, since I am convinced that radiologists will embrace AI to help us manage "routine" tasks quickly and efficiently, thus giving us more time to focus on things that really matter. For that is exactly what AI software solutions will do: take over repetitive and simple tasks. I do not share the bleak and defeatist vision of the future for radiologists. History teaches us that the arrival of new technology tends to increase, rather than reduce, the need for human personnel. More than a hundred years ago, when automobiles started to replace horses as the preferred means of transportation of goods and people, some professions such as horseshoe smiths and saddlemakers became virtually extinct, but the car industry more than made up for this loss by creating new means of employment and career opportunities.

Personally, I believe that the integration of AI into existing medical workflow is a very promising trend and we should embrace this exciting new prospect, rather than fight it or run away from it. In my opinion, AI will help a radiologist like a GPS guides the driver of a car. AI will offer proposals to the radiologists and help the doctor to make a better and more accurate diagnosis. But it will be the doctor who ultimately decides, as there are a number of factors that a machine which interprets imaging data sets cannot take into consideration, such as a patient's general state of health and family situation.

AI systems offer our profession a unique opportunity to make a new beginning, to re-invent what we do, to boost productivity and accuracy. I am convinced that AI can take over time-consuming routine tasks, freeing up time and resources to focus our attention on individual patients, and thereby moving from volume-based radiology toward value-based radiology. So, regarding the implementation of AI software into radiological practice, my closing message to all radiologists is: take charge of your own future, and embrace it with confidence, courage, and determination.

Prof. Dr. Paul M. Parizel, MD, PhD

Past President, European Society of Radiology (ESR)
Institutional Representative, European Board of Radiology (EBR)
Honorary President, African Society of Radiology (ASR)
Member, Royal Academy of Medicine of Belgium (RAMB)
Honorary Fellow, Royal Australian and New Zealand College of Radiologists (RANZCR)

Honorary and Founding Member, Russian Society of Neuroradiology (RSNR)
Honorary Member, Serbian Society of Neuroradiology (SSNR)
Member of Honour, Romanian Society of Radiology and Medical Imaging (SRIM)
Miembro de Honor, Sociedad Española de Radiología Médica (SERAM)
Honorary Member, European Society of Neuroradiology (ESNR)
Membre d'Honneur, Société Française de Radiologie (SFR)
Honorary Member, Israel Radiological Association (ISRA)
Membre d'Honneur, Société Algérienne de Radiologie et d'Imagerie Médicale (SARIM)
Honorary Member, American Society of Neuroradiology (ASNR)
Schinz Medal, Swiss Society of Radiology (SSR)

Department of Radiology
Antwerp University Hospital (UZA)
University of Antwerp (UA)
Edegem, Belgium

Preface

An increasing number of unanswered questions made us unsettled about the evolution of radiology and its current ability to keep up with the pace of healthcare transformation. The answers to these questions hopefully can be found in this book, which we have started in 2017 from the inside of the European Society of Medical Imaging Informatics (EuSoMII). The first EuSoMII Academy on AI in Rotterdam in November 2017 (actually the 35th annual meeting of EuroPACS-EuSOMII) attracted a plethora of bright minds and speakers who inspired us to publish a book about artificial intelligence (AI) in medical imaging. The preparation of this book originates from a brainstorm launched by Erik Ranschaert and Paul Algra immediately after the annual meeting. The idea was to provide a bright picture of existing ideas, concepts, and practical examples of AI applications in radiology and to connect professional worlds and communities of health care and data science. The work on the book progressed very well, and the harmonious whole of the authors' insightful and practical chapters has really pleased and inspired the editors.

The main activities of the EuSoMII society are centered around medical imaging informatics (MII), a.k.a. radiology informatics or imaging informatics, which is a subspecialty of biomedical informatics. Its purpose is to improve the efficiency, accuracy, usability, and reliability of medical imaging services within the healthcare enterprise. Imaging informatics covers processes for the acquisition, manipulation, analysis, storage, distribution, retrieval, and use of imaging data. Its area of interest is therefore very wide and includes topics ranging from radiology information systems (RIS), picture archiving and communication system (PACS), shared workflow, advanced visualization, and computer-aided diagnosis (CAD) to biobanks, computer vision, augmented reality/virtual reality (AR/VR), and 3D modeling. MII exists at the intersection of several broad fields: biological science, clinical services, information science, medical physics, biomedical engineering, cognitive science, and computer science. The IT solutions used in other industries are also relevant for application in the medical field, with the main intention of achieving a higher level of efficiency and safety in health care.

The role of a professional society is indisputable, as it is a driving force to bring the ideas forward and to share the early results of research for the common good. Another important role of such society is to become a guide in changes in a specialty that at the same time preserves the core of the specialism and follows ethical principles. All this precisely describes

EuSoMII's role in supporting the acceptance of advanced imaging informatics and the optimal integration in the radiological community.

The European Society of Medical Imaging Informatics (EuSoMII) is a professional healthcare organization that provides its members and the radiological community with up-to-date information on the latest innovations and achievements in medical IT by supporting education, research, development of standards, and networking related to a top tier of IT solutions in radiology, pathology, cardiology, neurology, and other imaging-based subspecialties. The guiding principles of EuSoMII are multidisciplinary and international collaboration, joining forces and efforts to make radiology stronger and increase our specialty's value. A multidisciplinary group consisting of radiologists, physicists, radiology technicians, IT experts, and other biomedical informatics professionals represents its target audience.

EuSoMII's vision is to reach an optimal integration of information and communication technologies (ICT) with medical imaging professions for increasing the quality and safety of diagnostics and therapeutics. EuSoMII aims to become a leading think tank for new developments in ICT related to medical imaging, enabling distribution of best practices within the professional community. In its current role and format EuSoMII is a driving force behind Imaging Informatics Subcommittee of the European Society of Radiology (ESR), a leading provider of training and teaching on imaging informatics in Europe, a partner of the Society for Imaging Informatics in Medicine (SIIM), Computer Applications in Radiology and Surgery (CARS), Healthcare Information and Management Systems Society (HIMSS), the Medical Image Computing and Computer Assisted Intervention Society (MICCAI), European Federation of Organizations for Medical Physics (EFOMP), and many medical subspecialty societies.

The structure of the AIMI book multi-directionally develops all aspects of artificial intelligence applications in radiology and allied specialties. It starts from the role of medical imaging computing, informatics, and machine learning in health care, proceeds into the principles of deep learning (DL) and neural networks (NN) in imaging, provides guidance on how to develop AI applications, and presents a medical imaging data readiness (MIDaR) scale for machine learning tasks in radiology. Further on the book emphasizes several significant medical imaging AI domains for developers, such as the value of structured reporting, validation of AI applications, enterprise imaging, imaging biomarkers, and image biobanks. Practical use cases of AI in radiology are outlined in detail for the areas of chest pathology, cardiovascular diseases, breast cancer, neurological diseases, and clinical trials support and for applications beyond imaging. Economic and legal aspects of AI are elaborated by presenting a regulatory infrastructure, a perspective on the market and economics, and the importance of an AI ecosystem for radiology. Finally, the book addresses advantages and risks of AI for radiologists, balancing them by presenting a "rethinking" of radiology as a medical specialty. The AIMI book is a journey along a highway of healthcare innovations where radiologists and data scientists travel in one direction guided by the principles of medical quality and safety for the patients.

Our outstanding team of authors and editors is honored and happy to welcome you to join the shaping of the future of diagnostic imaging. We deeply appreciate and value the contributions of the authors and coauthors. We would like to thank the whole AIMI book team for their ingenuity, creativity, originality, professionalism, openness for discussion, and constructive critique. Thank you, dear coauthors and coeditors, for sharing your knowledge, experience, vision, and values.

Tilburg, The Netherlands Erik R. Ranschaert
Moscow, Russia Sergey Morozov

Contents

Part I Introduction

Part II Technology: Getting Started

Part III Technology: Developing A.I. Applications

Part IV Big Data in Medicine

Part V Practical Use Cases of A.I. in Radiology

Part VI Quality, Regulatory and Ethical Issues

Part I

Introduction

1 Introduction: Game Changers in Radiology

Sergey Morozov, Erik Ranschaert, and Paul Algra

1.1 Era of Changes

Medicine in general and radiology as a specialty are experiencing significant changes, associated with extensive introduction of informatics, machine learning, biobanks, and personalized medicine. It is important that radiologists develop an understanding not only of computer technologies but also of a wider field of information technologies and informatics in order to keep up with the ongoing digital transformation of our society, including healthcare. However, the current situation seems to be rather characterized by a general radiologists' reluctance to changes. The media is full with publications and cautions on an approaching disappearance of radiology as a specialty, its merge or acquisition by other specialties, and even a decreasing number of radiology residency applications. Indeed, the power of media and irrational emotions is strong. As far back as in the 1960s, Amos Tversky, cofounder of behavioral economics and coauthor of Nobel laureate Daniel Kahneman, said, in response to a question on a contribution of his studies of human bias and irrationality of choice to the development of artificial intelligence (AI): "My colleagues, they study artificial intelligence; me, I study natural stupidity" [1]. His words are no less significant today when the healthcare market is already observing collapses of businesses aiming to "substitute a medical doctor by an algorithm."

The media pressure creates an obvious hype around AI, which has not only negative but also a positive effect on the healthcare. Computer scientists and entrepreneurs start asking radiologists uncomfortable questions. What is a value of your specialty? What actually do you do on a routine basis? Which tasks do you perform as a radiologist? How do you analyze and interpret images? How do you report them? Which standards do you follow? They also start measuring a time for performing various tasks in a diagnostic workflow. And a bottleneck of radiology productivity becomes unearthed, being often attributable to a shortage of qualified radiologists or ineffective use of a radiologist's time.

But let us consider a difference between radiologists and cardiologists: the latter are much less concerned about a threat of AI. They really use it intensively to improve their efficiency and value. Why are radiologists so hesitating? What are we

S. Morozov (✉)
Research and Practical Center of Medical Radiology, Moscow, Russia
e-mail: smorozov@post.harvard.edu

E. Ranschaert
ETZ Hospital, Tilburg, The Netherlands

P. Algra
Department of Radiology, Northwest Hospital Group, Alkmaar, The Netherlands

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_1

waiting for? What are the incentives to go for AI, and why do we think differently about it? Why do we see it as a threat instead of a very valuable and useful tool to improve our services both from an economic/financial/management point of view and a patient-oriented perspective?

1.2 Perspectives

In the recent years, deep learning (DL) algorithms based on neural networks have become widely available and have significantly contributed to the (still ongoing) hype on the role of DL in radiology. However, an AI on itself is a very broad concept which includes lots of conventional technologies and algorithms. Even working with formula in Excel tables might be classified as such. One essential principle of programming and development remains however the same for all technologies under the umbrella of AI: "garbage in is garbage out." Therefore it's essential for any new software tool to be developed as a product with a clearly defined target or clinical use case.

On the other hand, the preparedness of healthcare providers to change their routine methods also depends on their openness to solutions coming from other industries, such as retail, aviation services, and hospitality industry. Technologies from computer games and investment business automation are and can be applied in healthcare. The recent advances that have been made in the development of DL algorithms for medical imaging, for example, are partly due to the availability of highly advanced GPUs (graphics processing units), which were developed in the computer games industry. We are confronted with a progressive but accelerating transition of healthcare from art and craft to a more industrial concept of working, based upon scientific and technological progress. This doesn't necessarily imply the replacement of human doctors by machines but a reinforcement of healthcare by scaling its processes and introducing quality control systems. Solutions from other industries mentioned can bring new ideas and business processes not only for increasing the value of healthcare services and patient satisfaction but also for minimizing the costs and making healthcare more accessible to a wider range of patients or even on a larger scale and in areas or in populations that are deprived of medical care.

1.3 Opportunities for the Future

The key healthcare development opportunities can be assigned to three factors: a further integration of information technologies (IT) and systems in healthcare, the connection of separate healthcare providers into networks able to share digital information, and the standardization of medical procedures and their digital formats. All of these will incentivize and facilitate the further development and deployment of AI-based tools for optimization of workflow and value in healthcare. In addition applications for telemedicine and teleradiology, by which low-cost primary care and diagnostics can be provided in remote areas, will be a stimulus for developing such tools. In the coming years, we will probably also see the further deployment of analytical and predictive medical tools that are provided in a B2C model to the average consumer. The generation Y became inured to searching the Internet, taking tests via social media, and receiving services on the spot. Therefore medical care is not excluded from this trend. The further development of such new technologies is however limited by several factors such as the lack of structured data, the existing legislations and regulations on privacy and security, the skepticism of many medical specialists, and the reluctance of medical community to change, as well as the patients' resistance to "extreme standardization and IT penetration instead of a human talk to a doctor."

1.4 Conclusion

Artificial intelligence is capable of revolutionizing healthcare industry by the expedited development of personalized and automated diagnostics, new diagnostic data-based methods, imaging-guided robot-assisted surgery, tele-monitoring of chronic conditions, support of correct medical decisions, as well as systematic monitoring of potential diagnostic errors. Expertise, wisdom, human attitude, care, empathy, mutual understanding, and support lie at the very base of the medical profession and cannot be automated. Professional medical societies should lead this transformation while preserving and ensuring the quality and safety of new diagnostic algorithms. Let artificial intelligence help us.

Reference

1. Lewis M. The undoing project: a friendship that changed our minds. New York: W. W. Norton & Company; 2017. 368p.

Part II

Technology: Getting Started

The Role of Medical Image Computing and Machine Learning in Healthcare

2

Frederik Maes, David Robben, Dirk Vandermeulen, and Paul Suetens

2.1 Introduction

Due to continuing technological advances in medical image acquisition, novel imaging modalities are being introduced in medical practices, such as multi-slice (volumetric) and multi-energy CT, multi-parametric and multi-frame (dynamic) MRI, multi-dimensional (3D+time) US, multi-planar interventional imaging, or multi-modal (hybrid) PET/CT and PET/MRI imaging technologies [1]. The analysis of the large amounts of imaging data created by these modalities has become a tremendous challenge and a real bottleneck for diagnosis, therapy planning and follow-up, and biomedical research. At the same time, the general adoption of digital picture archiving and communication systems (PACS) in radiology, and their integration within the overall hospital information system, makes that large databases of medical images and associated relevant medical information of patients (including demographics, clinical findings, blood tests, pathology, genomics, proteomics) are being built up. It is to be expected that such databases will become more and more accessible for research purposes, provided that technical challenges and privacy issues can be properly dealt with. The availability of well-documented medical imaging "big data" offers new opportunities for groupwise analyses within specific subject groups, such as characterization of normal and abnormal variation between subjects and detection of individual patient anomalies (computer-aided diagnosis), discovery of early markers of disease onset and progression (imaging biomarkers), optimal therapy selection and prediction of therapy outcome (radiomics in radiotherapy), and correlation of genotype and phenotype related findings (imaging genetics). In order to optimally exploit all available imaging data and to support the effective use of "big data" involving medical images in the context of personalized medicine, reliable computer-aided image analysis becomes indispensable to extract and quantify the relevant information from the imaging data, to fuse complementary information and to support the interpretation thereof.

2.2 Medical Image Analysis

Medical image analysis involves measurements in medical images, i.e., the extraction of relevant quantitative information from the images. Manual measurements by human experts in large 3D medical imaging datasets (in particular by radiologists in clinical practice) are not only tedious and time-consuming and thus impractical

F. Maes (✉) · D. Robben · D. Vandermeulen · P. Suetens
KU Leuven, Department of ESAT/PSI, Leuven, Belgium
e-mail: frederik.maes@kuleuven.be

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_2

in clinical routine, but also subject to significant intra- and inter-observer variability, which undermines the significance of the clinical findings derived from them. There is, therefore, great need for more efficient, reliable, and well-validated automated or semi-automated methods for medical image analysis to enable computer-aided image interpretation in routine clinical practice in a large number of applications. Which information needs to be quantified from the images is of course highly application specific. While many applications in computer vision involve the detection or recognition of an object in an image, whereby the precise geometry of the objects is often not relevant (e.g., image classification, object recognition) or may be known a priori (e.g., machine vision), medical image analysis often concerns the quantification of specific geometric features of the objects of interest (e.g., their position, extent, size, volume, shape, symmetry, etc.), the assessment of anatomical changes over time (e.g., organ motion, tissue deformation, growth, lesion evolution, atrophy, aging, etc.), or the detection and characterization of morphological variation between subjects (e.g., normal versus abnormal development, genotype related variability, pathology, etc.). The analysis of 3D shape and shape variability of anatomical objects in images is thus a fundamental problem in medical image analysis. Apart from morphometry, quantification of local or regional contrast or contrast differences is of interest in many applications, in particular in functional imaging, such as fMRI, PET, or MR diffusion and perfusion imaging.

Within the wide variety of medical imaging applications, most image analysis problems involve a combination of the following basic tasks [2].

2.2.1 Image Segmentation

Image segmentation involves the detection of the objects of interest in the image and defining their boundaries, i.e., discriminating between the image voxels that belong to a particular object and those that do not belong to the object. Image segmentation is a prerequisite for quantification of the geometric properties of the object, in particular its volume or shape. Image segmentation can be performed in different ways: boundary-wise by delineating the contour or surface of the object in one (2D) or multiple (3D) image slices; region-wise by grouping voxels that are likely to belong to the same object into one or multiple regions; or voxel-wise by assigning each voxel in the image as belonging to a particular object, tissue class, or background. Class labels assigned to a voxel can be probabilistic, resulting in a soft or fuzzy segmentation of the image. Accurate 3D segmentation of complex shaped objects in medical images is usually complicated by the limited resolution of the images (leading to loss of detail and contrast due to partial volume artifacts) and by the fact that the resolution is often not isotropic (mostly multi-slice 2D instead of truly 3D acquisitions). Hence, interpolation is usually needed to fill in the missing information in the data. In clinical practice, precise 3D measurements (e.g., volumetry) may be too tedious and time-consuming, such that often a simplified, approximate 2D or 1D analysis is used instead (e.g., for estimation of lesion size).

2.2.2 Image Registration

Image registration involves determining the spatial relationship between different images, i.e., establishing spatial correspondences between images or image matching, in particular based on the image content itself [3]. Different images acquired at different time points (e.g., before and after treatment), or with different modalities (e.g., CT, MRI, PET brain images), or even from different subjects (e.g., diseased versus healthy) often contain complementary information that has to be fused and analyzed jointly, preferably at the voxel level to make use of the full resolution of the images. Image registration is needed to compensate for a priori unknown differences in patient positioning in the scanner, for organ or tissue deformations between different time points, or for anatomical variation between subjects. After proper registration, the images can be resampled onto a common geometric space and fused, i.e., spatially corresponding voxels can be precisely overlaid, which drastically facilitates the joint analysis of the images. In some cases, when deformations are ignorable, the registration solution can be represented as an affine transformation matrix with a small number of parameters, but in general a more complex transformation in the form of a locally flexible deformation field is needed to accommodate for non-rigid distortions between the images.

2.2.3 Image Visualization

The information that is extracted from the images ideally needs to be presented in the most optimal way to support diagnosis and therapy planning, i.e., such that the correct interpretation by the user of all relevant image data is maximally facilitated for a specific application. For 3D medical images, 2D multi-planar visualization is not well suited to assess structural relationships within and between objects in 3D, for which true 3D visualization approaches are to be preferred. To this end, either surface rendering or volume rendering can be applied. Surface rendering assumes that a 3D segmentation of the objects of interest is available and renders these within a 3D scene under specified lighting conditions (e.g., ambient light, point light sources) by assigning material properties to each surface or surface element that specify its specular and diffuse light reflection, transmission, scattering, etc. Volume rendering instead renders the image voxels directly by specifying suitable transfer functions that assign each voxel a color and opacity depending on their intensity. While in principle volume rendering does not require a prior segmentation of the objects of interest, in practice a prior segmentation of the image is often applied such that the transfer functions can be made spatially dependent and object specific, which allows to discriminate between voxels with similar intensity belonging to different objects. In clinical applications such as image-based surgery planning or image-guided intra-operative navigation, additional tools need to be provided to manipulate the objects in the 3D scene (e.g., cutting), to add virtual objects to the scene (e.g., implants), or to fuse the virtual reality scene with real-world images (e.g., intra-operative images). While such augmented reality techniques can improve the integrated presentation of all available information during an intervention (e.g., using a head-mounted display), their introduction in clinical practice is far from trivial.

Image segmentation, registration, and visualization should not be seen as separate subproblems in medical image analysis that can be addressed independently, each using a specific set of strategies. On the contrary, they are usually intertwined and an optimal solution for a particular image analysis problem can only be achieved by considering segmentation, registration, and visualization jointly. For instance, image registration can be used as a computational strategy for image segmentation by matching the image to be segmented to a similar image (e.g., from a different patient, or an atlas template) that has been previously segmented (i.e., atlas-based segmentation). Vice versa, image registration can benefit from the fact that a prior segmentation of similar structures in each image is already available, as these provide strong clues to guide the registration process. Joint visualization of multiple different images, acquired, for instance, pre-operatively and intra-operatively, requires that registration issues between all images residing in different coordinate systems have been resolved. This in turn is facilitated when suitable visualization and manipulation tools are available to verify and adjust the registration interactively by visual feedback (i.e., visual matching). Moreover, in applications involving image-guided treatment, the pre-operative treatment plan needs to be transferred onto the patient during the intervention and the intra-operative position of the instruments needs to be tracked in the images. This registration problem typically requires additional hardware to be installed in the treatment room (e.g., an optical tracking system).

2.3 Challenges

Medical image analysis is complicated by different factors, in particular the complexity of the data, the complexity of the objects of interest, and the complex validation.

2.3.1 Complexity of the Data

Medical images are typically 3D tomographic images. The 3D nature of the images provides additional information, but also an additional dimension of complexity. Instead of processing the data in 2D slice by slice, 3D processing is usually more effective as it allows to take spatial relationships in all three dimensions into account, provided that the resolution of the data in-plane and out-plane is comparable. Medical images are based on different physical principles and the quantification of the images is complicated by the ambiguity that is induced by the intrinsic limitations of the image acquisition process, in particular limited resolution, lack of contrast, noise, and the presence of artifacts. Moreover, many applications involve the analysis of complementary information provided by multiple images, for instance, to correlate anatomical and functional information, to assess changes over time or differences between subjects. It is clear that the variable, *multi-X* nature of the images to be analyzed (multi-dimensional, multi-modal, multi-temporal, multi-parametric, multi-subject, multi-center) poses specific challenges.

2.3.2 Complexity of the Objects of Interest

The objects of interest in medical images are typically anatomical structures (sometimes also other structures, e.g., implants), either normal or pathological (e.g., lesions), that can be rigid (e.g., bony structures) or flexible to some extent (e.g., soft tissue organs). Anatomical structures may exhibit complex shape, such as the cortical surface of the brain, the cerebral and coronary vessels, or the bronchial tree in the lung. Such complex shapes cannot easily be described by a mathematical model. Moreover, anatomical structures can show large intra-subject shape variability, due to internal soft tissue deformations (e.g., breathing-related motion, bowel activity), as well as inter-subject variability, due to normal biological variation and pathological changes. In general, the appearance of similar structures in different images (of the same subject at different time points or from different subjects) can show significant variability, both in shape and in intensity. Computational strategies for medical image analysis need to take this variability into account and be sufficiently robust to perform well under a variety of conditions.

2.3.3 Complexity of the Validation

Medical image analysis involves the quantification of internal structures of interest in real-world clinical images that are not readily accessible from the outside. Hence, assessment of absolute accuracy is often impossible in most applications, due to lack of ground truth. As an alternative, a known hardware phantom that mimics the relevant objects of interest could be imaged, but the realism of such a phantom compared to the actual in vivo situation is often questionable. Moreover, a hardware phantom usually constitutes a fairly rigid design that is not well apt to be adapted to different variable anatomical instances. Instead, the use of a software phantom in combination with a computational tool that generates simulated images based on a model of the imaging process provides more flexibility, with respect to both the imaged scene and the image acquisition setup itself. But again, such simulated images often fail to capture the full complexity of real data. In clinical practice, ground truth is typically provided by manual analysis by a clinical expert, for instance, manual delineation in case of image segmentation or manual annotation of corresponding anatomical landmarks in case of image registration. As already mentioned, such manual analysis is subject to intra- and inter-observer variability, which should be taken into account when validating (semi-)automated methods. Apart from accuracy, precision, consistency, and robustness of the method are to be considered as well when evaluating its clinical usability.

2.4 Medical Image Computing

Medical image computing, which is a branch of scientific computing at the intersection of medical imaging, computer vision, and machine learning, aims at developing computational strategies for medical image analysis that can cope with the complexity of medical imaging data to enable (semi-)automated analysis with sufficient accuracy and robustness. Such strategies rely on mathematical models that incorporate prior knowledge about the typical appearance of the objects of interest in the images, including photometric properties (e.g., intensity, contrast, texture), geometric properties (e.g., position, shape, motion), and context (e.g., relations to other objects) [4]. Model-based image analysis involves the construction of an appropriate parameterized representation for the model, the derivation of an objective function for assessing the goodness of fit of the model to the data, and the selection of a suitable optimization strategy for finding the optimal parameters of the model instance that best fits the image data. The models need to be sufficiently flexible to account for image appearance variations, due to, e.g., variability in the image acquisition itself, normal biological shape variability, motion and deformation, and pathology. The flexibility of the model is determined by the specific representation that is chosen for the model, its parameterization and number of degrees of freedom, and by the constraints imposed on its parameters.

Simplistic models based on generic, heuristic assumptions about the appearance of the objects in the images, for instance, about their intensity homogeneity and (dis)similarity or their boundary continuity and smoothness, are in general not suited for medical image analysis applications, as they are not powerful enough to capture the full complexity of the problem (apart from few exceptions, such as segmentation of bony structures in CT). Instead, more sophisticated approaches are needed that incorporate application-specific information about the images to be analyzed. A natural and powerful strategy is to construct suitable models from the data itself by analysis of a representative set of previously analyzed images. Such statistical models could in principle ensure that the degrees of freedom of the model are optimally tuned to the relevant variation within the data of each specific application, provided that the training set of previously analyzed images from which the model is constructed is large enough and representative for the population of subjects being considered. Instead of making use of a generic parameterization of the model that is applicable to many different applications, the construction of application-specific statistical models allows to decrease the number of relevant parameters of the model by exploiting correlations in the data, for instance, by adopting a multi-variate Gaussian model for the underlying distribution of the data or by using dimensionality reduction techniques such as principal component analysis. Instead of postulating a specific analytical form for the statistical model, more general supervised data-driven approaches can also be used to infer the relationship between a vector of specific features extracted from the data and the desired quantification outcome. To this end, various machine learning strategies for feature-based classification, such as support vector machines or random decision forests, can be used [5].

Recent advances in supervised learning of models from training data, especially deep learning based on convolutional neural networks, have shown great promise for many problems in computer vision, including image classification, object recognition, and segmentation. Deep learning also shows great promise for medical imaging applications [6]. The analysis problem is formulated as a classification task based on a large set of non-specified local features that are assumed to be homogeneous within and between all images separately and that are to be learned by the neural network itself. Neural networks define highly complex function classes and large amounts of data are typically necessary for them to converge to a stable solution with good generalization ability. This requirement is usually not met in medical image analysis, where the availability of training data is limited, which poses additional challenges that are not addressed in the majority of the machine learning literature. Moreover, the analysis of shape and shape variability, which is a fundamental problem in medical image analysis, typically includes dispersed non-local patterns derived from dense spatial correspondences between heterogeneous images analyzed jointly, for which it is not evident how this problem could be formulated as a classification problem using current neural network architectures. In addition, model-based analysis aims at avoiding heuristics and implicit assumptions about the data as much as possible by making such assumptions explicit using a parametric model. While deep learning seems to comply with this paradigm by avoiding heuristic feature construction and selection, plenty of heuristics are embedded in the actual implementation of the chosen neural network architecture, in the optimization strategy and in the sampling of the training data presented to the network, which complicates the interpretation of the model and the optimization of its performance.

2.5 Model-Based Image Analysis

Model-based image analysis makes use of a model of the image appearance of the objects of interest in the images. The model represents prior knowledge about the geometric, photometric, and contextual properties of the objects of interest in the images. The model should be sufficiently specific to deal with ambiguity in the imaging data and at the same time sufficiently flexible to accommodate for normal biological shape variability, pathological changes, or non-rigid deformations of the objects of interest. The model is fitted to the image data using a suitable measure for the goodness of fit. This can be generally formulated as follows. Let I be the image data and $M(\theta)$ the representation of the model with parameters θ. Fitting the model $M(\theta)$ to the image data I involves finding the model instance $M(\theta^*)$ with parameters θ^* for which the a posteriori probability $\text{Prob}(M(\theta)|I)$ is maximized:

$$\theta^* = \arg\max_\theta \text{Prob}(M(\theta)|I) \tag{2.1}$$

Using Bayes' rule, the a posteriori probability can be written as

$$\text{Prob}(M(\theta)|I) = \frac{\text{Prob}(I|M(\theta)) \cdot \text{Prob}(M(\theta))}{\text{Prob}(I)} \tag{2.2}$$

with $\text{Prob}(I|M(\theta))$ the likelihood of observing the data I given the model, $\text{Prob}(M(\theta))$ the prior probability of the model instance with parameters θ, and $\text{Prob}(I)$ the probability of observing the data I. Because the latter is independent of $M(\theta)$, it follows that the optimal model parameters should satisfy

$$\theta^* = \arg\max_\theta (\text{Prob}(I|M(\theta)) \cdot \text{Prob}(M(\theta))) \tag{2.3}$$

2.5.1 Energy Minimization

Instead of estimating and maximizing $\text{Prob}(M(\theta)|I)$ directly, the model fitting can as well be performed by estimating the prior probability $\text{Prob}(M(\theta))$ and the data likelihood $\text{Prob}(I|M(\theta))$ and maximizing their product. The maximum is preserved by taking the logarithm of the right-hand side (as the logarithm is monotonically increasing):

$$\theta^* = \arg\max_\theta \log(\text{Prob}(I|M(\theta)) \cdot \text{Prob}(M(\theta))) \tag{2.4}$$

$$= \arg\max_\theta \log(\text{Prob}(I|M(\theta))) + \log(\text{Prob}(M(\theta))) \tag{2.5}$$

By adopting a Gibbs distribution for both the prior and the data likelihood, this optimization problem can be formulated as energy minimization:

$$\text{Prob}(M(\theta)) = \frac{\exp(-E_{\text{int}}(\theta))}{Z_{\text{int}}} \tag{2.6}$$

$$\mathrm{Prob}(I|M(\theta)) = \frac{\exp(-E_{\mathrm{ext}}(\theta|I))}{Z_{\mathrm{ext}}(I)} \tag{2.7}$$

with Z_{int} and $Z_{\mathrm{ext}}(I)$ normalization constants (independent of θ) and E_{int} and E_{ext} the internal and external energy, respectively, such that

$$\theta^* = \arg\max_\theta \log\left(\frac{\exp(-E_{\mathrm{int}}(\theta))}{Z_{\mathrm{int}}}\right) + \log\left(\frac{\exp(-E_{\mathrm{ext}}(\theta|I))}{Z_{\mathrm{ext}}}\right) \tag{2.8}$$

$$= \arg\min_\theta \left(E_{\mathrm{int}}(\theta) + E_{\mathrm{ext}}(\theta|I)\right) \tag{2.9}$$

The internal energy E_{int} of the model instance θ is a measure for its a priori likelihood, while the external energy E_{ext} represents the goodness of fit of the model instance θ to the data. Note that E_{ext} depends on the image data I, while E_{int} is independent of I. The actual formulation of the energy terms in the objective function E depends on the choice of the representation for the model $M(\theta)$ and its parameterization and on the specific prior knowledge about the appearance of the objects of interest in the images that is to be captured by the model. In practice, the energy minimization formalism is very versatile and makes it easy to implement various different deterministic or statistical, purely mathematical or biomechanically inspired, heuristic or learned constraints (or penalties) on the model [7].

In practice, both energy terms are not absolute but relative and need to be weighted with respect to each other, which can be made explicit by introducing a weight γ in the energy function:

$$E(\theta|I,\gamma) = E_{\mathrm{ext}}(\theta|I) + \gamma E_{\mathrm{int}}(\theta) \tag{2.10}$$

$$\theta^* = \arg\min_\theta E(\theta|I,\gamma) \tag{2.11}$$

By tuning the value of γ, the optimal model instance can be driven towards favoring more data congruency (γ small) or towards more model fidelity (γ large). Additional parameters may be embedded in the definition of E_{int} and E_{ext} itself. Tuning of such hyperparameters for optimal performance for the given application is an intrinsic difficulty of any model-based image analysis approach.

The solution of this optimization problem requires a suitable optimization strategy. Typically, this can be done using variational calculus and iterative gradient descent on E w.r.t. the model parameters θ, starting from a suitable initialization θ_0 that is sufficiently close to the actual optimal value θ^*. In practice, such local search procedure yields a local optimum θ^+ that is not guaranteed to be globally optimal and that depends on the initialization and on parameters of the optimization procedure itself (e.g., step size, stop criterion). Alternatively, by discretizing the solution space θ, the problem can be formulated as a discrete optimization problem on a graph, for which efficient global optimization strategies are available, such as dynamic programming or min cut/max flow algorithms.

2.5.2 Classification/Regression

For the energy function $E(\theta|I,\gamma)$ to be relevant and sufficiently informative for a specific image analysis problem, realistic and sufficiently sophisticated formulations of the data likelihood $\mathrm{Prob}(I|M(\theta))$ and the prior $\mathrm{Prob}(M(\theta))$ are needed. These can in principle be derived by statistical analysis of a set of similar, previously analyzed images $\mathscr{T} = \{(I_i,\theta_i), i = 1\ldots n\}$. In that case, a direct estimate $\mathscr{P}(\theta|I,\mathscr{T})$ of the posterior probability $\mathrm{Prob}(M(\theta)|I)$ may then as well be derived from the training data $\mathscr{T}$, such that the optimal model instance $M(\theta^*)$ may be obtained by maximization of $\mathscr{P}$:

$$\theta^* = \arg\max_\theta \mathscr{P}(\theta|I,\mathscr{T}) = \Phi(I|\mathscr{T}) \tag{2.12}$$

The function Φ maps every input I onto the most likely output θ^* based on the given training data $\mathscr{T}$. The output θ can be defined as a single value for the entire image (e.g., in case of image classification), or voxel-wise (e.g., in case of segmentation). In case θ takes on discrete values (e.g., object class labels), the function Φ acts as a classifier, while if θ takes on continuous values (e.g., coordinates), Φ acts as a regression function. In practice, the dimensionality of the image I (i.e., the number of voxels) is high, while the number of training samples n is usually much smaller, such that the estimation of Φ is ill-conditioned. Hence, the problem is often simplified by estimating a function Φ_f based on a limited number of pre-defined features $f = \mathscr{F}(I)$ that are computed from the image I:

$$\theta_f^* = \arg\max_\theta \mathscr{P}(\theta|\mathscr{F}(I), \mathscr{T}) = \Phi_f(f|\mathscr{T}) \tag{2.13}$$

Different machine learning strategies, such as k-nearest neighbors, support vector machines, or random forest decision trees, can be applied to construct the mapping Φ_f for a given training set $\mathscr{T}$ and a given set of features $\mathscr{F}$. During training, the parameters w that define the mapping $\Phi_f(f|w)$, whose representation depends on the chosen learning strategy, are iteratively refined such that estimation performance is maximized on the training set itself, using a measure L (a cost or loss function) that evaluates the difference between the given ground truth θ_i for each image I_i in the training set and the estimated $\theta_i'(w) = \Phi_f(\mathscr{F}(I_i)|w)$:

$$\mathscr{L}(w|\mathscr{T}) = \sum_{i\in\mathscr{T}} L(\theta_i, \theta_i'(w)) \tag{2.14}$$

$$w^* = \arg\min_w \mathscr{L}(w|\mathscr{T}) \tag{2.15}$$

By cross-validation against a separate validation set $\mathscr{V}$ of additional instances $\{(I_j, \theta_j), j = 1\ldots m\}$ that are not used for training, the generalization potential of the learned mapping $\Phi_f(f|w^*)$ to new, previously unseen cases can be assessed. Large differences in performance between training and validation sets ($\mathscr{L}(w^*|\mathscr{T}) << \mathscr{L}(w^*|\mathscr{V})$) are an indication that the mapping $\Phi_f(w^*)$ is overfitted and that a simpler, less flexible, and more regularized mapping would be more appropriate.

Feature vector-based classification/regression is a very flexible approach for image analysis, in the sense that multiple, separately computed sets of features f_k, related to different object properties (e.g., contrast, texture, geometry, context, etc.) or computed from different subparts of the data (e.g., multi-parametric MRI), can be simply combined by concatenating them into an aggregated feature vector $f = (f_1, f_2, \ldots)$. Moreover, additional, non-image related features (e.g., clinical parameters, genetic information, etc.) can be easily incorporated in the same way. However, extending the feature vector also increases the dimensionality of the problem and may necessitate proper prior scaling of the different feature ranges, which makes a robust estimation of Φ_f more complex and increases the risk of overfitting. Hence, in practice, a mechanism for optimal feature selection will be typically applied to reduce the dimensionality of the feature space by only maintaining the most relevant features (or combinations thereof) yielding optimal validation performance.

A drawback of conventional feature vector classification/regression is that the initial set of features $\mathscr{F}(I)$ is pre-defined by the user, which is typically done heuristically and therefore likely suboptimal. Current state-of-the-art machine learning approaches, in particular deep learning using convolutional neural networks with several consecutive hidden layers, alleviate this problem by performing optimal feature computation and selection during training, thus effectively estimating the mapping Φ directly from the data I itself. Due to the large number of parameters in such networks, different aspects related to network architecture, optimization, regularization, data sampling, and augmentation have to be carefully considered.

2.6 Computational Strategies

Model-based computational strategies for medical image computing can be broadly classified as either flexible shape fitting or pixel classification.

2.6.1 Flexible Shape Fitting

Flexible shape fitting makes use of a more or less global parametric model of the image appearance of the object, including photometric, geometric, and contextual properties, that is fitted to the actual image data by optimization of an objective function that evaluates the goodness of fit of the model instance. This objective function is often formulated as an energy function and finding the optimal model instance is solved as an energy minimization process. The energy function consists of both external, data-dependent energy terms, which aim at driving the geometry of the model instance to deform towards relevant photometric features of interest in the image, and internal, data-independent energy terms, which serve to constrain the flexibility of the model and account for prior knowledge about the shape of the object. The external energy of the model can be based on local boundary features or more global regional features, and can be defined heuristically, be derived from a statistical modeling of these features, or rely on a specific model of the image acquisition. The internal energy can be based on deterministic penalties imposed on the shape of the model instance (e.g., spatial smoothness), on a (partial or pseudo-realistic) biomechanical modeling of the object (e.g., elastic deformation behavior), or on statistical constraints (e.g., the expected shape of the object and its shape variability). Contextual information, for instance, the relative position of different objects, can also be accounted for by incorporating cross-terms in the internal energy functions of these objects. Typically, the balance between both energy contributions, i.e., between fidelity to the model and congruency to the data, is controlled by user-specified weights that need to be tuned for each specific application.

The geometry of the model can be represented explicitly, i.e., landmark-based, by representing the model as a set of discrete points defined by their 3D coordinates in some suitable space, typically arranged in a graph (a 1D curve, a 2D surface, a 3D mesh, etc.), with or without underlying continuous analytical parameterization (e.g., a piece-wise polynomial function). Alternatively, the geometry of the model can be represented implicitly, i.e., image-based, by representing the model as a picture, e.g., a gray value image, label image, probability map, or distance map, whereby the geometry of the model is intrinsically related to the regular spatial grid on which the image is defined. In case of a landmark-based shape model, the shape is altered by modifying the coordinates of the landmarks, either individually or jointly. In case of an image-based shape model, the shape is altered by modifying the intensities of the image, either directly or indirectly, namely by a transformation (affine or non-rigid) of the underlying image grid. Examples of flexible shape fitting strategies using an explicit shape representation include, for instance, active contours (“snakes”) [8], active shape models [9], and active appearance models [10]. Examples of flexible shape fitting strategies using an implicit shape representation include level sets [11], graph cuts [12], eigenfaces [13], and intensity-based non-rigid registration.

While deterministic constraints imposed on the flexibility of the model are necessarily largely heuristic in nature, statistical models aim at avoiding heuristics by learning suitable model constraints from the data itself, based on a representative training set of examples, typically derived from a database of similar images acquired from different subjects (see Fig. 2.1 for an example). A popular strategy for landmark-based statistical shape modeling are point distribution models (PDM) [14]. A PDM is constructed by statistical analysis of the observed variations in the locations of corresponding landmark points defined on all object shapes in a representative training set of shape instances, after appropriate spatial normalization of all shapes to a common coordinate space to eliminate irrelevant, pose-related variability. In practice, the construction of PDMs is complicated by the fact that a sufficiently large training set of previously segmented shape instances needs to be available, which usually involves manual delineation in 3D images, which is tedious and time-consuming for a large collection of images, error-prone and subject to intra- and inter-rater variability and uncertainty. Moreover, a sufficiently large

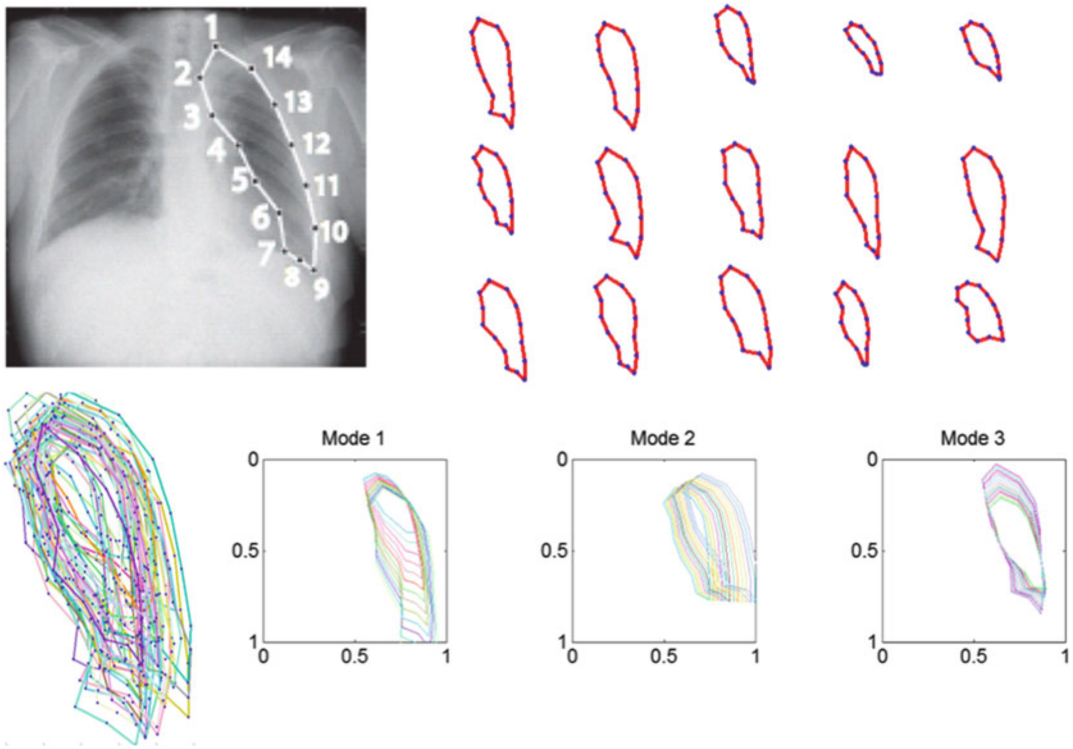


Fig. 2.1 Flexible shape fitting: example of a statistical shape model of the left lung in a 2D thorax radiography. The model is constructed from a set of example shapes that were manually delineated in images of different subjects. All shapes are represented by the same number of corresponding landmarks. The shapes are first aligned to eliminate irrelevant differences in pose between different subjects, after which principal component analysis is applied on the landmark coordinates to yield the mean shape and the main modes of variation around the mean. This shape model can be used for automated segmentation of the lung in new images by augmenting it with a photometric model of the local intensity patterns around each landmark

set of point correspondences needs to have been established between all shape instances, which by itself is a non-trivial problem for which different strategies have been proposed [15].

In case of image-based models, fitting of the model to new data can be formulated as an image registration problem. Image registration establishes dense spatial correspondences (i.e., a coordinate transformation) between two images based on a suitable local or global similarity measure, such that prior information from one can be propagated onto the other or vice versa (if the transformation is invertible). While registration using an affine transformation only compensates for global differences in pose and scaling between both images, most applications require a more flexible, non-rigid transformation to accommodate for local shape differences. While affine image registration can usually be achieved automatically and robustly in a variety of applications using maximization of mutual information [16, 17], non-rigid image registration is ill-posed due to the large number of degrees of freedom and ambiguity in the images, for instance, in homogeneous regions. Regularization of the registration problem is therefore required to constrain the solution space to include only deformation fields that are physically acceptable and to smoothly propagate or extrapolate the registration solution from sites with salient registration features towards regions where registration clues are absent or ambiguous. One popular approach is the use of analytical basis functions to represent the deformation field, especially B-splines with local support [18]. The flexibility of the deformation and the number of degrees of freedom is controlled by the spacing between spline control points, for which a multi-resolution strategy is typically adopted. Smoothness of the deformation field is intrinsic to the parameterization at scales smaller than the control point spacing and may be imposed at larger scales by penalizing high spline curvature or excessive local volume changes. Alternatively, the deformation field can be represented as a 3D, discrete vector field that is typically regularized by adopting a physics-based elastic or viscous fluid model or a diffeomorphic deformation framework [19]. In theory, a statistical deformation model could be derived in a similar way as for landmark-based models, but this is complicated in practice by the large number of degrees of freedom in non-rigid registration, requiring a large number of examples to capture fine-scaled statistically meaningful correlations.

Image registration is frequently used in medical image analysis for inter-subject spatial normalization, for the construction of mean shape templates (atlases), for atlas-based segmentation, for quantification of local shape differences and characterization of shape variability between groups, and for spatio-temporal analysis of motion or disease evolution. Atlas-based image segmentation makes use of a prior model in the form of an atlas, typically consisting of a gray value template and associated binary or probabilistic label images, either derived from a single subject or from a mean-shape average of multiple subjects. To avoid bias in the analysis introduced by the atlas, the atlas may be stratified or specifically constructed for the specific population of subjects of interest (e.g., age, disease status). Alternatively, multiple suitable atlases may be selected from a collection of images, each applied separately to generate a segmentation and the resulting segmentations fused, for instance, using a majority voting scheme.

2.6.2 Pixel Classification

Pixel classification aims at assigning an object label or its probability to each voxel in the image individually, mainly based on local intensity information alone. The classification can be supervised or unsupervised.

Model-based unsupervised classification adopts a parametric model for the expected intensities of the objects of interest, typically a Gaussian mixture model, and estimates the optimal parameters of the model and the classification simultaneously by maximizing the posterior probability of the labels given the data and the model, for instance, using the expectation-maximization (EM) algorithm [20]. Local spatial constraints can be imposed on the classification by formulating them as a Markov random field. In addition, more global a priori spatial information about the objects of interest can be included as prior probability maps, typically derived from an atlas that is first registered to the images and that also serve to initiate the EM algorithm, such that the procedure becomes completely automated. Atlas-based classification and classification-based atlas registration can be unified in a single algorithm, such that both are iteratively refined. By extending this approach to the simultaneous segmentation, atlas construction and clustering of a population of multiple, possibly heterogeneous images, automated data-driven detection of morphological differences between subpopulations becomes feasible using this framework.

Intensity-based classification of pixels can be extended to more general feature-based classification of individual pixels (e.g., for segmentation) or entire images (e.g., for disease staging) whereby a vector of image-derived features, possibly augmented with non-imaging features (e.g., patient age, gender, clinical findings, genetic information, etc.) is computed for each pixel or per image. As the dimensionality of the feature vector increases and the features themselves are more diverse, the adoption of an underlying model for the feature distribution (e.g., multivariate Gaussian) is less justified. Hence, instead of assuming a specific feature model, supervised classification methods estimate decision boundaries between different object classes in the high-dimensional feature space based on a training set of positive and negative samples of each class. Different generic classifiers can be used for

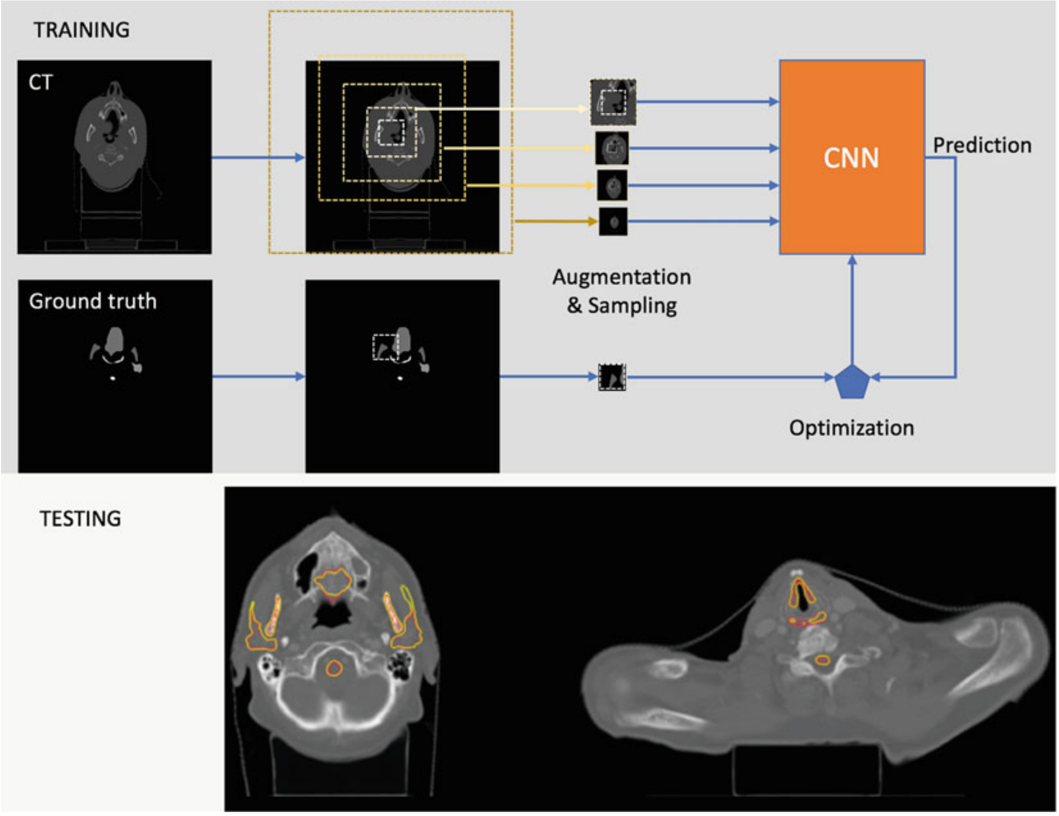


Fig. 2.2 Pixel classification: example of a CNN trained for organ at risk delineation in a head and neck CT scan used for treatment planning in radiotherapy. The neural network predicts the object labels in a small patch around each voxel based on local intensity information at different scales and is trained by comparing its prediction (yellow) to the manual ground truth (red) using cross-entropy as loss function. The classifier generates delineations of multiple organs simultaneously: parotid glands, oral cavity, mandible, brainstem (left), glottic larynx, esophagus, spinal cord (right)

this purpose, such as support vector machines, random forests, or neural networks. A drawback is that suitable features need to have been defined in advance, such that the performance of these methods is impacted by feature selection and by the need for dimensionality reduction of the feature vectors to avoid overfitting in case of limited training data.

Deep learning deals with the issue of optimal feature selection by using a neural network with multiple hidden layers to learn the optimal features simultaneously while training the classifier such that overall classification performance is optimized [21]. Due to the very large number of parameters in such networks, regularization is required to avoid overfitting and to ensure generalization to unseen data. Deep neural networks, and especially convolutional neural networks, are an active topic of research in medical image analysis (see Fig. 2.2 for an example), as these approaches currently achieve the best performance for object detection and segmentation in many applications [22].

2.7 Fundamental Issues

When selecting or designing an appropriate computational strategy for model-based image analysis in a particular application, some fundamental choices or issues related to the model representation and fitting have to be considered.

2.7.1 Explicit Versus Implicit Representation of Geometry

Object shape can be described using points, contours or meshes with explicitly specified coordinates, or as a picture with implicit geometry implied by the image grid. Flexible model-to-image fitting using an explicit geometric model representation is typically driven by local photometric features along or inside the object boundary, while fitting an implicit iconic model (e.g., an atlas) involves deformable image registration. While statistical models could be built for each representation independently, unified hybrid explicit/implicit models have also been conceived in which variability in object boundary shape and iconic deformations are considered jointly, for instance, by combining explicit landmark-based and implicit picture-based shape models in a single strategy for model-based segmentation and registration. The implicit model provides dense correspondences between a pictorial shape template and the image to be segmented by non-rigid registration constrained by statistical deformation modes. The explicit model provides point correspondences at landmark points (defined in the space of the implicit model) by flexible shape fitting constrained by a statistical model of shape variability (e.g., a PDM). While the explicit model is driven by intensity information at the object boundary specifically, the implicit model would bring the overall intensity appearance of the object into account. A matching strategy incorporating both models is thus expected to be more robust than each of the models separately. Moreover, by imposing that both models must be consistent at the boundary points, matching information can be propagated from one model onto the other and vice versa. Also, training of the landmark-based and picture-based shape models can best be performed simultaneously to exploit likely correlations between both.

2.7.2 Global Versus Local Representations of Appearance

A global model representation is based on parameters that each affect the overall (photometric and/or geometric) appearance of the object. This is advantageous for imposing global shape constraints and for propagating evidence about a suitable model fit from image regions with salient clues towards regions where such clues are absent. Local representations on the other hand have parameters that each describe the object appearance only locally, thus providing additional flexibility to adapt to local deviations. A unified hybrid local/global model can provide control at small scale, while at the same time imposing constraints over larger scales. Ideally, such multi-scale representation would be learned automatically from training data in order to describe the observed variability locally at the most optimal scales.

To tackle this problem, hybrid models that can represent shape variability at multiple scales are needed. For instance, in case of landmark-based shape modeling, shape constraints can be imposed between any two landmarks in a graph-like structure, both locally between adjacent landmarks and globally between more distal landmarks. The concatenation of multiple such partial statistical constraints in a single objective function may be advantageous over a single global PDM, as it requires less training samples and provides more flexibility to deal with local shape distortions. While such dependencies can be simply assumed and imposed deterministically based on application-specific heuristics, such heuristics can be avoided by learning local and global spatial correlations from the data itself. During model fitting, such dependencies need not be treated equivalently, but could be weighted to put more emphasis on shape constraints between the most correlated landmark points.

When assessing groupwise morphological variation between populations of subjects based on dense spatial correspondences established by non-rigid registration between all images, regional shape-related imaging biomarkers can be discovered based on features of interest that are derived from the deformation fields (e.g., the local Jacobian determinant of the deformation, possibly at multiple scales) obtained by a previously applied deterministic, groupwise image registration process. However, these deformation fields are likely biased by the implicit assumptions (e.g., degrees of freedom, similarity measure, regularization) made during this preprocessing stage. Such bias could be avoided by establishing dense spatial correspondences directly from the image data without having to adopt a specific deformation model.

2.7.3 Deterministic Versus Statistical Models

Shape models in medical image analysis are often deterministic, i.e., mainly heuristic, for instance, an object represented as a 2-D flexible curve with intrinsic smoothness properties, a 3-D elastic mesh, or a deformable template with physically based deformation properties. The objective function used for fitting the model to an image is usually formulated as an energy function or similarity measure, which typically consists of a weighted combination of an external energy term, assessing the agreement of the fitted model instance with relevant photometric features derived from the image data, and one or more internal energy terms or penalty functions, imposing geometric constraints on the model. Such deterministic models are often too generic in nature and their proper behavior overly depends on the tuning of ad hoc parameters and on suitable initialization. The behavior of the model can be made more robust by constraining its flexibility by incorporating application-specific knowledge about the expected variability in image appearance of the object of interest as deduced from a representative ensemble of exemplars, i.e., a statistical appearance model. However, the initial training of such models typically relies on a deterministic approach for extracting suitable photometric features and for establishing spatial correspondences between different exemplars in the training set. The bias introduced in the model by these implicit assumptions adopted during model construction may not be ignored in case the available training data is limited. A unified hybrid deterministic/statistical modeling and model fitting approach would be able to initiate itself from small data ensembles based on generic ad-hoc assumptions that could be gradually re-assessed as the model is built up and traded for more specific knowledge about probable variability in the data as more exemplars become available.

2.7.4 Data Congruency Versus Model Fidelity

Objective functions to be optimized during model fitting typically show a trade-off between assuring fidelity of the model instance to the expected object properties versus maximizing the goodness-of-fit of the model instance to the data. Typically, with deterministic modeling approaches, this trade-off becomes apparent in the form of weight parameters that have to be set heuristically to balance the influence of different terms in the objective function in order to obtain suitable behavior of the model in a specific application. Such heuristics would not be needed if the photometric and geometric variability of the object and their interdependence would be fully modeled statistically based on joint probability models derived from actual data, instead of relying on simplifying deterministic assumptions. Model fitting is then formulated as maximizing the posterior probability of the model parameters given the observed data and their a-priori distribution. This model fitting should be robust to deviations from the expected photometric and geometric variability, due to imaging artifacts and/or pathological conditions. In practice, however, the construction of suitable priors is complicated by the limited amount of training data.

2.8 Conclusion

The evolution towards a more personalized medicine requires the analysis of a large amount of imaging (and non-imaging) data. Robust automated methods are essential for a more efficient and more accurate analysis of multi-modality images in clinical practice in the context of early diagnosis, optimal treatment planning, and treatment follow-up. Medical image computing benefits from advances in machine learning to develop data-driven model-based image analysis strategies that are less biased by heuristic assumptions about the appearance of the objects in the images. Supervised learning using deep convolutional neural networks appears promising for various applications in medical image analysis, although the large number of parameters in these networks and the limited amount of training data available in most applications pose specific challenges.

References

1. Suetens P. Fundamentals of medical imaging. 3rd ed. Cambridge: Cambridge University Press; 2017.
2. Bankman IN. Handbook of medical imaging: processing and analysis. San Diego: Academic; 2000.
3. Hajnal JV, Hill DLG, Hawkes DJ. Medical image registration. Boca Raton: CRC Press; 2001.
4. Suetens P, Fua P, Hanson AJ. Computational strategies for object recognition. ACM Comput Surv. 1992;24(1):5–61.
5. Prince SJD. Computer vision: models, learning, and inference. Cambridge: Cambridge University Press; 2012.
6. Greenspan H, van Ginneken B, Summers RM. Deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans Med Imaging. 2016;35(5):1153–9.
7. Zhao F, Xie X. Energy minimization in medical image analysis: methodologies & applications. Int J Numer Methods Biomed Eng. 2015;32:1–63.
8. Kass M, Witkin A, Terzopoulos D. Snakes: active contour models. Int J Comput Vis. 1988;1(4):321–31.
9. Cootes TF, Taylor CJ, Cooper DH, Graham J. Active shape models: their training and application. Comput Vis Image Underst. 1995;61(1):38–59.
10. Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. IEEE Trans Med Imaging. 2001;23(6):681–5.
11. Osher S, Paragios N. Geometric level set methods in imaging, vision, and graphics. Berlin: Springer; 2003.
12. Boykov Y, Veksler O, Zabih R. Fast approximate energy minimization via graph cuts. IEEE Trans Pattern Anal Mach Intell. 2001;23(11):1222–39.
13. Turk M, Pentland AP. Eigenfaces for recognition. J Cogn Neurosci. 1991;3(1):71–96.
14. Heimann T, Meinzer H-P. Statistical shape models for 3D medical image segmentation: a review. Med Image Anal. 2009;13:543–63.
15. Davies R, Twining C, Taylor C. Statistical models of shape: optimisation and evaluation. Berlin: Springer; 2008.
16. Maes F, Collignon A, Vandermeulen D, Marchal G, Suetens P. Multimodality image registration by maximization of mutual information. IEEE Trans Med Imaging. 1997;16:187–98.
17. Pluim JPW, Maintz JBA, Viergever MA. Mutual-information-based registration of medical images: a survey. IEEE Trans Med Imaging. 2003;22(8):986–1004.
18. Rueckert D, Sonoda LI, Hayes C, Hill DL, Leach MO, Hawkes DJ. Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans Med Imaging. 1996;18(8):712–21.
19. Ashburner J. A fast diffeomorphic image registration algorithm. Neuroimage. 2007;38(1):95–113.
20. Van Leemput K, Maes F, Vandermeulen D, Suetens P. Automated model-based tissue classification of MR images of the brain. IEEE Trans Med Imaging. 1999;18(10):897–908.
21. LeCun Y, Bengio Y, Hinton GE. Deep learning. Nature. 2015;521:436–44.
22. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.

3 A Deeper Understanding of Deep Learning

Bart M. ter Haar Romeny

3.1 Introduction

It was a struggle for decades, but finally computer-aided diagnosis (CAD) begins to work. Today we are confronted with more and more examples where the performance of deep learning systems comes close to human performance. At RSNA 2017 more than 100 companies showcased AI products, but it has not penetrated yet on the working floor. One reason may be the unfamiliarity with the concepts, fear of giving a lot of responsibility to a seemingly "black box," and not having time yet to get familiar with it by testing it. In this chapter we discuss the engineering technicalities of AI in some depth, as well as the relation to biological perception and some of the highly dynamic AI development world.

There are numerous high-quality reviews covering the current state of the art of deep learning in radiological application areas and its impact, see, e.g., [12, 20, 30]. The field now covers virtually all fields where human recognition plays a role, like language translation, genetic analysis, social media analysis, but in this chapter the focus is on visual analysis: image recognition and classification.

A solid mathematical theory or model of the internal functioning of the neural networks is still lacking, or at least only partly understood. The performance is in some areas so impressive, that the field is taken by storm: not only radiologists but also long-time computer vision scientists take turns in their career.

In this chapter I will especially focus on the intuitive mechanisms behind an important class of networks for images, i.e., *convolutional neural networks*, how they work, how they are related to human visual perception, and discuss some insightful models and terminology. Terms in italics are explained in the glossary.

The paper is organized as follows: After a discussion of a classical CAD processing pipeline, we look at the concept of contextual processing. Then we study the deep layers of our human visual system, and the layers of a general convolutional neural network (CNN), as this is the type of network most effective in imaging. We give examples of current network architectures, pointers on the web for further study, and a medical application in large-scale retinal screening for diabetic retinopathy. We then dive somewhat deeper in the mechanism of self-organization of the filters, both in physiology and artificial networks. The brain is still far outperforming our computers in terms of energy efficiency and speed, so we discuss some roadmaps where the field might go.

B. M. ter Haar Romeny (✉)
Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: B.M.terHaarRomeny@tue.nl

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_3

The paper is concluded with a discussion and some recommendations especially for radiology.

3.2 Computer-Aided Diagnosis, the Classical Approaches

The classical way to do pattern recognition for CAD is by designing hand-crafted detectors of low-level image features, like edges, lines, corners, etc. This is done with banks of special filters (also called kernels or templates). The filters are designed to look like the tiny pieces of structure they are supposed to detect. The shifting of a small filter over an image, row by row, is called *convolution*. With the many filters many properties are measured at each position or in small image patches, giving rise to a high-dimensional feature space. Clusters in this feature space are separated by the so-called classifiers in the desired classes, e.g., healthy or disease. A typical pipeline of processing stages is shown in Fig. 3.1.

3.3 Artificial Intelligence

Already decades ago people have tried to implement the intelligence of our human brain. First attempts focused on the so-called expert systems, with rule-based reasoning, like the "if-else" statements in programming. An illustrative example from 1984 is the case of US Campbell Soup Company: Aldo Camino, an expert with 46 years of experience, knew everything about the complex 22 m high sterilizers, which heated 68,000 cans of soup to 120 °C. If it went wrong, a lot of soup was lost. Aldo knew everything: "if this valve ticks, and the temperature there is too low, that valve must be opened further," etc. He flew from factory to factory, but was about to retire. It was decided to record his full knowledge in a large set of AI rules. Later this form of AI got stuck; it turned out to be impossible to keep discovering more rules and add them to the system. These relatively simple networks actually disappointed, and at the end of the 1990s this field was virtually given up.

3.4 Neural Networks

An important new family of ideas mimicked synaptic connections by weights in an artificial neural network (ANN). The neuron sums the weighted inputs and the output needs to pass a threshold, see Fig. 3.2. Learning is accomplished by iteratively adapting the weights in such a way that the error at the output is minimized. Typically the networks were shallow, not deep,

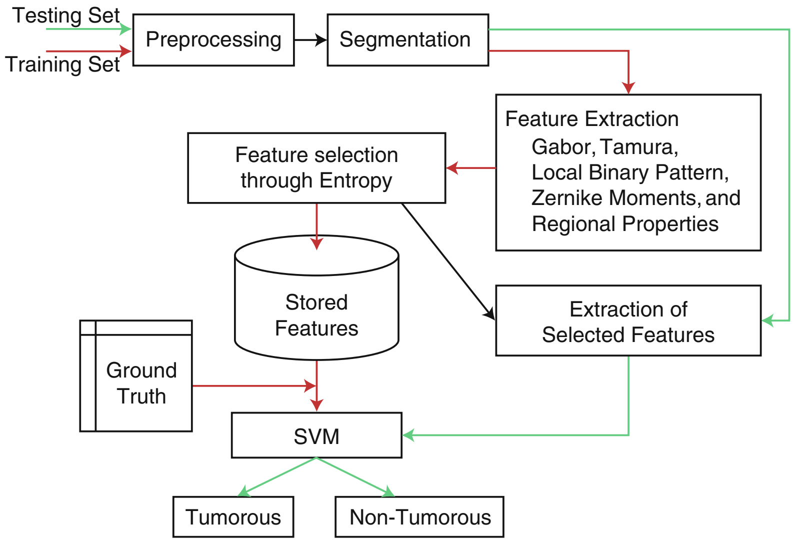


Fig. 3.1 A typical classical processing pipeline for computer-aided diagnosis. Many features are detected with hand-crafted filters

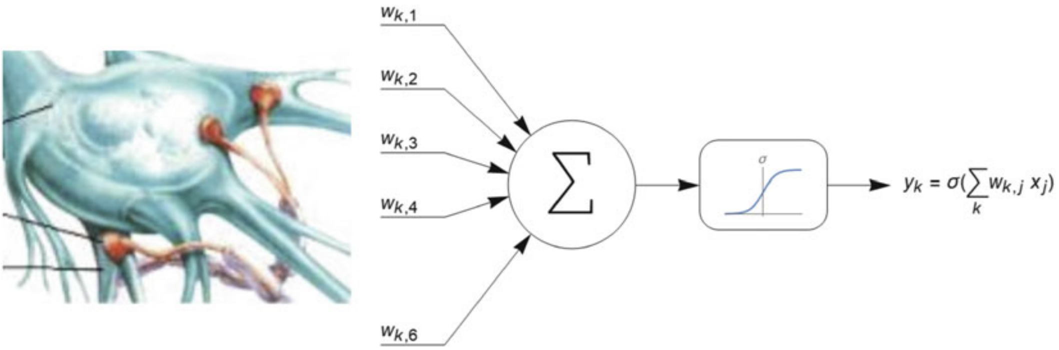


Fig. 3.2 Left: synapses connect between neurons. A larger synapse is a larger weight. Repetitions let synapses grow, the basis of learning. Right: a simple artificial neural network, summing weighted inputs and passing the output through a nonlinear threshold

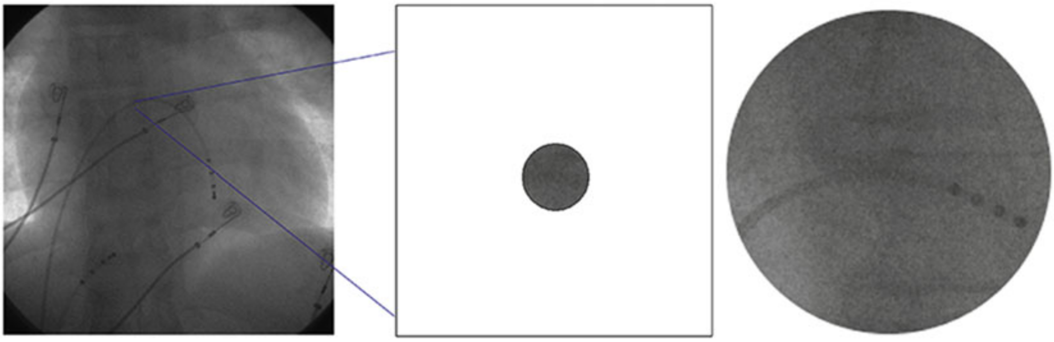


Fig. 3.3 The role of context is clear in this example of detection of a catheter in fluoroscopy. With only local detection, i.e., with a key-hole view, like in the middle zoomed-out section, nothing can be discriminated, but when the same small area is enriched with context, the catheter becomes clear

only a few (typically three) layers: input layer, hidden layer, and output layer (for layer locations see also Fig. 3.4). The weights were not known, but had to be learned. After each offered example, the error deviation at the end was used to adjust an initial random assignment of weights to more optimal values. After many, many learning cycles the error could be made lower and lower. However, also these ANNs only gave about 75% success rate.

The key idea turned out to make a *deep neural network*, i.e., with many more (10–100) internal layers [18, 19]. The functional mechanism is essentially to exploit the use of *context*. Human vision is known to make extensive use of the context around a given structure for recognition, see, e.g., the example in Fig. 3.3. The phenomenon is known as Gestalt since the 1930s [28, 29], but a mathematical theory for it was never fully established. The second insight is that it should be done by *incremental expansion of the contextual region*, in small steps, i.e., layer by layer. See Fig. 3.4. We were too greedy to do it all in only a few layers. Our human visual system also works with multiple layers, see Fig. 3.5. It is the most extensively studied brain area and functionality. It is estimated that one quarter of our brain, located in the back of our head, is for vision. So we are very much visual machines, which is abundantly clear by our use of images in everything we do. It is estimated that our visual system has at least 11 deep layers, in two major pathways, the dorsal parietal pathway for "what" and the ventral temporal pathway for "where."

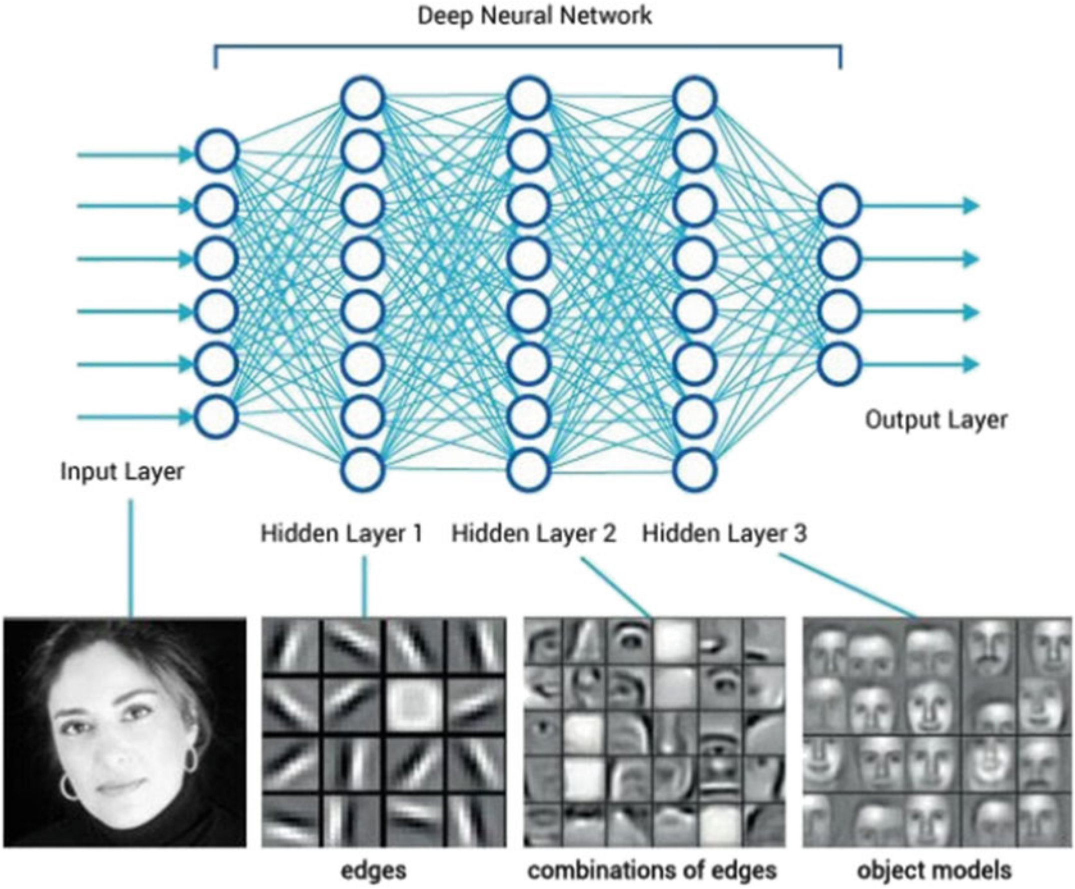


Fig. 3.4 A deep neural network to be trained for face recognition. In the first layer simple features are detected, like edges and lines. In the next layer the context is larger, and they are combined to parts, e.g., noses and mouths. The next layer looks larger again, and has learned faces, etc. The last layer of the network is the classifier, giving the probability to belong to a pre-destined number of classes as output

3.5 Convolutional Neural Networks

In 2012 the famous annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC, [14]) was won by the convolutional neural network *"AlexNet,"* by Krizhevski et al. [18], with an error rate of 15.3%, which was a stunning 10.8% better than the runner-up. See Fig. 3.6. The numbers next to the layers specify the structure of the network, described below.

It is instructive to study this particular network in more detail: the input image is 224 × 224 pixels, 3 bytes per pixel (RGB color). The convolution filters in the first layer are 11 × 11 pixels, and are shifted over the input image in steps (*stride*) of 4 pixels. The filtered output image of layer 1 has thus a size of 55 × 55 pixels. These first filters find edges and lines, and come in many orientations and scales (in this case 48). The 48 filtered output images of layer 1 form a data cube of 55 × 55 × 48, which is mathematically called a *tensor*[1] (a high-dimensional matrix). Because such a tensor is a serious data explosion, it is decimated with a so-called max-pooling layer: only the maximum of a 5 × 5 template is kept as input for layer 3. In the third layer 128 filters

[1] This explains the name of Google's deep learning software *TensorFlow* [11].

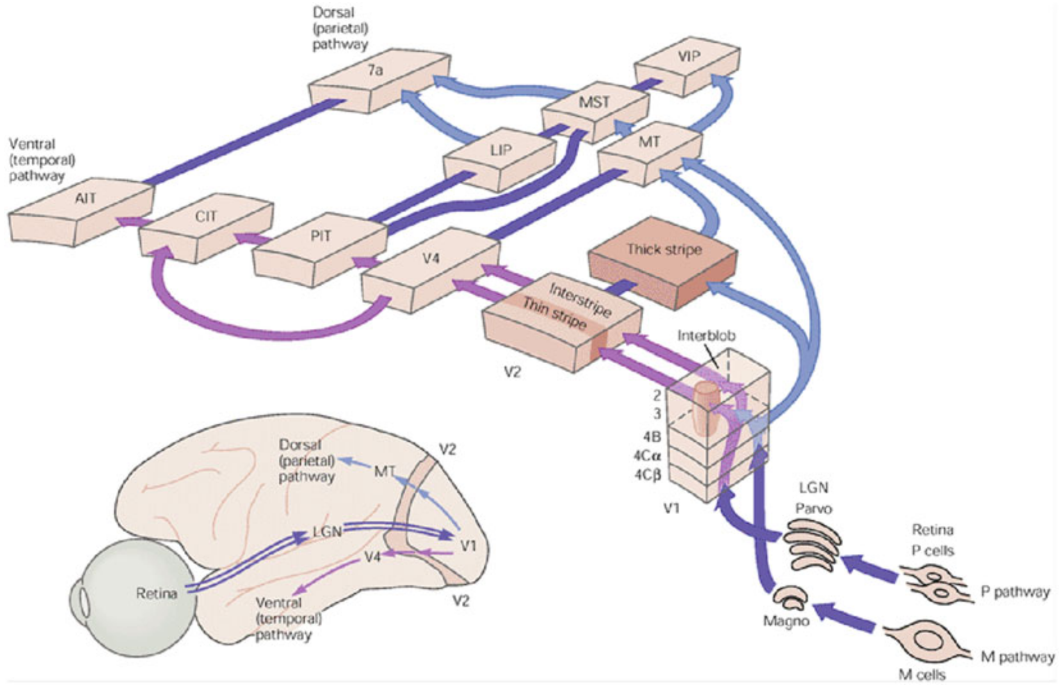


Fig. 3.5 Stages of the human visual system. From the retina (lower right) the signal travels to the lateral geniculate nucleus (LGN) in the thalamus, then to the primary visual cortex (V1) in the occipital brain, to V2, V3, V4, etc. From [16]

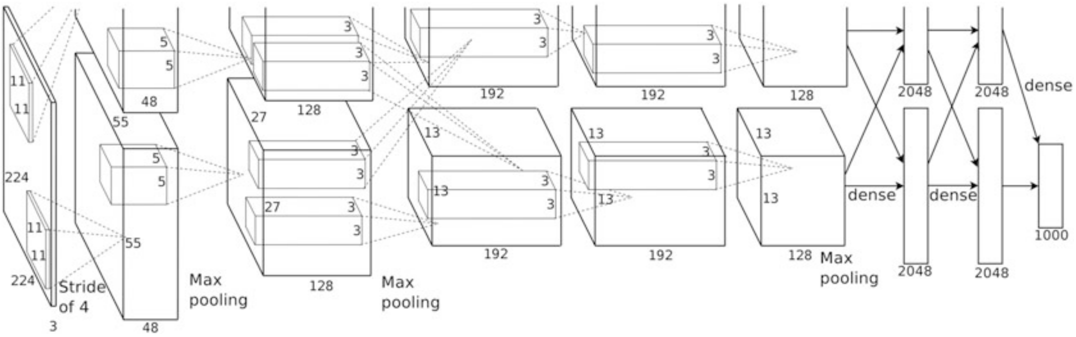


Fig. 3.6 AlexNet, a famous deep convolutional neural network. The numbers indicate the size and number of the applied filters. The *forward flow* is from left to right, where the complexity of the filters increases per stage, and the *error backpropagation* is from right to left, where the many internal network weights are adjusted one by one by minimizing the resulting output error (or *loss function*) by endless training cycles. From [18]

of size 3 × 3 are applied, max pooling again, etc. Layers can be stacked as Lego. The last layers are *fully connected layers*, which form the final classifier. Those can also be classical classifiers, such as a *support vector machine* (SVM) or a *random forest*. In our figure the three fully connected layers have as output 1000 categories [18].

The training (i.e., the actual learning) of the network is the labor- and time-consuming part. The (for a large network typically millions of) weights of the connections need to be adjusted using the training data, such that the classification error is minimized. This process is called *error backpropagation*. Powerful optimization techniques exist to do this optimally fast, using techniques as *gradient descent*. Training a large network can take hours, sometimes even weeks, even on very powerful computers.

When the full training data are exploited, the network is validated with the test data, i.e., to establish how well it actually works, specified in, e.g., sensitivity and specificity.

The trained network can be deployed as the electronic expert: image in, classification out. This process, the production stage, is fast. Most deep learning neural networks are written in the programming language *Python*, see also [8].

3.6 Why Now?

Three main reasons may explain the explosion of AI:

(I) Big Data Today big data are everywhere. Really big data. It is the primary key for the successful deep learning applications we see today.

The more data the better, e.g., for CAD, general language translation, for image classification, for self-driving cars, face recognition, etc. There is a clear trend on the internet of "the winner takes all." The big companies of today (Google, Facebook, Amazon, Microsoft, Uber, Baidu, Apple, TenCent, etc.) offer many services for free, such as Google Translate, Google Photos, Facebook face tagging, Google Streetview, to name a few. They pay to get your data.

Some examples: Over 1.2 billion photos are uploaded to Google Photos every day, which now stores 13.7 petabytes (10^{15}) of image information. In comparison, 350 million photos are uploaded every day to Facebook.

Every time one of its 1.65 billion users uploads a photo to Facebook and tags someone,[2] that person is helping the facial recognition algorithm. The company claims to be able to accurately identify a person 98% of the time.

Virtually all these big companies are developing new health care applications, starting from where the most data is available, e.g., *internet of things* (wearables, apps, medical devices, and sensors), electronic health records, and fundus photos to detect diabetic retinopathy [13], see also Sect. 3.7.

It is clear that in radiology the huge local PACS data repositories contain the big data for efficient training of our profession's AI networks, but much organizational and legislative work still has to be gone through.

(II) Graphical Processing Units (GPUs) Computer games need screen updates some 60 times per second, and the regular computer processors were not fast enough. A GPU is a single additional computing chip, located on the *game card* in a PC or laptop with typically hundreds (today up to 3800) of *parallel* processors. It is ideal hardware for deep learning training and deployment. The by far largest supplier is the US-based NVIDIA company. GPUs can be stacked in servers (see Fig. 3.7). They are typically programmed in the language CUDA.

(III) Smart Network Architectures A multitude of deep neural network architectures have been developed for specific purposes: *convolutional neural networks* for images, *residual neural networks* for faster training, *recurrent neural networks* for temporal and sequential data, *U-nets* for biomedical image segmentation, *generative adversarial networks* (GANs) to generate new images from a learned specific style, etc.

3.7 Example: Screening for Diabetic Retinopathy

A high-volume screening application, where brain-inspired quantitative geometric methods [26] are combined with deep neural networks have proven to be effective, is the Sino-Dutch RetinaCheck project [27]. In today's China 11.6% of the population is affected with

[2] The Facebook "like" button has been pressed 1.13 trillion times.

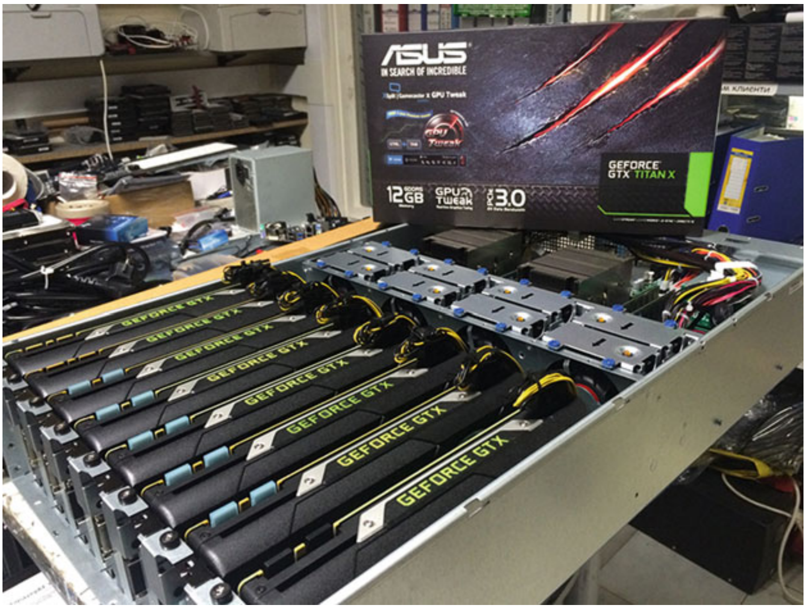


Fig. 3.7 Affordable mini supercomputer for AI: 19″ rack server with 8 NVIDIA's Titan Xp GPUs, with 8 × 3840 = 30,720 parallel cores. Price around 35 K€

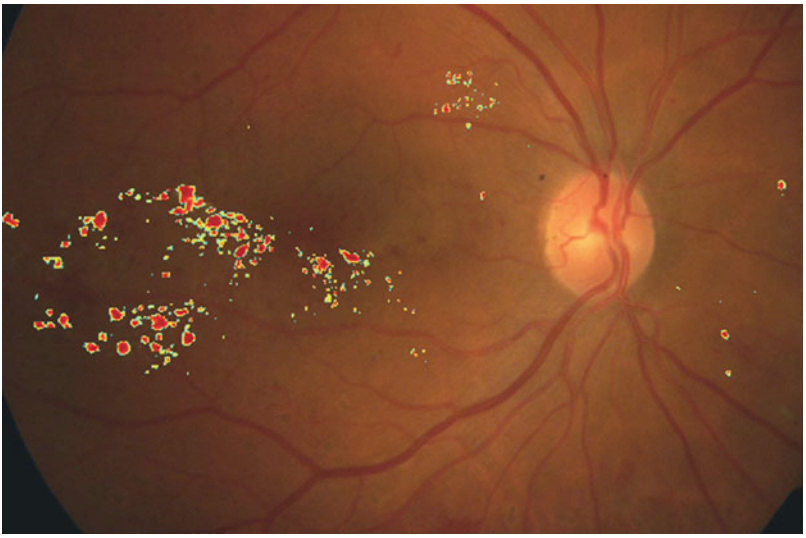


Fig. 3.8 Fully automatic exudate detection in color retinal fundus images with a residual deep neural network. Training is on image level, classification is on pixel level, surpassing the time-consuming annotation step. From [1]

diabetes, and 4% will develop total blindness. To detect early signs of microvascular dysfunction, fundus images are made at high resolution (12 megapixels) and low cost. By means of residual nets exudates could be automatically detected [1], see Fig. 3.8. In this application pathology in the training images was specified at the image level ("there is something somewhere"), while the network's output classified on the pixel level ("there is pathology in those specific pixels"), saving substantial annotation time.

3.8 Pointers on the Web

Information on deep learning on the web is overwhelming. Some useful resources are:

- **Grand challenges** in biomedical image analysis: a challenge is a public contest on a large database. This is now the norm on high-end medical image analysis conferences, e.g., MICCAI:

URL: http/grand-challenge.org/all_challenges/

- **Tutorial papers**, like [6, 9]
- **NVIDIA.com** blog: every day updated with new application areas of its GPUs:
  URL: http/blogs.nvidia.com
- **Medium.com**: scientific writing is more and more accompanied by blogs, written in magazine-style exploiting interactive multimedia, rated by readers:
  URL: www.medium.com
- **arXiv.org**: scientific publishing is too slow for modern computer science. All deep learning research papers are first claimed on this huge pre-publications server and resource. As such, it is a non-peer-reviewed publication platform, which means that it is useful for "insiders" but not as a reference, e.g., the evaluation of software for clinical purposes:
  URL: www.arxiv.org
- **ConvNetJS**: Karpathy's interactive demos of deep learning teaching examples, running on any browser, with visualizations of what the layers do:
  URL: http/cs.stanford.edu/people/karpathy/convnetjs/

3.9 A Comparison with Brain Research

The visual system, retina and visual cortex, is one of the most extensively studied areas of the brain [16, 17, 21]. Modern techniques like voltage sensitive dyes, optogenetics, diffusion-weighted MRI brain connectivity studies, and nano-scale crowd-segmentation of neurons [5] show neural anatomy ánd function from subcellular scale to functional regions.

It is a bit amazing that the worlds of AI and biology are still strikingly separate. Biological papers use few mathematics, AI papers often lack modern biological insights.

3.9.1 Brain Efficiency

Despite the fact that AI seems powerful, there is still a lot to be learned from the brain. A huge difference between our brain and modern computing and datacenters is that the brain is much more efficient: it uses 25 W only, while computing server centers typically use megawatts. Brain's neurons fire action potentials in the 1–6 kHz range, while every computer today works on processor clock speeds of 2–3 GHz. There is still huge room for efficiency improvement for our AI implementations!

What are the differences? We are still at the beginning of this long journey, but a few directions are becoming clear. Most AI methods exploit *supervised learning*, but the brain seems to learn much with *unsupervised learning*.

3.9.2 Visual Learning

It is instructive to study what happens in the very first stages of vision, the visual front-end [26]. Hubel and Wiesel, Nobel Prize winners 1981, found that the *receptive fields* (the neuronal correlates of filters) of the retinal ganglion cells have a center-surround structure. This is always explained as "lateral inhibition" or "surround suppression," but it must be the physiological implementation of local background subtraction (*batch normalization*) which is so essential for proper AI network functioning [15].

Hubel and Wiesel also found cells, higher up in the visual primary cortex V1, with receptive fields acting as simple edge and line detectors, the so-called simple cells. They can be modeled as mathematical operators: they take derivatives of the images: the first derivative extracts edges (differences between nextdoor pixels), the second order extracts lines, etc. It is no coincidence that the resulting filters in the first layer of any trained deep neural net resemble these V1 filters, see Fig. 3.9.

To introduce a bit of mathematics: the theory called *principal component analysis* (PCA) can express any variable data as a weighted sum of just a few basis components (the "principal components"). If we run the PCA algorithm on multiple small patches from a radiological image, e.g., a HRCT of lung with extensive fibrosis, we get the filters depicted in Fig. 3.9. These filters

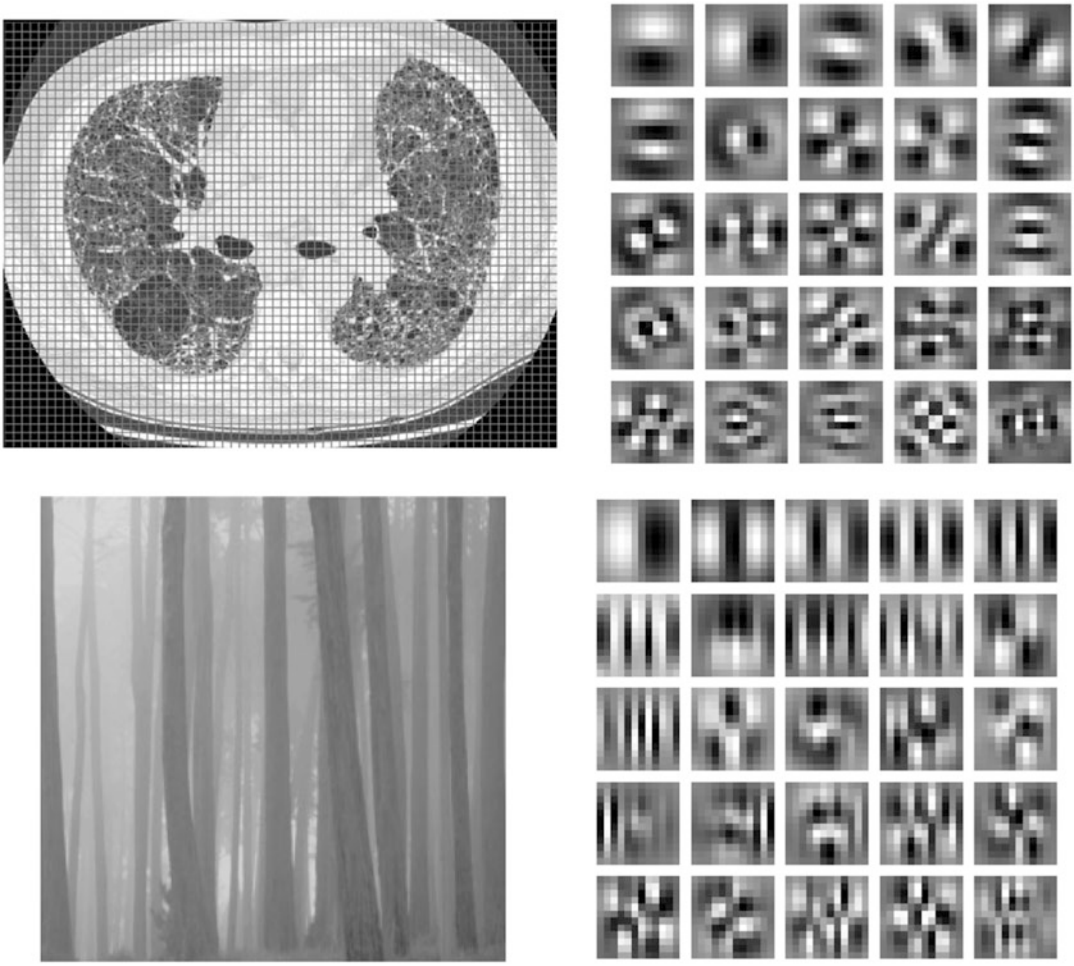


Fig. 3.9 Principal component analysis on image patches. Top left: original HRCT (source: Wikipedia HRCT) with the patches marked. Top right: resulting filters. Filters that show a black-white division measure the difference between two neighboring pixels, i.e., edges (first row, first filter: horizontal edge detection; first row, second filter: vertical edge detection). Bottom row: if only vertical structures are learned by PCA, as from this image with vertical trees, primarily filters result that detect the structure of vertical edges. Lesson: the filters are created by the data. From [26]

are the same as when learned by error backpropagation, but now acquired much faster.

Blakemore experimented with the development of receptive fields in a kitten from birth [3]. After being raised with only exposure to horizontal lines for 3 months, it had not developed receptive fields (filters) for vertical lines, and it could not discriminate a vertical stick. Lesson: also in brains: the filters are created by the data. See Fig. 3.10 and the movie [4].

We have the same deficiency for recognizing faces upside down, as they do not appear in our normal daily visual experience. See Fig. 3.11: Finding faces, even in an artistic drawing, is relatively easy for us,[3] can be much assisted by CAD algorithms, but recognition fails when the same image is turned upside down.

The lesson of the above is that our brain, and deep neural networks, learn the filters from the data. They do not need to be designed, or prewired. In this way just the right filter banks are

[3] The search term "faces everywhere" in Google Images gives many common objects in which faces are perceived.

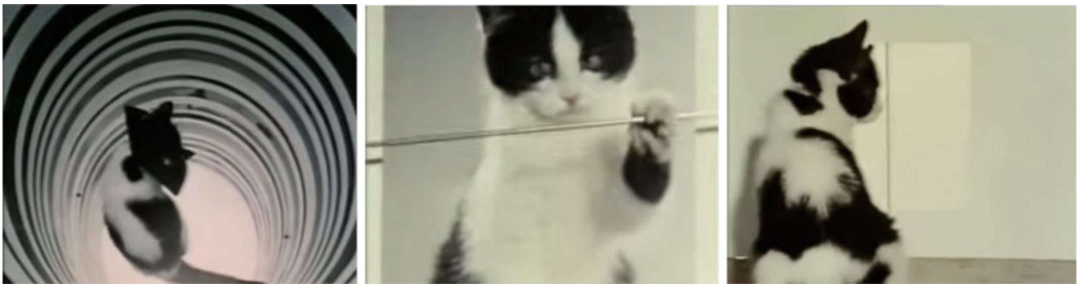


Fig. 3.10 Blakemore's cat. After 3 months seeing only horizontal lines from birth, it could not see a vertical stick. From [3,4]

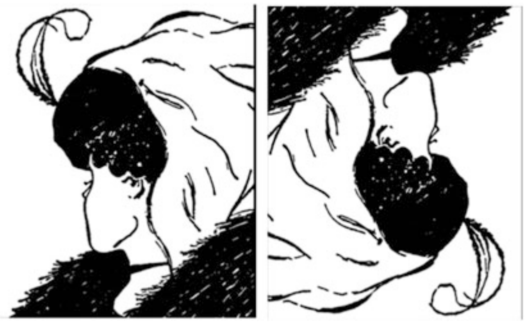


Fig. 3.11 Face detection. Left: A famous illusion, showing a young and old lady. Right: when the image is turned upside down, face recognition is much more difficult. We have never learned faces upside down. Art from [7]

found, with just the right filters: not too many, not too few.

3.9.3 Foveated Vision

It is sure that our brain exploits many energy saving strategies. One example is foveated vision. The resolution on our retina is not isotropic (the same everywhere), i.e., in the middle, at our fovea, we see sharper than at the periphery. There is a radial linear decrease of acuity with eccentricity, see Fig. 3.12. What might be the reason for this?

It is interesting how much we can learn from other fields, such as autonomous robot exploration. Here energy saving is also a key issue. We take an argument from Erik Nelson's 3D Simultaneous Localization and Mapping (3D SLAM) experiment [23]. The energy saving argument is that it is a waste to process all pixels equivalently, it is only needed in the attention areas. We have an excellent internal representation of our environment in memory in our higher visual centers, and only a few updates with our scanning movable eye is enough. This phenomenon is exploited in the 3D SLAM experiment on Fig. 3.13, where an environment is scanned with a so-called LIDAR, a laser-scanning distance measuring device on the head of the researcher (see the YouTube movie [22]). This sequential scan and the storage in memory turns out to be hugely effective in terms of processing and storage. This process can even run on a smartphone (it also fits into a mosquito brain, etc.).

Maybe this is the reason why we have only 1 million fibers in our optic nerve, but 150 million receptors (rods and cones)?

The non-isotropic retina was introduced into deep CNNs by Ghafoorian et al. [10], see

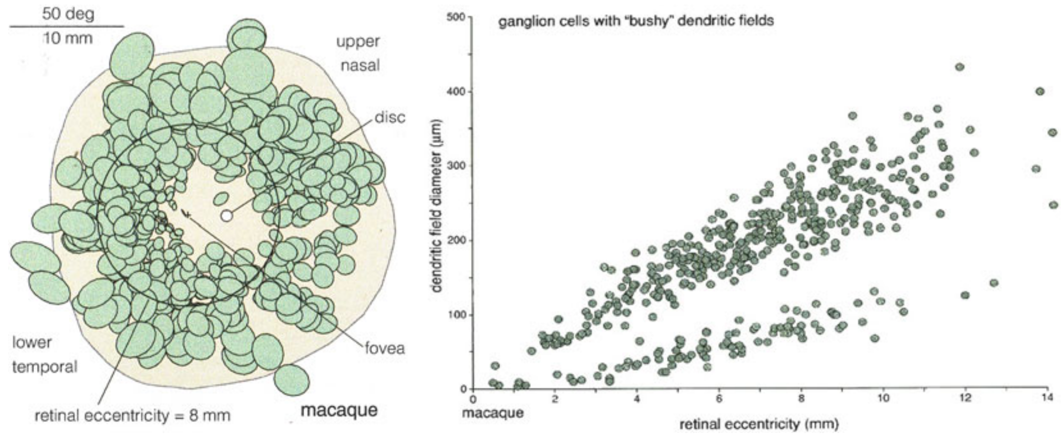


Fig. 3.12 The diameter of retinal receptive fields increases linearly with eccentricity (macaque). Left: dendritic field size mapped on the retina. Right: diameter as function of eccentricity; top cluster: parasol cells (for motion), bottom cluster: midget cells (for shape). Adapted from [25]

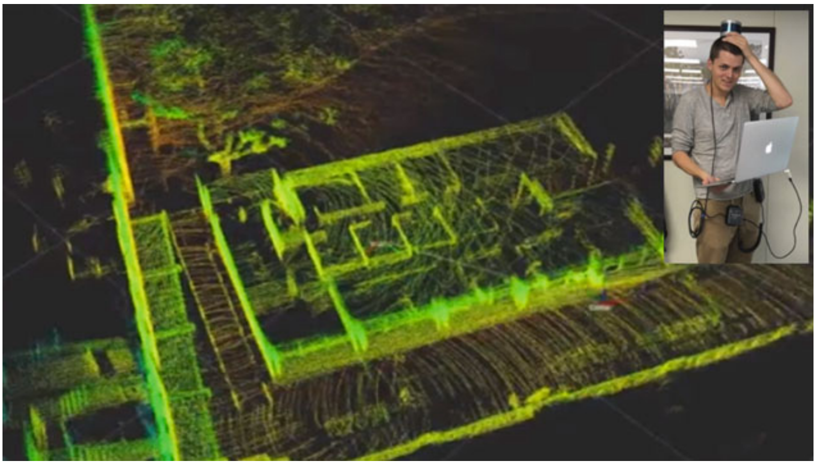


Fig. 3.13 3D Simultaneous Localization and Mapping (3D-SLAM), with a portable LIDAR distance scanner, mimicking our scanning fovea in a low-resolution retina. This is a very efficient operation, saving processing power and memory storage. From the YouTube movie [22], see also [23]

Fig. 3.14. The convolution filters have decreasing resolution from the center of the filters. The proposed method outperformed identical CNNs with uniform patches of the same size (*Dice coefficient* 0.780 ↔ 0.736), and got very close to the performance of an independent human expert (Dice coefficient 0.796).

In short, these examples point to an important lesson that much still can be learned, especially from cross-disciplinary fields, to optimize our current low-efficiency deep neural networks. In other words: there is still much to come in AI, we are still at the beginning of this revolution.

3.10 Conclusions and Recommendations

The key to a well-performing CNN is a deep network topology, with incrementally increasing contextual analysis. Up to millions of network weights can be properly adjusted, using error

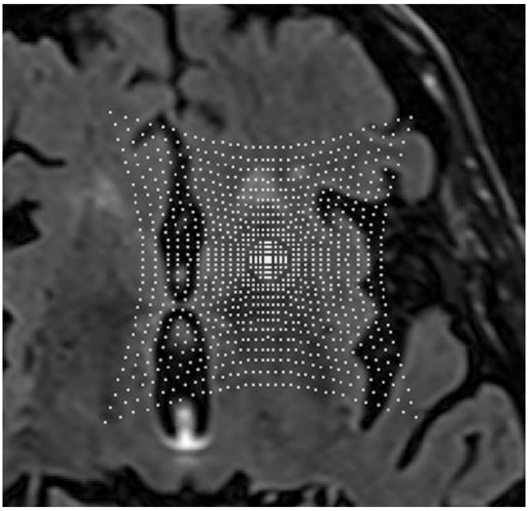


Fig. 3.14 Foveated convolution kernel. The resolution of the filter decreases with distance from the center, making it more efficient over a larger region. From [10]

backpropagation, by training the network with huge amounts of data. The trained network is stored: it is the expert now. For images the winning networks are convolutional neural networks.

When do CNNs work, when not? CNNs can learn faster from much more data: they finally will beat the experts; however, up till now only in specific cases: where context learning is evident, and when enough training data are available. Important new application areas include fields where people find difficulties, e.g., in judging the huge images of histology and pathology. Metadata from the electronic patient record (EPR) do help by combining them in the training, in the way that they assist humans. If a CNN needs to be designed, it is always good to compare a CNN to a human expert. E.g. estimating the necessary number of training samples is about equal to how many case studies human experts need to see during their professional training.

Big data is what it is all about. Actually, big data is more important than the specific network architecture. The networks come in a few classes, big data in radiology is still mostly hidden in protected PACS environments.

The big companies active in AI today will surely enter the healthcare market aggressively, knowing that the winner takes it all. It is a role of the national societies to develop a national strategy how to deal with the golden data. Many new applications will be developed in close harmony between radiology departments, research institutes, and companies, starting with data from population imaging and screening studies.

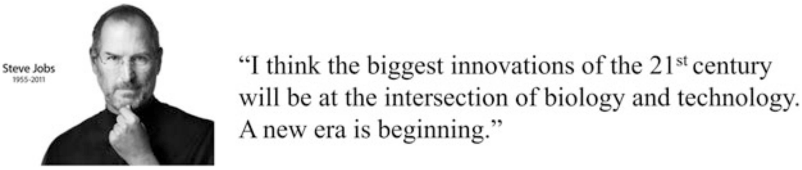
It is interesting to note that deep CNNs also work in 3D, i.e., with 3D voxel filters, convolving with the 3D data. Here AI has a notch. Actually, convolutions can be implemented for any number of dimensions.

The fast introduction of AI in radiology is evident. The main societies (RSNA, ESR, EUSOMII, SIIM, MICCAI, ACR (which founded the Data Science Institute—DSI [2]), etc.) all abound on attention for application reports, tutorials, challenges, legislative studies, and have workshops to discuss where and why AI applications clearly fail. The computer vision world is now fully dedicated to developing new deep learning paradigms and algorithms, new network topologies, validation data sets, challenges, and open source shared software. Hospitals and companies started having joined hands-on sessions, where multiple vendors can be compared in clinical practice and workflow [24].

A good advice for high-end radiology department is to hire local AI specialists, who can teach, develop, and critically judge. Modern biomedical engineers today are fully equipped with state-of-the-art knowledge of the AI field. In the Netherlands we are very well equipped with knowledgeable medical image analysis groups in virtually all University Medical Centers: Nijmegen, Rotterdam, Utrecht, Eindhoven, Delft, Leiden, Amsterdam, etc.

And radiologists themselves should consider making imaging informatics become a part of their training. They should have basic knowledge of machine learning and deep learning, to cope with new products from industry, and demand a prominent coordinating role in the providing and validation of the precious large-scale (ground truth) data.

Realizing the immense efficiency that the human brain is exhibiting, it is sure that we will witness many new brain-inspired discoveries. And it works both ways: also modern brain research is much helped with breakthroughs in AI. It is appropriate to end with a famous aphorism by Steve Jobs:



"I think the biggest innovations of the 21st century will be at the intersection of biology and technology. A new era is beginning."

3.11 Take Home Messages

- In Radiology, AI is here to stay. The successes are already impressive in many application areas, and a rapid expansion is seen. It finally works.
- Hire AI specialists in the larger academic Radiology departments, and incorporate imaging informatics in the radiologist's training program.
- New paradigms are needed in AI research. As the brain is very energy efficient, and works with low-frequency neurons, much inspiration is still to come from biology.
- However, the worlds of computer vision and biological vision are quite separated, and the cross-fertilization can mutually benefit.

References

1. Abbasi-Sureshjani S, Dashtbozorg B, ter Haar Romeny BM, Fleuret F. Boosted exudate segmentation in retinal images using residual nets. In: Proceedings of ophthalmic medical image analysis OMIA 2017, at MICCAI 2017, Québec City. Cham: Springer; 2017. p. 210–8
2. American College of Radiology. Data Science Institute; 2018. The ACR DSI is collaborating with radiology professionals, industry leaders, government agencies, patients, and other stakeholders to facilitate the development and implementation of artificial intelligence (AI) applications that will help radiology professionals provide improved medical care. www.acrdsi.org.
3. Blakemore C, Cooper GF. Development of the brain depends on the visual environment. Nature. 1970;228:477–8.
4. Blakemore C, Cooper GF. Development of the brain depends on the visual environment; 1970. Movie: https://www.youtube.com/watch?v=QzkMo45pcUo.
5. Briggman KL, Helmstaedter M, Denk W. Wiring specificity in the direction-selectivity circuit of the retina. Nature. 2011;471:183–8.
6. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, Kadoury S, Tang A. Deep learning: a primer for radiologists. RadioGraphics. 2017;37(7):2113–31.
7. Doolittle B, MacLay E. The forest has eyes. Seymour: Greenwich Workshop Press; 1998.
8. Eremenko K, de Ponteves H. Deep learning A-Z: hands-on artificial neural networks in Python, Udemy Inc. https://www.udemy.com/deeplearning. Most popular course 2018.
9. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. RadioGraphics. 2017;37(2):505–15.
10. Ghafoorian M, Karssemeijer N, Heskes T, van Uder IWM, de Leeuw FE, Marchiori E, van Ginneken B, Platel B. Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation. In: 13th international symposium on biomedical imaging (ISBI), April 2016. New York: IEEE; 2016. p. 1414–7.
11. Google Inc. TensorFlow; 2018. An open source machine learning framework. www.tensorflow.org.
12. Greenspan H, van Ginneken B, Summers RM. Guest editorial: deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans Med Imaging. 2016;35(5):1153–9.
13. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. J Am Med Assoc. 2016;316(22):2402–10.
14. ImageNet. Large scale visual recognition challenge (ILSVRC), 2010–2017. ILSVRC evaluates algorithms for object detection and image classification at large scale: 150000 photographs, 1000 classes. http://www.image-net.org/challenges/LSVRC/.
15. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd international conference on machine learning, Lille. Vol 37; 2015. https://arxiv.org/abs/1502.03167.

16. Kandel ER, Schwartz JH, Jessell TM. Principles of neural science. 5th ed. New York: McGraw-Hill; 2013.
17. Kolb H, Fernandez E, Nelson R, editors. Webvision: the organization of the retina and visual system. University of Utah Health Sciences Center, Salt Lake City (UT); 1995. https://www.ncbi.nlm.nih.gov/books/NBK11530/.
18. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in neural information processing systems 25. Red Hook: Curran Associates, Inc.; 2012. p. 1097–105.
19. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
20. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.
21. Masland RH. The fundamental plan of the retina. Nat Neurosci. 2001;4:877–86.
22. Nelson E. Wide-are indoor and outdoor real-time 3D SLAM; 2016. Movie: https://www.youtube.com/watch?v=08GTGfNneCI. UC Berkeley, Department of EECS/Lawrence Berkeley National Laboratory.
23. Nelson E, Corah M, Michael N. Environment model adaptation for mobile robot exploration. Auton Robot. 2018;42(2):257–72.
24. Radboud UMC, Nijmegen, the Netherlands. Diagnostic Image Analysis Group: EuSoMII workshop hands-on with AI in radiology, April 21, 2018. http/diagnijmegen.nl/index.php/Hands-on-AI, www.linkedin.com/pulse/eusomii-hands-on-ai-workshop-nijmegen-great-success-erik-r-/.
25. Rodieck RW. The first steps in seeing. Sunderland: Sinauer Associates, Inc.; 1998.
26. ter Haar Romeny BM. Front-end vision and multiscale image analysis. Computational imaging and vision series. Vol. 27. Berlin: Springer; 2003.
27. ter Haar Romeny BM, Bekkers EJ, Zhang J, Abbasi-Sureshjani S, Huang F, Duits R, Dashtbozorg B, et al. Brain-inspired algorithms for retinal image analysis. Mach Vis Appl. 2016;27(8):1117–35.
28. Wagemans J, Elder JH, Kubovy M, Palmer SE, Peterson MA, Singh M, von der Heydt R. A century of gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. Psychol Bull. 2012;138(6):1172–2012.
29. Wertheimer M. Laws of organization in perceptual forms (partial translation). In: Ellis WB, editor. A sourcebook of gestalt psychology. San Diego: Harcourt, Brace; 1938. p. 71–88.
30. Zhou SK, Greenspan H, Shen D, editors. Deep learning for medical image analysis. Cambridge: Academic; 2017.

Deep Learning and Machine Learning in Imaging: Basic Principles

4

Bradley J. Erickson

4.1 Introduction

Machine learning is a component of artificial intelligence that primarily focuses on finding patterns. In most medical imaging applications, this is further narrowed to finding patterns in imaging using a number of examples from which the algorithm figures out the pattern. While this is simple in concept, and there are some great tools that can make it simple to implement, assuring that the result is a robust and accurate system can be very difficult, with many subtle challenges along the way. By the end of this chapter, you should be familiar with some of the common machine learning algorithms, some of the ways in which these algorithms can be "fooled," and ways to make them more robust and then finish with a discussion of traditional machine learning and deep learning.

4.2 Features and Classes

As noted above, this chapter will focus on supervised machine learning, which is the type of machine learning in which known examples are used to train an algorithm to properly classify future/unseen examples into the correct classes. For example, one might have a collection of chest X-rays, some of which are known to harbor malignant nodules and others that are known to have no malignant nodules. A goal might be to develop a machine learning algorithm that will correctly identify the chest X-rays that have malignant nodules with high sensitivity and specificity.

Given a collection of labeled images (e.g., labeled as to whether they have malignant nodules or not), a first task is to compute "features" that are strong indicators of malignancy or lack thereof. Usually, more than one such feature is calculated, and the set of features computed for one example is referred to as a feature vector. It is critical to note here that there must be appropriate preprocessing of images to make the features as reproducible as possible. This typically means applying intensity and other normalization steps to the images. Exactly which features are calculated depends on the intuition and experience of the investigator. Often many features are calculated, and then a "feature reduction" or "feature selection" step is performed in which duplicative or non-informative features are removed from the feature vectors. It is important to have as few features as possible but keep as many as are needed to get the best performance.

B. J. Erickson (✉)
Department of Radiology, Mayo Clinic, Rochester, MN, USA
e-mail: bje@mayo.edu

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_4

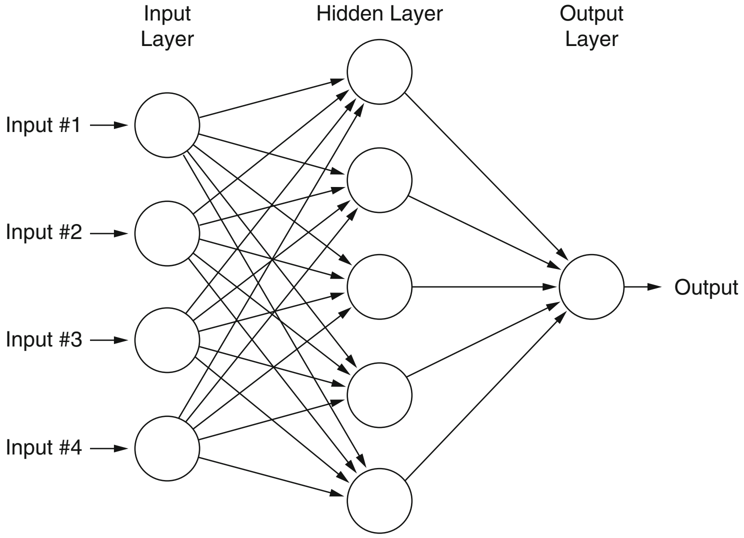


Fig. 4.1 Architectural diagram of a neural network. The input layer has as many nodes as the number of features. The values are then multiplied by weights (represented by lines connecting the input nodes to the next "layer" of nodes). These nodes receive the products of the prior layer nodes/weights, sum them up, and then apply an "activation function" to that sum, which determines the output value for this node. This continues for each layer until the final layer, which is called the output layer, where the final decision is made. The layers between the input and output layers are referred to as "hidden" layers. Traditional neural networks had 1–2 hidden layers, while current "deep" neural networks often have 10s to 100s of layers

A "class" refers to the type of finding present: in the chest X-ray example described above, there are two classes, malignant nodule(s) present and no malignant nodules present. Classes could also be used for segmentation: class 1 is the object of interest (like the pixels that constitute the liver), and class 2 is everything else. The most common classifier separates two classes, but it is possible to build classifiers that can directly classify more than two classes at a time.

Once the features have been selected, they are used as input to a machine learning algorithm. There are many different types of algorithms available, including neural nets, decision trees, support vector machines, Bayes networks, and many more, some of which are variants or combinations of these.

4.3 Neural Networks

Neural networks were perhaps the earliest form of machine learning and were based on our understanding of how the brain and its neurons work. When applied to imaging, one common approach is to have each feature that is provided to the network multiplied by a "weight" which is a floating-point number typically ranging from −1.0 to +1.0 (see Fig. 4.1). The product of each input feature value times the weight is passed to each node in the next layer. Each node in this next layer will sum these products and then apply an "activation function" that converts the input value to an output value. This output value is then multiplied by weights and passed to the next layer, and this continues until the final output layer is reached, where the decision/class is determined for the feature vector example provided. In the early days, the activation function used was often the hyperbolic tangent function (Fig. 4.2), because that approximated what was observed in biological neurons.

A critical element of the neural network was to "learn," and that was accomplished by adjusting the weights that connected the nodes. Backpropagation is the general term used for taking the error observed at the output and adjusting the weights to reduce the error for the next set of

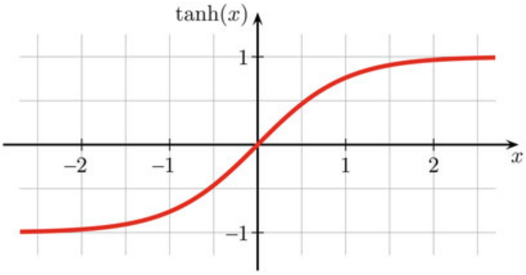


Fig. 4.2 Hyperbolic tangent function. This function approximates the activation function seen in live neurons and was heavily used in early neural networks. Nonlinearity of activation functions is an important component of effective learning

examples provided to the network. While this is simple in theory, it becomes very difficult when the network is anything beyond just a few nodes. One reason is the “vanishing gradients” problem. When an output is in error, one typically uses the amount of error or “gradient” to determine how much to change the weights. But which of the many weights that contributed to the output should be changed? If one shares the gradient across all nodes, the amount of change to any one node becomes quite small—so small it doesn’t really have an impact, and thus the system is not really learning. This led to rather poor performance of neural networks and caused them to be largely abandoned in the 1970s, and techniques other than neural networks gained more attention.

4.4 Support Vector Machines

The SVM algorithm was invented by Vladimir Vapnik and Alexey Chervonenkis in 1963 [1]. Two key concepts of the SVM are the plane that separates two classes (which is known as the “support vector”) and the challenge of mapping points from their original space to a space that allows them to be separated by a plane. The name SVM indicates that the concept of the separating plane is central to the SVM method. Since most problems cannot be solved with a simple 2D linear separator, the SVM algorithm constructs a hyperplane or set of hyperplanes in a high-dimensional space. If the data are linearly separable, it is possible to compute two hyperplanes that separate the two classes of data, so that the distance between them is maximized.

Many, if not most, cases of classification tasks are not linearly separable. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space. To keep the computational load reasonable, the mappings used by SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function that suits the problem. The hyperplanes in the higher-dimensional space are defined as the set of points whose dot product with a vector in that space is constant.

The solution to the challenge of mapping points in a way that would allow them to be separated by a plane was developed almost 30 years later, when Vapnik and others identified a way to create nonlinear classifiers by applying the kernel trick to maximum-margin hyperplanes [2]. Even with this remapping of points, it is rare to have a plane that will separate all points, and so it is necessary to allow for some points to be on the “wrong side” of the plane. The current standard incarnation to address this challenge (known as a soft margin) was proposed by Cortes and Vapnik in 1993 and published in 1995 [3]. SVMs have a parameter that specifies the amount of penalty for a point being misclassified and how much the penalty increases as a function of the distance from the plane. These parameters are also referred to as “hyperparameters” because the weights are more typically referred to as parameters. Selecting and adjusting hyperparameters are still very much an art that requires experience with both machine learning algorithms and images. It is also important to note here that while the above is largely focused on classification, SVMs can also be used for regression and outlier identification.

4.5 Decision Trees

While SVMs can be difficult to understand, decision trees are very straightforward, and that is their great appeal—it is very easy to understand how a decision tree is making its decisions. As the name implies, a decision consists of a series of decisions. For instance, one might begin with a decision like "Is the pixel value greater than 100?" All pixels greater than 100 go to one branch of the decision tree, and the others go to the other branch. By organizing a series of such binary decisions, one may accomplish very complex tasks. Of course, this simplicity also limits the power of the decision tree.

A decision tree typically is constructed to make the most "important" decisions first, and that should result in the fewest decisions having to be made. If one has a collection of examples, each with several features, the first step is to calculate the entropy of each feature across the samples, allowing us to calculate each feature's information gain. Information gain calculates the expected reduction in entropy due to sorting on the attribute. The feature with the greatest information gain is selected and the threshold that results in the best separation. The Gini Index [4] is another option for selecting features and is a measure of how often a randomly chosen element would be incorrectly identified: an attribute with lower Gini index would be the one selected.

An important weakness of the decision tree is that it will fit the data set provided, and particularly the later decisions will depend highly on the exact data given, which means it will overfit the training data. One way that people have attempted to reduce the chance of overfitting is to create "random forests." As the name implies, this involves combining many decision trees, and the way those many trees are made is by randomly pulling sets of examples out of the full data set. By doing this, the dominant features/decisions are still found in nearly all of the trees, but the last decisions, which are based on just a few examples and thus likely to be noisy and overfit, will be more variable and effectively cancel out in the forest.

4.6 Bayes Network

Bayesian networks arguably are not really machine learning but rather focus on using probabilities learned from the training data to predict the classes/outcome. "Bayes law" states that the probability of A given B (where A is the desired prediction and B is the set of input features) is equal to the probability of B given A, times the probability of A, divided by the probability of B.

$$p\,(B\,|A\,) = \frac{p\,(A\,|B\,)\;\;p(B)}{p(A)}$$

This basic formulation can be expanded to address more complex situations, such as when there are multiple known probabilities (elements of the feature vector), and more complex relationships between those features than a simple linear connection. A Bayesian network represents a set of variables and their conditional dependencies as a directed acyclic graph which is a structure in which once a node has been traversed, one can never go back to it (Fig. 4.3). For instance, given a set of features for some group of images, and the probabilities that each feature is present when disease is or is not present. Once those probabilities and connections are computed (learned), one can then compute the probability of disease in some new feature vector by passing it through the network.

In theory, the features (probabilities) should be independent, but often they are not. In addition, nonlinearities in the relationships are usually difficult to represent, and highly nonlinear situations don't perform as well. Despite these limitations, and even when these assumptions are violated, Bayesian networks can perform very well and thus are a tool that people should be aware of.

4.7 Deep Learning

Deep learning has received much attention recently because several factors came together that have enabled substantial leaps in performance. A popular contest for measuring machine learning

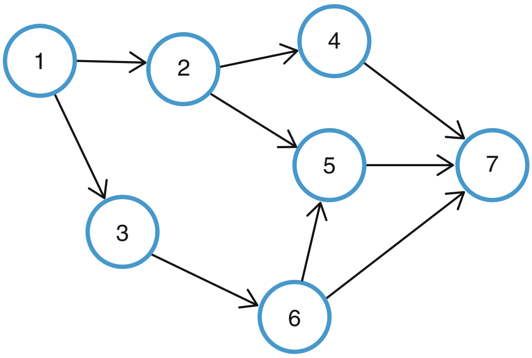


Fig. 4.3 Directed acyclic graph. This describes a series of computations that can only progress from the input to an output without any loops. For instance, in this diagram, one can only go to nodes that have higher numbers, starting with node #1 and ending with node #7

performance is the ImageNet challenge (https://www.kaggle.com/c/imagenet-object-detection-challenge), in which machine learning algorithms are given a large collection of images and labels indicating what is present in the image (e.g., "frog" or "airplane"). Traditional machine learning methods like SVMs and random forests could achieve 70–75% performance in this challenge—it was typical to see an improvement of around 1% year over year. Then in 2012, the team of Geoffrey Hinton, Ilya Sutskever, and Alex Krizhevsky from the University of Toronto submitted a deep learning algorithm called AlexNet which beat the field by more than 10%, with an error rate that was 41% better than the second-place finisher [5]. In 2013, another deep learning method resulted in another 10% gain over that. Deep learning methods now dominate this challenge, and performance is above 98% (which is better than human performance on the test set). While ImageNet has received the most attention, there are other challenges for text and speech recognition, and deep learning methods have similarly resulted in dramatic performance improvements.

4.7.1 Deep Learning Layers

Fully Connected Layers Deep learning gets its name because it uses neural networks with many layers. The traditional neural network consists of nodes and connectors that simply multiply and apply an activation function. These layers are usually referred to as "fully connected" layers and are often employed near the end of the deep learning network.

Convolutional Layers For image-based tasks, it is quite common to use several layers of convolutions at the input. The elements of the kernels are also a part of the learning process, and thus, these systems learn the features critical to successful training.

Pooling Layers Many of the modern architectures have convolutions followed by a "pooling layer," in which the outputs of adjacent convolutions are combined into a single output. The most common pooling function is the "max pool" which simply finds the maximum value for its "window" and passes that on to the next layer, which is often another convolution.

Activation Layers A key component of learning is to have nonlinearity in the system, and that is the primary function of activation layers. Early neural networks used sigmoidal-shaped functions such as hyperbolic tangent function because that is similar to what biological neurons used. However, it appears now that much simpler functions perform better for most deep learning systems. A popular activation function is the rectified linear unit or ReLU, which outputs a "0" for any negative input and outputs the input if the input is positive, thus acting like a rectifier, and hence its name. Modifications of this layer include leaky ReLU, Gaussian ReLUs, and exponential ReLUs in which some nonzero output is used for a negative input.

Output Layer The final output layer is a special case of an activation function, and for that, more sophisticated layer types are often used. If the task is regression (e.g., estimating the age of the patient based on a hand X-ray), a linear output (e.g., a floating-point value rather than a binary "yes" vs. "no") is appropriate. If the task is classification (e.g., the lesion is a cancer rather than benign), the softmax function often performs well. The softmax function will take a vector of values (from the prior layer) and convert them to an arbitrarily sized output vector (the number of possible classes), and the sum of all the output values is one. If the output allows for multiple binaries (e.g., multiple objects present), then a simple sigmoid can work. The cost function used is also an important factor in choosing the output layer.

Residual Layer There are some additional layer types that are being developed and applied to great advantage. One recent example is the "residual layer" which gets its name because it uses a "bypass" layer that is essentially the identity function, and then the output of a layer or group of layers is compared with that identity function. This effectively forces the non-bypass layers to do better than an identity function and can therefore learn more effectively with fewer layers. This is important because the reduction in layers both reduces the number of potential parameters to adjust when learning and also reduces the chance of overfitting to the training data.

4.7.2 Deep Learning Architectures

Deep learning systems can include many different types of layers, in various sequences, each layer has a number of parameters, such as how many nodes or other layer-specific configurations such as size of a convolution kernel or size of the pooling window. The selection and arrangement of these layers are referred to as the architecture of the deep learning system.

Convolutional neural networks (CNNs) are a common deep learning architecture for images, particularly for image classification tasks. The series of convolutions and max pooling layers at the start effectively find low-level (high-resolution) features in the images. As the max pools reduce the resolution, lower-level features are combined into higher-level features that represent more complex objects. Thus, the first layer(s) may find things like points, lines, and edges. These are combined by later layers to identify boats or cars or faces. It is common to employ fully connected networks just before the final, so that the high-level features are weighted to determine which type of object is present in an image. The original AlexNet [5], VGGNet [6], and GoogLeNet [7] are all examples of CNNs. More recent variants of CNNs with specialized layers include ResNet [8], ResNeXt [9], and region-based CNN [10].

U-Nets [11] are a special form of CNN that get its name from the appearance of the architecture diagram (Fig. 4.4). The key element of U-Net is that at the bottom of the "U," the image is reduced to the key component that one is looking for. Once that key component is recognized, the refinement to original resolution is achieved with "bypass layers" in which the upscaling uses pixel data from the higher-resolution versions to refine the key component until the original resolution is achieved. SegNet [12] is one specific example of a U-Net.

Fully connected networks (FCNs) using traditional backpropagation can not only be used as a part of a more complex architecture but may also be used as the only type of layer for an entire network [13]. Essentially, these are the original neural network architectures; though in the deep learning space, these typically have many more layers and nodes. The advantage of FCNs are that they are very general—they can be applied to images, 1D signals, text, and in fact, pretty much any type of input. The challenge is that the generality usually requires more training data for the system to get good results. FCNs also naturally enable mixed data types—e.g., where you want to perform classification on images but want to include other data like age or gender or blood tests as additional input. This can be done with other network architectures but is completely natural to an FCN.

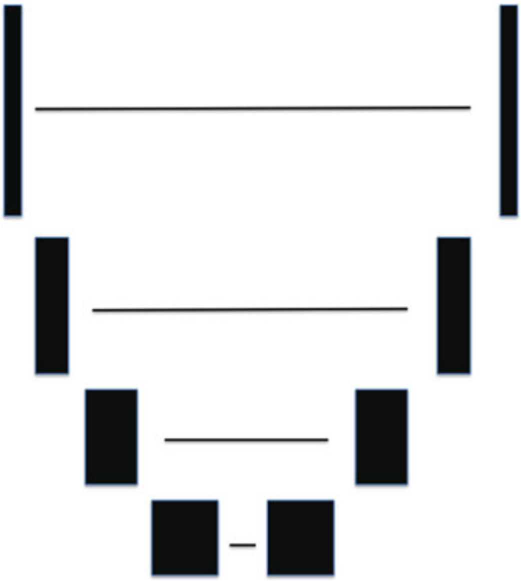


Fig. 4.4 Architectural diagram of U-Net. At the upper left, the full resolution is provided to the network. The next layer is reduced spatial resolution (the next box down and to the right). This reduction in resolution continues, while the larger features of the object being segmented become more apparent. At the bottom, the lowest-resolution image is found, and then this is fed to networks that then refine that low-resolution version (boxes up and to the right). The lines connecting the boxes to the left represent the component where the higher-resolution images are used until one is back to the original full resolution

Generalized adversarial networks (GANs) are a very different type of network that are designed to *create* images rather than classify or segment them [14]. This may seem like a useless function in radiology, as image devices like CT or MR scanners are usually the only source of images. However, GANs have gained attention in the radiology world (and throughout the machine learning field) because they can create images that look very real. This can be useful in medicine for a few purposes, including the creation of additional training and testing images, and they may be useful for providing insight into both how deep learning works and also how disease pathology might be better detected with imaging devices. Early work on GANs focused on creating images that would fool deep learning systems. One famous example was where the addition of carefully crafted noise was added to a picture of a panda, and the addition of this noise caused the image classifier to change from a correct classification ("panda") to an incorrect classification ("gibbon"). This caused quite a stir in the imaging community and did raise awareness of the potential for these systems to fail.

Such networks have also proven useful for improving the robustness of deep learning systems. Because GANs can identify the weakest points (i.e., the ways that the system can be most easily mistaken), they can help developers to make the network more robust. The specific aspects that make the system can be "fortified" to make them more resistant to real-world examples that may be similar to the GAN examples and thus make the system perform better in the real world.

4.8 Conclusion

While machine learning has been applied to medical images for decades, recent success in the application of deep learning methods to medical images has resulted in great enthusiasm about its potential. Furthermore, our understanding of how these systems work is still undergoing rapid development, and it is likely that this will continue. Today, it is expected that radiologists understand how CT and MR scanners work, even though almost none will ever design or build one. It is considered an important knowledge because it provides insight into how disease might be manifest and also may help to discern disease from artifact. In that same way, deep learning methods almost certainly will become a required part of training for those that utilize medical imaging, because it is critical that physicians understand how to best apply these methods in current clinical practice in order to avoid errant use.

References

1. Vapnik V. Pattern recognition using generalized portrait method. Autom Remote Control. 1963;24:774–80.
2. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceed-

ings of the fifth annual workshop on Computational learning theory – COLT '92. 1992. https://doi.org/10.1145/130385.130401.

3. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
4. Gini C. Variabilita e Mutabilita. J R Stat Soc. 1913;76:326.
5. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, CJC B, Bottou L, Weinberger KQ, editors. Advances in neural information processing systems 25. Red Hook, NY: Curran Associates; 2012. p. 1097–105.
6. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv [cs.CV].
7. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Computer vision and pattern recognition (CVPR). 2015. http://arxiv.org/abs/1409.4842
8. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Bajcsy, editor. Proceedings of the IEEE conference on computer visions and pattern recognition. Los Alamitos, CA: Conference Publishing Services; 2016. p. 770–8.
9. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. 2016. arXiv [cs.CV].
10. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. 2015. arXiv [cs.CV].
11. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical image computing and computer-assisted intervention – MICCAI 2015. Cham: Springer International Publishing; 2015. p. 234–41.
12. Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. 2015. arXiv [cs.CV].
13. LeCun Y, Boser BE, Denker JS, Henderson D, Howard RE, Hubbard WE, Jackel LD. Handwritten digit recognition with a back-propagation network. In: Touretzky DS, editor. Advances in neural information processing systems 2. San Mateo, CA: Morgan-Kaufmann; 1990. p. 396–404.
14. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, editors. Advances in neural information processing systems 27. Red Hook, NY: Curran Associates; 2014. p. 2672–80.

Part III

Technology: Developing A.I. Applications

5 How to Develop Artificial Intelligence Applications

Angel Alberich-Bayarri, Ana Jiménez Pastor,
Rafael López González, and Fabio García Castro

5.1 Introduction

Artificial Intelligence (AI) is one of the most flourishing topics in many aspects of industry, science and technology, being pointed as the main driver of the fourth industrial revolution. Among the multiple progress that data science and AI have introduced in daily-life applications, speech and face recognition, self-driving cars and natural language processing must be highlighted.

Beyond these applications, if there is a field where AI is introducing disruptive innovations, it is healthcare, where doctors have to handle a large set of information in every clinical episode. The increased computational capabilities, thanks to the progressive growth in the performance of graphics processing units (GPU), combined with the potential for pattern recognition of deep artificial neural networks (ANN), have allowed for the management of huge amounts of data with an efficiency that was not possible a decade ago.

In the field of radiology, the main inflexion point in the success of AI was the growing success of a specific type of ANN called convolutional neural networks (CNN) that are specifically suitable to analyse unstructured bidimensional information like images. The main difference in comparison with conventional ANN is that the hidden layers of these architectures are formed by convolutional layers that apply a filter (convolution) to the input data. The advantage over using traditional network architectures is that the convolution allows for a significant reduction of the number of free parameters of the network and does not require for the previous extraction of hand-crafted features. These networks that in the field of radiology are named as deep learning (DL) (although deep learning is also used in non-convolutional architectures) were mainly proposed for real applications by Yann Lecun in 1998, with the introduction of the network architecture LeNet-5 [1] that was used to recognize handwritten digits in bank cheques. However, the lack of computational capabilities limited the application of CNN to images with a few number of pixels (i.e. 32×32), far from the high spatial resolutions and matrix dimensions used in radiology and in digital photography. It was not until 2012 when in the annual ImageNet Large Scale Visual Recognition

A. Alberich-Bayarri (✉)
Biomedical Imaging Research Group (GIBI2^30), La Fe Health Research Institute, Valencia, Spain

Quantitative Imaging Biomarkers in Medicine (QUIBIM S.L.), Valencia, Spain
e-mail: alberich_ang@gva.es

A. J. Pastor · R. L. González · F. G. Castro
Quantitative Imaging Biomarkers in Medicine (QUIBIM S.L.), Valencia, Spain
e-mail: anajimenez@quibim.com; fabiogarcia@quibim.com

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_5

Challenge (ILSVRC), a novel CNN architecture outperformed other image processing algorithms by decreasing classification error rate from 25% (2011) to 15%. Although the concepts and the building blocks were not new by themselves, their combination resulted in a highly effective network called AlexNet named after the first of the three authors of the work: Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton [2]; moreover this network showed the benefits of training a CNN by using GPUs. Their work has had a very important impact in the computer vision community, with more than 4000 sites per year in the 5 years following the publication, and opened the door to the application of CNN for the analysis and classification of radiological images. Deep learning is currently in the top of the hype cycle for emerging technologies [3]. Even more, it has been the source of concern about the future of the profession, and AI and specifically the domain of machine learning (ML) in radiology became one of the main trends of the two most important annual conferences: the annual meeting of the Radiological Society of North America (RSNA) and the European Society of Radiology (ESR). These algorithms, however, go beyond CNN architectures [4], since ML is a complete knowledge domain with different techniques designed to learn patterns from data. All of them have demonstrated to be highly specific, being useful to solve repetitive and rule-driven problems without clinical context with human-like performance, and must be understood more as a complement than a substitute of the radiologist. The quantity and heterogeneity of information to be evaluated by radiologists' mind during the image interpretation process are high. Radiology is not only about image recognition but a high amount of contextual information (patient characteristics and habits, clinical status, previous clinical episodes, lab test results, previous examinations, imaging findings, among others) [5]. A future value-driven integration of AI in radiology is expected where technology, really solving unmet clinical needs, will reach clinical practice implementation to reduce the interpretation times needed for complex imaging studies and to diminish the number of repetitive time-consuming tasks, which will therefore improve the workflow efficiency. Although it remains a matter of discussion due to the lack of current legal coverage, a shared responsibility scheme between radiologists and AI software vendors, regarding potential mistakes in image interpretations due to improper performance of the algorithm, is envisaged.

In this chapter, the main applications of AI in radiology together with a methodology on how to implement them in clinical routine will be introduced, demystifying all the hype surrounding the technology and showing how we can get the highest value from it to improve radiologists' daily workflows.

5.2 Applications of AI in Radiology

The applications of AI in radiology go quite far beyond the intuitive use for automating image interpretation [6], with functions in image acquisition, management and population imaging that will probably be more prolific in the coming years due to the value that would be provided in optimizing daily practice workflows.

- *Image acquisition*:
  - *Creation of study protocols*: the creation of some patient-specific acquisition protocols largely depends on the clinical indications for the imaging procedure. A significant amount of data from the patient is taken into consideration at the same time, including results of other diagnostics, such as blood tests and previous examinations performed. Even more, the examination duration may be different due to specific image series to be acquired. AI will allow for creation of ad hoc protocols in special situations by taking into account current disease guidelines, differential diagnosis and the required data such as previous image acquisitions and lab test results. Furthermore, this new protocol development task could be synchronized with the scheduling in the agenda of the machine to

allocate the appropriate image acquisition time.

- *Optimized MR and CT image quality*: MR machines include methods to shorten times and improve signal homogeneity; likewise, CT machines include image filters and radiation dose reduction functionalities like dose modulation. However, these algorithms and models have been mainly hand-crafted, not evaluating the optimum parameter configuration in every patient to guarantee a maximum image quality by the continuous monitoring of spatial resolution, contrast and signal-to-noise ratio (SNR). AI will help to automatically extract image quality indicators as they are generated and store relationships between protocol parameters and quality to train new algorithms of optimization.
- *Assessment of image quality*: the continuous assessment of image quality within a radiology department would allow for the detection of potential anomalies in the scanners even if they are occasional. AI will help both in the automated quality assurance tests performed phantom-less using patients data and in the application of algorithms for the detection of 'abnormal behaviours' of image quality in a specific machine using methods similar to those being used in the banking sector to detect suspicious or fraudulent operations.

- *Image interpretation*:
  - *Automated hanging protocols*: radiologists dedicate an enormous amount of time to organize the images in the interpretation process. Although most picture archiving and communication system (PACS) solutions include the option to configure different profiles, certain intelligence would help to arrange series and images according to the indications for the examination. AI will help to not only load the most relevant series but also going to the slices in the specific organ or region anatomy relevant from the clinical data.
  - *Radiomics and imaging biomarkers analysis*: in the last decades, several models and techniques have been proposed for the extraction of imaging biomarkers from tissues and organs, with applications in diffuse diseases [i.e. Alzheimer, steatosis, chronic obstructive pulmonary disease (COPD)] and in the characterization of focal lesions (i.e. lesions in cancer). These advances suppose the most important breakthrough in image analysis. Nevertheless, quantification is still not integrated in clinical routine with well-known normality ranges, disease values as we do in blood test biomarkers. AI will significantly help the field of imaging biomarkers in two different steps: segmentation and data mining.

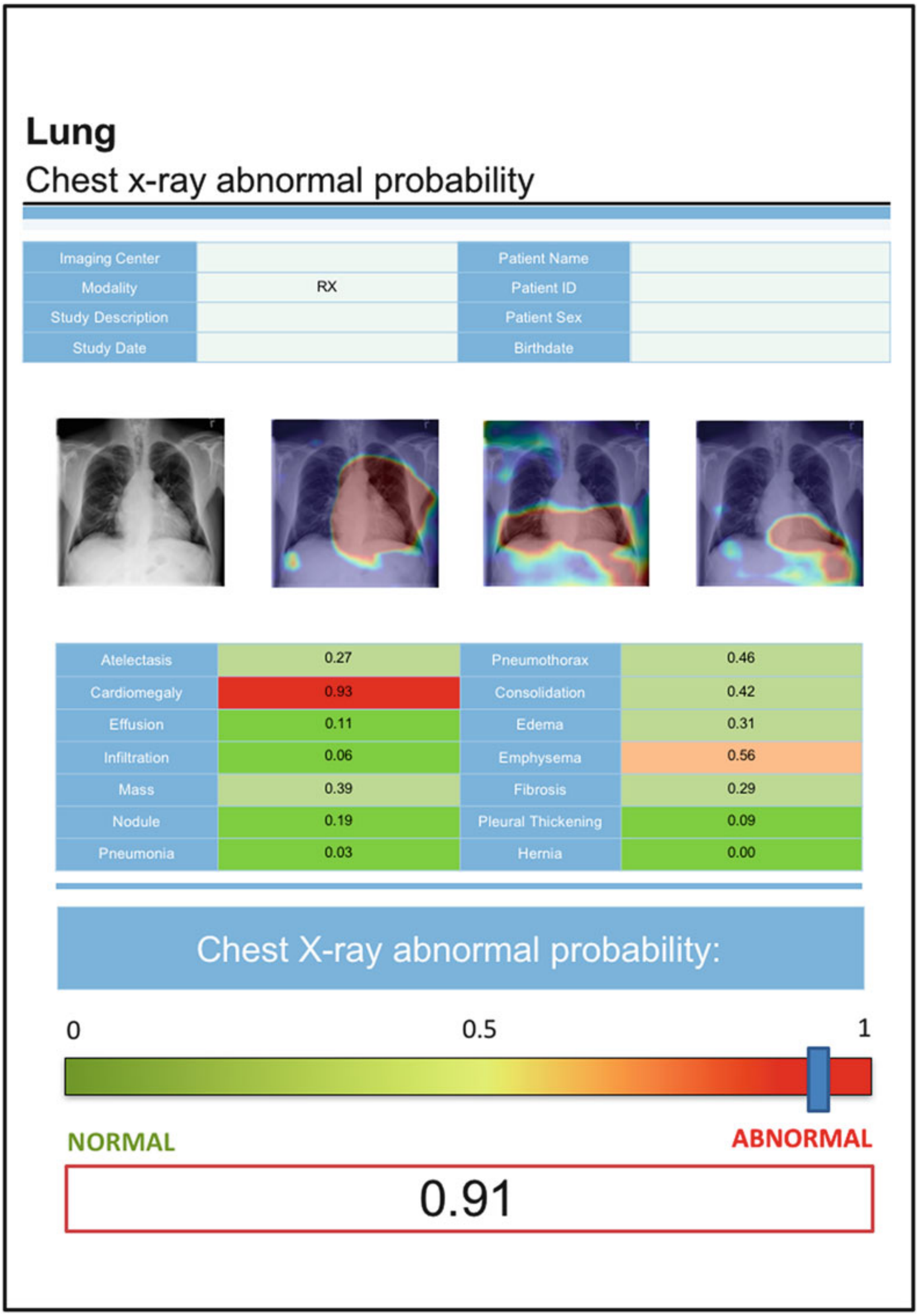
    One of the main steps of the analysis requiring human interaction is segmentation of organs, lesions or selection of specific areas as regions of interest (ROI) like the arterial input function (AIF) in perfusion studies. ROI selection is one of the steps hindering a complete integration of quantitative analysis tools within clinical routine. AI is allowing for highly accurate segmentations compared to human performance. The use of CNN and more specifically network architectures based on compression-expansion and designed for segmentation, like U-Net or V-Net (U and V letters are given by the shape of the architecture of the network in the first case, U-shape, and by the use of 3D volumes in the second, respectively), is providing DICE scores (indicator that ranges from 0 to 1 and measures the degree of overlap and accuracy between ground truth delineated by the expert and the computed segmentation) near to 1.

    In the field of radiomics, there is a need to introduce ML techniques to process all the quantitative information generated beyond basic descriptive statistics and to extract the relationship of biomarkers with clinical endpoints. For that purpose, methods for automated clustering of patient 'radiomics signatures' or 'imaging fingerprints' are applied, allowing for the detection of image-based phenotypes that might be related to diagnosis, prognosis or treatment response. This process allows for the translation from population-derived knowledge to the application in a single patient. For example, in patients with rectal cancer, it would be possible to detect with ML that a combination of specific texture features of the cancer is prone to a higher chance of recurrence. These conclusions can be translated to a new patient and evaluate the prognosis to relapse depending on the radiomics features.

  - *Automated image interpretation*: if there is an application that has raised the concern among the radiological community, this is image interpretation and the detection of findings. The excellent performance of CNN allows for training new algorithms able to classify studies according to a huge amount of image features that the network is able to extract. The main limitation for the generation of such classifiers is the lack of annotated studies. In fact, this application requires huge amounts of labelled data to achieve good performance. Also, the networks can be trained to solve specific problems but lack in management of contextual information. Image interpretation is more than reading images but putting together all the information of the patient to achieve a proper diagnosis or clinical decision. Figure 5.1 shows an example of an output of a chest X-ray classifier developed with deep learning as a first read to prioritize relevant examinations and improve radiologist workflow.

- *Reporting*:
  - *Speech recognition*: although the field of speech recognition has evolved significantly with the application of deep learning technology for daily-life vocabulary and applications, speech recognition systems in radiology are mostly based on traditional hidden Markov models and dynamic time warping (DTW) that consist of handcrafted algorithms and are outperformed by AI data-driven algorithms like recurrent neural networks (RNN) [7]. AI will help to minimize the error rate in transcription of radiology reports assisted by speech recognition, special thanks to the better performance of RNN for outlier speech signals, where DTW fails. In any case, the field of speech recognition in radiology will dramatically change with the adoption of structured reporting on a routine basis, since all the information will be structured, diminishing the probability of error in matching the speech with the multiple-choice questions about findings. 'Ergonomic' software, with structured reporting forms that can be filled in an agile way by typing and using keyboard shortcuts, would also be the end of the field of speech recognition in radiology.
  - *Text translation*: reports translation to different languages or to different types of reports would be a major step benefiting patients that may want to share the reports with doctors from any country. AI will allow to have an ontology-based electronics health record (EHR) with the items translated to several languages. A specific functionality of AI will be the synthesis of natural text to resemble human reports.
  - *Automated annotation through keywords*: one of the main limitations of the development of AI solutions for image interpretation and classification is the lack of properly labelled or annotated studies. The automated conversion of clinical data and radiology reports into keywords from MeSH (Medical Subject Headings) to RadLex (Radiology Lexicon) dictionaries will allow to seamlessly label the examinations and make them useful for training new AI algorithms for image interpretation.

- *Knowledge extraction through data exploitation*:
  - *Processing radiology reports*: although there is a willingness of the radiological community to implement structured reporting solutions in daily routine, most of the centres still report the studies on prose text. For this reason, techniques like natural language processing (NLP) can



Lung

Chest x-ray abnormal probability

Imaging Center		Patient Name	
Modality	RX	Patient ID	
Study Description		Patient Sex	
Study Date		Birthdate	

Atelectasis	0.27	Pneumothorax	0.46
Cardiomegaly	0.93	Consolidation	0.42
Effusion	0.11	Edema	0.31
Infiltration	0.06	Emphysema	0.56
Mass	0.39	Fibrosis	0.29
Nodule	0.19	Pleural Thickening	0.09
Pneumonia	0.03	Hernia	0.00

Chest X-ray abnormal probability:

NORMAL ABNORMAL

0.91

Fig. 5.1 Example of a report at the output of a chest X-ray classifier algorithm based on deep learning through the use of CNN. It can be seen that an abnormality score is provided, together with the probability of having different typical findings. The algorithm is embedded in QUIBIM Precision® platform (QUIBIM SL, Valencia, Spain), accessible through cloud, and allows for the classification within 14 different types of findings and was trained with 112,120 annotated chest X-rays

be applied for the automated extraction of semantic information from free-text radiology reports already stored in radiology information systems (RIS), PACS and EHR. This method allows for automated annotation of cases by keywords extracted after the NLP application. One of the main challenges in the application of these algorithms is the proper detection of negations that have reported a sensitivity of 77.8% and a specificity of 94.5%, respectively [8].

  - *Image-based search engines*: although this has become quite effective for daily life images in search engines such as Google, the possibility of searching similar reference cases with radiological images is also a field of development and new start-up companies are appearing that offer this kind of product. This could be used to train young radiologists and to assess radiologists in the diagnosis process.
  - *Population health*: initiatives like the Euro-BioImaging project (http://www.eurobioimaging.eu/) and the publication of the position paper on imaging biobanks by the European Society of Radiology (ESR) [9] have fostered the creation of software platforms that allow the storage, analysis and annotation of images with associated clinical and context data. These biorepositories will allow for the application of AI algorithms in order to extract information about the relationship between image features and clinical endpoints (i.e. overall survival, time to progression, disease free survival, among others).

- *Management*:

  One of the most straightforward applications of AI is in the improvement of current business intelligence platforms for the management of hospital departments such as radiology. In the field of management in radiology, AI will allow for the optimization of imaging equipment utilization and appropriate scheduling of staff and examinations, tedious activities driven by specific rules that are a perfect problem to be solved by AI capability of pattern extraction.

5.3 Development of AI Applications in Radiology

Clinical Problem Definition As in any field of biomedical engineering, the golden rule for the development of successful and useful technological solutions is to clearly detect a clinical need and allocate time to achieve a requirements definition as detailed and specific as possible. For example, in the chest X-ray classifier shown before, the radiologist's immediate need is not an algorithm performing X-ray diagnosis but a method to rule out abnormal studies and prioritize them in the worklist, minimizing the number of unreported exams, which is a worldwide issue.

Engineering the AI Technology Once the clinical need is clear enough, the AI technological solution must be engineered, since some decisions will determine the type of ML technology or implementation to be used. In this step we have to choose between creating a classifier and a regressor. The output of the classifier will be a group of categories (i.e. normal vs. abnormal chest X-ray), while the output of the regressor will consist of continuous values (i.e. X, Y, Z positions of a ROI). The classifier or regressor technique to be evaluated among all the available models in ML will be specified [10]. At this point, it is also critical to evaluate the data collection and annotation procedures (i.e. the number of samples required or the labels needed to achieve an optimal performance) as well as the hardware requirements.

Dataset Collection After the proper design of the AI technology, the dataset of imaging studies should be collected; in this step it is essential to preserve the data format across all the files of the dataset in order to have consistency through the different image files.

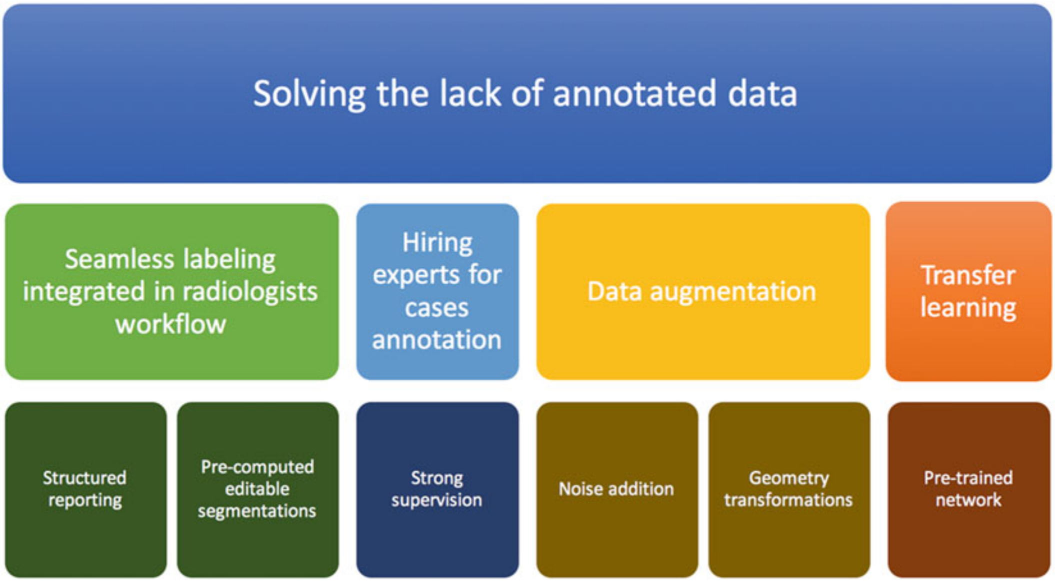

Solving the lack of annotated data
Seamless labeling integrated in radiologists workflow
Hiring experts for cases annotation
Data augmentation
Transfer learning
Structured reporting
Pre-computed editable segmentations
Strong supervision
Noise addition
Geometry transformations
Pre-trained network

Fig. 5.2 Strategies to overcome the lack of data annotation in radiology

Data Annotation The next step is the data annotation, and it is one of the most relevant parts of the whole ML development process. Annotated data has been considered with analogies such as the new gold or the new oil. As in other fields, data annotation in radiology is a big problem, since PACS systems are mainly an archive of the images and have not been designed to properly manage case annotations [5]. Due to these reasons, different strategies have been proposed to overcome the annotation deficit (see Fig. 5.2).

The best approach to facilitate the annotation of cases would be the creation of more AI-friendly PACS environments, which would allow radiologists to seamlessly label the cases. This, however, has a main drawback as it will only work for prospective imaging studies. In the case of image annotation with semantic information, it can be achieved by the integration of structured reporting in daily routine, where the case will automatically be linked to a set of predefined fields in the database containing the findings. With regard to graphical annotations, such as regions of interest (ROI) definitions, a seamless annotation would be achieved if the radiologist gets precomputed editable ROI, which can be easily modified and stored as a ROI validated by the expert. As an example, approximate ROI for focal lesions in the liver could be precomputed (as soon as the images are received in the PACS after the generation in the corresponding modality), and the radiologist would modify them to better adjust to lesion contours. After editing it, this ROI could be stored as an expert annotation that might be used later for training new AI algorithms.

However, in order to take advantage of all the information that already exists in imaging repositories and PACS, there is a need to perform retrospective annotations. For this, a significant number of start-ups and AI companies are hiring radiologists for annotating images. Imaging centres and radiology departments have also started to offer the service of annotating cases in a similar way to the Mechanical Turk services in Amazon for labelling daily life images (Amazon, WA, USA). This expert annotation is considered as a technique of ‘strong supervision’.

Annotation can be also performed by ‘weak supervision’ strategies, which allow the labelling of large databases in a cost-effective manner but decreasing the annotation accuracy. An example of these strategies is the use of NLP techniques to automatically extract labels from chest X-ray reports.

When the labelled data available is scarce, techniques like data augmentation can be used to artificially generate new images that can be combined with the original ones to enrich the variability of the dataset. The main strategies for data augmentation are noise addition and geometry transformation of the images (rotation, translation, zoom, etc.). In the case of ANNs, a technique that has shown excellent performance creating new data is generative adversarial networks (GAN) [10], which consist of a deconvolutional network acting as a generator (i.e. artificially generating new images) and a CNN that evaluates the degree of similarity between the generated and the real data. Backpropagation is an ANN training method that can be applied in both networks making the generator produce better images, while the discriminator becomes more skilled at labelling synthetic images [11].

A common technique used to minimize the lack of annotated cases in radiology is transfer learning, which consists of using pretrained ANN models in other domains (i.e. daily-life images) that must be retrained for the desired application in radiology. Transfer learning allows to reduce the training time to achieve good performance, since the weights of the network have not to be tuned from the beginning.

Training The training phase is one of the most relevant steps in AI. Even a statistical analysis should be performed to split the dataset into training and testing; typically, 80% of the cases in the whole dataset are used for training the AI models. Both input data and the corresponding labels or annotations will be used to update iteratively the weights of the model. In the case of CNN, not only the network weights but also the filter parameters will be tuned in order to increase the performance on the training data. Once the training is finished, all these weights and parameters are fixed and will remain unmodified along the next steps of the AI pipeline. The algorithm most frequently used in the training process is gradient descent, which is used to calculate the minimum of the cost function, that is, the difference between the obtained output and the expected one. The iteration in which the full dataset is passed forwards and backwards through an ANN is called an ‘epoch’. Since the whole training dataset cannot be passed all at once to the network, it is divided in batches. Therefore, in each epoch, all the different batches in which the training data is divided are evaluated sequentially by the network. Since in each epoch, the network must evaluate the entire dataset divided in batches, the training step can take a lot of computing time. However, this task is very suitable to be parallelized; it is advisable to run it over one or multiple GPUs, which carry out parallelization task orders of magnitude faster than regular microprocessors.

Testing Once the training is performed, new unseen images are used to evaluate the robustness and generalization of the final model. This step is much faster and requires less computing power since the test images are just evaluated once by means of a forward pass through the network. This is because the model weights are just updated and fixed during the training; therefore, there is no need of an iterative process afterwards. The model evaluation is carried out in different ways depending on the domain of the problem, for example, when dealing with a classification problem, the area under the curve (AUC) is broadly used in bibliography, or when facing a segmentation task, the Dice coefficient (DC) is a common approach.

In Fig. 5.3, a schematic summary of the training and testing processes for a CNN can be appreciated, where it can be observed that the training process requires high-performance computing requirements and training times in the order of hours, days or even months, while after the CNN has been properly tuned, the testing process can be executed even in consumer devices, and it takes seconds to give a result.

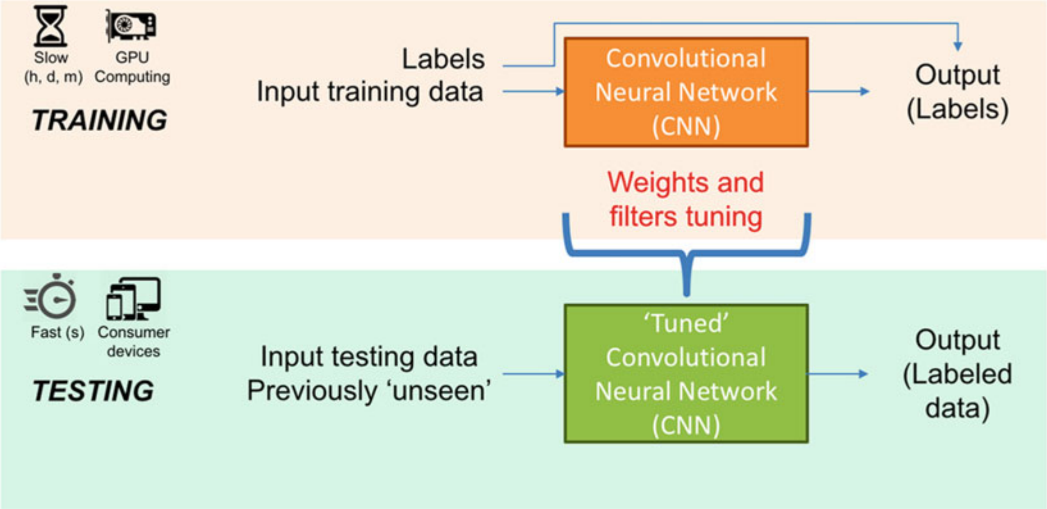

Slow (h, d, m)
GPU Computing
TRAINING
Labels
Input training data
Convolutional Neural Network (CNN)
Output (Labels)
Weights and filters tuning
Fast (s)
Consumer devices
TESTING
Input testing data Previously 'unseen'
'Tuned' Convolutional Neural Network (CNN)
Output (Labeled data)

Fig. 5.3 Summary of training test process in a CNN

5.4 Resources Framework

For the development of AI applications within radiology departments or medical imaging research groups, there is a need for a paradigm shift in the processes and in the professionals involved in the data workflows. From the authors' perspective, the future of radiology is linked to the inclusion of data scientists within radiology departments. These specialists must support their work with the appropriate hardware resources for high-performance computing and software libraries to develop new algorithms. However, these three components (data scientists, hardware and software) have to be fed with the appropriate labelled data. An example of the four key pieces to be incorporated to current radiology departments for the development of innovative AI applications can be seen in Fig. 5.4.

Expertise Dealing with AI computing infrastructure and processing algorithms is a challenging task that requires very specific profiles with knowledge in fields such as computer science, statistics, mathematics, image processing, machine learning, etc. This figure is currently known as data scientist.

When trying to solve an AI problem in the clinical domain, it is very important to reach a synergy between medical experts and data scientist in order to develop solutions that satisfy clinical needs in the most efficient way.

Computing Resources In most cases, medical images are large files (i.e. a whole-body CT scan is common to find volume sizes of 512×512×1024); therefore, processing these images is a challenging task which requires powerful hardware. In addition, AI techniques are high computing demandant itself. Hence, when using AI algorithms on medical images, the computing requirements increase. To deal with this need, it is recommended the use of GPUs to accelerate the calculations performed by the DL algorithms.

Software Resources In the last years, many advances in deep learning libraries have been accomplished. This progress has brought two important changes. On one hand, the new libraries ease significantly the development of DL architectures, allowing data scientists to develop models much faster, permitting the broadening of AI borders. On the other hand, these frameworks make seamless the shift from research to production environments. The programming language

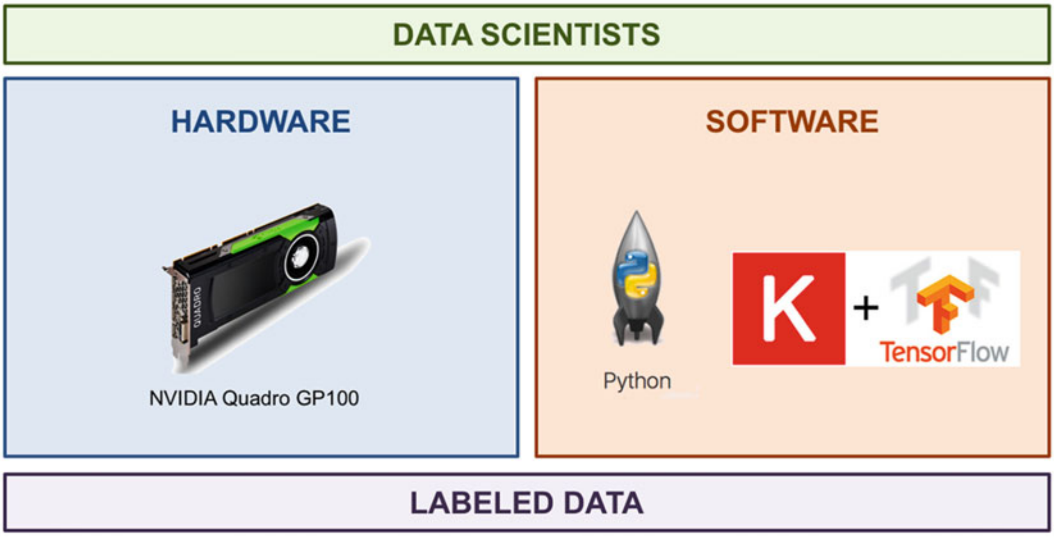


Fig. 5.4 Components required for the development of AI applications within radiology departments

most commonly used by the AI community is Python (Python Software Foundation) because the best DL libraries have been developed for this language (i.e. Theano, Tensorflow, Keras, Pytorch, MXNet, etc.).

Data Resources To develop AI algorithms, annotated datasets are required. The quality of the datasets has a great impact in the performance of the AI models. To ensure the quality of the dataset, it is important to keep a consistent structure between the collected samples. It is also fundamental that the inner variability of the dataset represents faithfully the whole population.

5.5 Conclusion

In this chapter, the main areas related with medical imaging where AI tools can have a large impact in the future have been presented. These tools can be used along the whole radiological workflow, going from the acquisition of the image to the reporting. A stepwise approach for the development of an AI model was also introduced, in which the first step has a high relevance, with the careful evaluation of the clinical and technical needs. The influence of these needs in the design of the data collection and annotation processes was also reviewed. It has also been introduced the two main steps in the model development which are training and testing. During training, the model parameters are learned from labelled data by means of an iterative process where lots of operations are performed; therefore, a high computing infrastructure is required. Finally, the resources needed to implement a successful solution were detailed, which are engineering and radiological expertise, computing and software infrastructures and large high-quality datasets. All this AI knowledge must be embraced by the radiological community in order to obtain efficient applications that allow to improve their work, not considering the technology as a threat but as the main driver of opportunities for the future specialists.

5.6 Summary/Take-Home Points

- AI is permeating all aspects of medical imaging, from image quality in acquisition, automated image classification, quantitative image analysis, reporting and management.
- AI developments have demonstrated to be highly specific, being useful to solve repetitive

and rule-driven problems without clinical context with human-like performance, and must be understood more as a complement than a substitute of the radiologist.

- Data scientist, appropriate hardware for computing, software tools and labelled imaging cases are the elements needed for successful development and implementation of AI algorithms with an impact in radiology.

References

1. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86:2278–324.
2. Krizhevsky A, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012:1097-1105.
3. Hype cycle for emerging technologies 2017. 15-08-2017. www.gartner.com. Accessed 1 May 2018.
4. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. Radiographics. 2017:160130–11.
5. Alberich-Bayarri A. Image interpretation. In: Medical radiology. Berlin: Springer; 2017.
6. Lakhani P, Prater AB, Hutson RK, et al. Machine learning in radiology: applications beyond image interpretation. J Am Coll Radiol. 2017:1–10.
7. Fohr D, Mella O, Illina I. New paradigm in speech recognition: deep neural networks. In: IEEE international conference on information systems and economic intelligence, Apr 2017, Marrakech, Morocco. 2017.
8. Cai T, Giannopoulos AA, Yu S, et al. Natural language processing technologies in radiology research and applications. Radiographics. 2016;36:176–91.
9. European Society of Radiology (ESR). ESR position paper on imaging biobanks. Insights Imaging. 2015;6:403–10.
10. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. 2014. arXiv:1406.2661.
11. Thaler SL. US Patent, 07454388, Device for the autonomous bootstrapping of useful information, 11/18/2008.

6 A Standardised Approach for Preparing Imaging Data for Machine Learning Tasks in Radiology

Hugh Harvey and Ben Glocker

6.1 Data, Data Everywhere?

The traditional paradigm of hypothesis-driven medical research largely rests on clinical studies involving cohorts of a few hundred or thousand patients. However, modern machine learning techniques benefit from exponentially larger volumes of data. It is often asked 'how much data is required to build an algorithm?'; a question which has no straight answer. Indeed, when dealing with neural networks and systems that are required to function accurately across numerous possible clinical scenarios, including rare conditions, it is difficult to calculate a suitably statistically powered number on which to rely. This is due to the need for thousands of examples per 'class' or entity being solved by algorithms that estimate complex, non-linear relationships between input and desired output. In short, the greater the number of classes (or conditions) to predicted, the more data is required.

The term 'big data' has been used since the 1990s to describe volumes of digital data in excess of those required for traditional scientific research. Globally, it is estimated that 2.5 quintillion bytes of digital data is produced every day, 90% of which are unstructured [1]. In radiology alone, exobytes of data are produced each year, with an ever increasing production velocity. It is clear that we are now surrounded by healthcare related data; however, the barriers to accessing it and harnessing it prevent us from utilising 'big data' to its maximum potential. In a world overflowing with radiological data, medical imaging researchers are paradoxically data starved.

It has been shown that algorithmic performance on computer vision tasks increases logarithmically based on volume of training data size [2, 3]. However, publicly available medical imaging data for algorithmic training has not significantly increased in size over the past few years, instead largely being held privately by hospitals, research units and industry in silos. While there are some publicly available imaging datasets available, these are not increasing in the orders-of-magnitude required for the machine learning sector to exploit them beyond narrow, and very specific tasks. Additionally, if everyone is training and validating on the same data, there is a danger of overfitting to what is available, and risking dangerous suboptimal performance when algorithms are introduced into the clinical wild. This relative plateauing of available training data is in stark contrast to the explosive increase in the volume of medical imaging data globally.

H. Harvey (✉)
Kheiron Medical Technologies, London, UK
e-mail: hugh@kheironmed.com

B. Glocker
Imperial College, London, UK

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_6

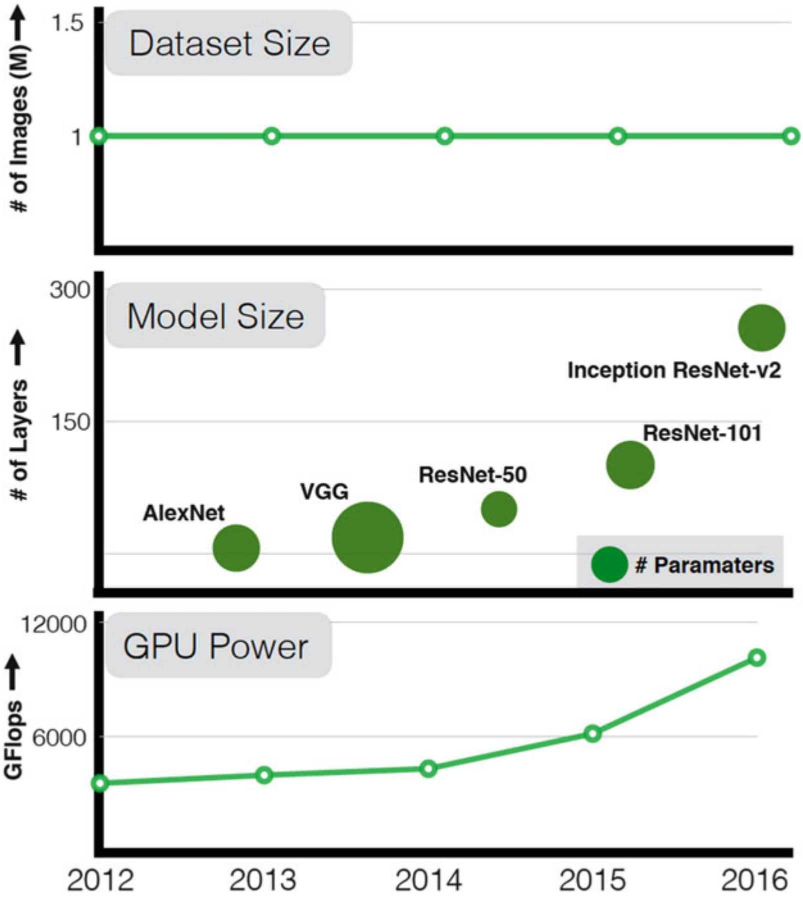


Fig. 6.1 Changes in dataset size, model size and GPU power over time. *Reproduced from* [2]

In the meantime, both GPU processing power and complexity of machine learning models have advanced rapidly (Fig. 6.1).

It is clear that larger conversations around the sharing of medical imaging data at scale are needed. These conversations require input from all parties, namely the data originators (patients), data controllers (healthcare providers) and data processors (machine learning researchers), and will be aided by setting out a common language in which to understand the underlying problems in medical imaging data sharing and data quality.

6.2 Not All Data Is Created Equal

Despite large strides in the introduction of digital PACS globally over the past few decades, and the existence and acceptance of international DICOM standards for the storage and transfer of medical imaging data, there remain significant barriers to large-scale big data sharing. Not all clinical providers have built or purchased infrastructure that allows for true interoperability with other systems. In addition, DICOM standards are only loosely adhered to, with large variation in the quality of DICOM metatags and other data points.

For instance, at a very high level, even the nomenclature of imaging studies is not standardised, one clinical site may refer to a CT study as ‘Chest, Abdomen and Pelvis’ as a DICOM header, and another as ‘Thorax to Pelvis’. Even more problematically, these headers may be different for each vendor hardware at the same clinical site. In some circumstances imaging studies are labelled completely incorrectly as technologists manually change the DICOM header inadvertently while attempting to optimise certain acquisition parameters (e.g. using an anatomically different acquisition protocol). This mismatch in simple image labelling affects the overall data quality, especially when researchers attempt to train and validate on multi-site and multi-vendor data. For example, Gueld et al. [4] found that it was impossible to automatically categorise medical images based solely on their DICOM metatags, as around 15% of all studies were labelled incorrectly due to human factors. Understanding this type of underlying discrepancy in the data (amongst many) is vital to ensuring that both researchers and clinical sharing sites can plan machine learning projects accordingly. Often, the amount of time dedicated to 'data cleaning' is disproportionate to the goal of the machine learning project being conducted. Indeed, many of today's algorithms can be trained in a matter of hours, whereas large-scale data cleaning can take months. In projects that are commonly constrained by funding periods and limited resources, a non-anticipated high burden of data curation at the start of the project can severely affect the entire project and lead to high risk of failure.

The FAIR Guiding Principles [5] state that 'good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process.' FAIR stands for findability, accessibility, interoperability and reusability, which are perceived as the four key factors affecting data quality. It is important to note that in the age of machine learning, 'reusability' not only refers to reuse by humans, but also by machines. To that effect, it is important to consider how to make data machine readable in order to make best use of modern technologies.

Good quality data management should enable both human and machine interrogators to establish data's identity, usefulness and accessibility quickly and easily. The reality however is far from this at present, especially at clinical sites not used to large-scale research. Additionally, hardware vendors in medical imaging have not been incentivised to support good data stewardship in terms of interoperability, as it can be perceived as damaging the competitive edge provided by being a proprietary solution. It is also worth considering that many clinicians and technical staff are completely untrained in the principles of effective data management, and therefore one should not underestimate the human factors involved in creating variability, errors and omissions in data.

6.3 The MIDaR Scale

There is no standard definition of what encompasses a baseline medical imaging dataset for machine learning. Kohli et al. [6] have previously well described an outline of the minimum requirements of imaging metadata (Table 6.1), with the understanding that these may significantly change depending on the clinical use-case and data type. This list, due to its semantic nature and variability dependent on clinical scenario, is not a useful one for introducing a common reference standard which non-data engineers can understand. More useful, but high level, is their statement that 'the ideal medical image dataset for an ML application has adequate data volume, annotation, truth, and reusability'. It is these factors, plus accessibility and interoperability, as highlighted by the FAIR principles, that are generally considered the key factors of medical image data quality. However, there is no standardised method to describe all of these factors, nor a recognised structure in which to group medical image data at a high level into objectively defined categories of 'readiness' for machine learning.

In a position paper on data readiness [7], Neil Lawrence proposed a three point scale to better allow inter-disciplinary conversations on the inherent readiness of data. Inspired by this schema, but taking into the account additional and specific requirements relating to research on medical imaging data, as well as reflecting the inverse relationship between volume of data and its readiness, we therefore propose a four-point medical imaging data readiness (MIDaR) scale.

Table 6.1 Baseline medical image metadata for machine learning tasks

Baseline metadata to catalogue medical image data
1. Image types
(a) Modality
(b) Resolution
(c) Number of images total and by series
2. Number of imaging examinations
3. Image examination source(s)
4. Image acquisition parameters
5. Image storage parameters (e.g. compression amount and type)
6. Annotation
(a) Type
(b) What is annotated, and how
7. Context
8. How is ground truth defined and labelled
9. Associated data
(a) Demographic
(b) Clinical
(c) Lab
(d) Genomic
(e) Timeline
(f) Social media
10. Date range of image exam acquisition
11. Log of dataset use
12. Who owns the data
13. Who is responsible for the data
14. Allowable usage
15. Access parameters
(a) Accessibility
(b) Costs and business agreements
16. Case distribution
(a) % Normals vs abnormals
(b) Summary of abnormal examinations
(i) Number of examinations with each pathology
Many of these are semantic, with further subcategories not listed here

Reproduced from [6]

The MIDaR scale (Fig. 6.2) is designed to objectively clarify ‘data readiness’ for all interested parties, including researchers seeking imaging data and clinical providers and patients aiming to share their imaging data. It is hoped that the MIDaR scale will be used globally during collaborative academic and business conversations, so that everyone can more easily understand and quickly appraise the relevant stages of data readiness for machine learning in relation to their AI development projects. Data refinement is a task that all AI researchers must acknowledge, and its ‘cost’ in terms of resources and time must be taken into account from the beginning of any AI project. By clarifying the stages of data refinement it is hoped that the often monumental task of preparing data for AI method development can be made more bite-sized and approachable.

6.3.1 MIDaR Level D

In order to begin the data refinement process it is prudent to start early discussions with the data controllers. The key points to consider are defining data quantity, quality, anonymisation and access. The first level of the MIDaR scale describes medical imaging data in its ‘natural habitat’: the clinical PACS system. Level D data is that which represents only its initial intended purpose of acting as a record of clinical activity, with no further consideration for research of any kind. While

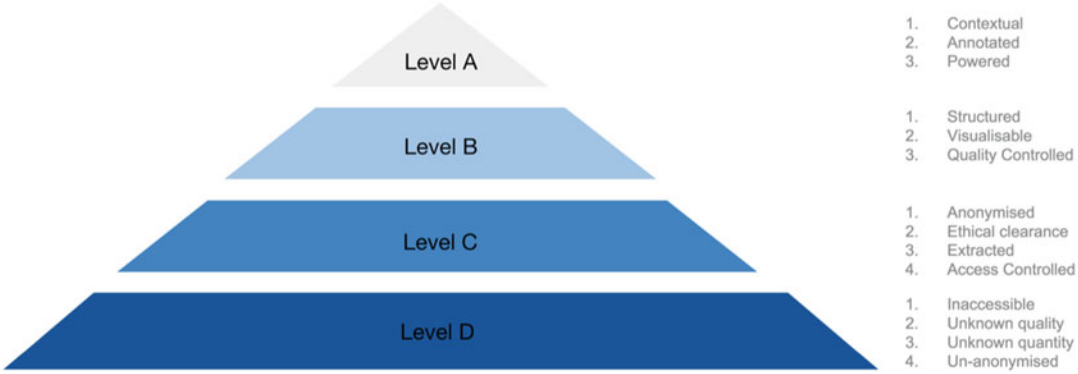


Fig. 6.2 Four-point medical imaging data readiness scale (MIDaR)

this type of data is the most abundant globally, with exobytes laying dormant in clinical systems, it is the least valuable per unit of all the data readiness levels, and unsuitable for supervised or semi-supervised machine learning tasks.

Level D data is defined by the following attributes:

Contains Patient Identifiable Information For machine learning development, medical data must be fully anonymised according to most laws governing data standards. That means any personally identifying information (PII) needs to be removed prior to any further use for research and method development. Level D data, however, is explicitly linked to patient identifiers for clinical purposes, and is therefore unsuitable for general release for algorithmic training and development by third parties. Therefore, data controllers need to implement mechanisms that can remove any sensitive information. Such mechanisms can often be entirely automated, typically using batch processing with scripts that define which data fields are to be removed from the DICOM header. In some situations, PII is hard coded or burnt into the image information as overlays (e.g. in ultrasound images). For these cases, more advanced image processing routines need to be employed to remove such information directly from the imaging data. Various levels of de-identification are covered within DICOM Supplement 142 concerning Clinical Trials de-identification [8], and the Cancer Imaging Archive also maintains a list of metadata which should be hashed, removed or fuzzed [9].

Unverified in Quantity The amount of medical data in any given PACS system at any given time is usually an unknown. Additionally, subcategorisation of data (e.g. by DICOM headers, modality or body location) is not possible with any accuracy. Estimates of data quantity may be available, with large error margins to be expected. Quantity of available data is often based on crude ‘guesstimates’ or even ‘hearsay’. Determining how much data of a particular type (e.g. trauma head CT scans) is actually available can require significant effort. For project planning, however, it is crucial to obtain reliable information which will influence, for example, the type of machine learning that can be considered.

Unverified in Quality The quality of the underlying data is not known, nor is it easily possible to perform checks on quality. It may be present in multiple formats, with varying degrees of compatibility, and will contain multiple errors, both in veracity of data present and omission of data.

Inaccessible to Researchers Level D data is only accessible to clinicians granted access by the representative healthcare provider, and to the individual patient or carer. Inaccessibility to researchers may be due to any combination of ethical, social, monetary and privacy concerns. Making data accessible for research involves a number of steps, and most importantly concerns patient consent and ethical approval.

Gaining access to level D data is in itself often a large challenge for medical image analysis researchers. Some large-scale institutions have started creating 'databanks'—mirror images of live clinical PACS systems with basic PII removed. This acts as a sandbox environment in which approved researchers may play-test early in the image data acquisition phase. Indeed, if these institutions were to take the extra steps needed to convert these image banks into level C data, then the barriers to entry for researchers would be significantly reduced, leading to faster development of machine learning applications. In an ideal setting, entire nations worth of data would be converted to this format, allowing for an entirely new industry of generalisable algorithms to be built on its foundation.

6.3.2 MIDaR Level C

Level C data represents 'wild' data that has been anonymised and made accessible via ethical approval, data extraction and access control. However, the data itself is still subject to errors, omissions, noise and artefacts affecting both image and metadata quality.

The steps required to refine level D data to level C are as follows:

Ethical Approval Typically local ethics committee approval will be required, which may be a lengthy process depending on the set up of the local organisation. Ethics committees may require evidence that data access is only being granted for the minimum required dataset for research, which can conflict with the tenet that machine learning systems rely heavily on vast amounts of data. When handling large amounts of retrospective data it is neither feasible nor practical to expect researchers to gain individual consent for every case. For this reason, many larger institutions with a track-record of big data research have implemented 'opt-out' consent models for all their clinical data, enabling faster ethical approvals and access to data at scale. Each geographic territory has its own local laws and regulations surrounding ethical approval for scientific research. European countries abide by the European ethics review procedure for research funded by the EU (2013) which indicates the main points of attention for the ethics review procedure as a part of the 7th Framework Programme (FP7) [10]. In the UK, the most commonly used system is the integrated research application system (IRAS) [11] which is used to describe research aims and methodologies prior to submission to the Research Ethics Service (RES) [12]. In the USA, researchers must apply to Institutional Review Boards (IRBs) with research proposals designed, reviewed, approved, and implemented in accord with accepted ethical principles and the US Department of Health and Human Services (45 CFR 46) and US Food and Drug Administration (21 CFR 50 and 56) regulations for the protection of human subjects [13].

Data Extraction In order to extract significant quantities of clean data from a working PACS environment, a thorough knowledge of the underlying vendor files formats and media storage methods is required. Most PACS vendors do not store DICOM files within their data centers, with many using binary large objects (BLOB's) to encode instances at the study level, or in many cases even using proprietary methods for compressing image and metadata. This data must also be cleansed against patient demographics updates received by HL7 messages. Not only must data be converted from vendor storage into DICOM 3.0 Part 10 format, but updated information concerning PatientID, PatientName, PatientSex, PatientBirthDate, BodyPartExamined and StudyDescription must be extracted from the working PACS database and applied to the 'stale' data stored on spinning disk or tape as it is extracted. This direct access to stored data is required given the size of PACS environments, where it is not uncommon to find 1 petabyte archives storing 10 million studies in up to 2 billion instances total. As PACS Query/Retrieve interfaces are often limited to retrieval speeds of less than 5–10,000 studies/day, using Query/Retrieve at this scale would lead to multi-year migrations. In addition to standard metatags, OEM vendors also extensively utilise DICOM Private Tags which contain highly clinically relevant data which often should not be stripped out. There are often equal numbers of DICOM tags and vendor private tags in a file system, and only by combing through the data can particular private tags be identified and a decision made as to include them or not.

Access Control ML researchers typically reside outside of the clinical PACS environment, and are more often than not in a geographically different location. For this reason access control to ethically approved and de-personalised data is required, particularly when live or non-explicitly consented patient data is in use. There are numerous variations of access control systems, many of which rely on at least 256bit SSH encryption as a backbone for data security during transfer. Detailed description of the various methodologies for access control is beyond the scope of this chapter.

6.3.3 MIDaR Level B

At Level B, the quantity and quality of relevant datasets are fully accounted for, and large-scale errors in data structure and format have been resolved. Task-specific data is separated from unwanted or poor quality data.

To get to Level B, the processes of data selection, visualisation, and quality control are performed on Level C data.

Data Selection For example, when developing an algorithm to spot lines and tubes on inpatient ICU studies, a proprietary set of chest X-rays from a hospital will include PA and AP films, as well as those from both inpatient and outpatient settings. Such an algorithm will not benefit from outpatient films as they are very unlikely to contain relevant information pertinent to the task, and they can be discarded. During visualisation and quality control processes the relevant films for the task can be tagged as such, and then selected out for the purposes of creating a dataset ready for labelling. It is interesting to note, that machine learning itself may be used to help with these tasks. For example, a relatively simple image classifier that recognises the type of scan and imaged body part could be employed in the context of image retrieval and automated categorisation of scans. Santosh et al. published work on automated categorisation of frontal and lateral chest films [14], for example.

Quality Control Structuring data in homogenised and machine readable formats will further refine it towards the end goal. It is no surprise that automated quality control of imaging data is an active field of research. Again, machine learning methods can be employed for this task to automatically determine image quality, for example, to check for motion artefacts or whether an organ of interest is fully visible.

Image noise is detrimental to knowledge extraction from the data and can spoil models obtained using noisy data, compared to models trained on clean data for the same problem. Figure 6.3 demonstrates common artefacts found in CT brain imaging, including those attained at the image acquisition phase. Figure 6.4 demonstrates the preferred quality of image data given the same clinical task. Common artefacts such as beam hardening, partial voluming, metallic star signal and under-sampling are desirable to remove from any training dataset. In many cases, manual review of the images is necessary to ensure adequate quality control, which requires both expert readers and time, and may be financially costly.

Data Visualisation Once the selected data has been formatted and structured, it becomes ‘visualisable’; that is, researchers are able to run statistical and quality tests over the data. This has many useful functions, such as quantitative assessment of data quality, volume per variable, checking for underlying distributions (e.g. uniformity or skew) and bias. It is only at this stage will researchers begin to be able to understand any underlying patterns or biases in the data that may or may not affect the models they are to train.

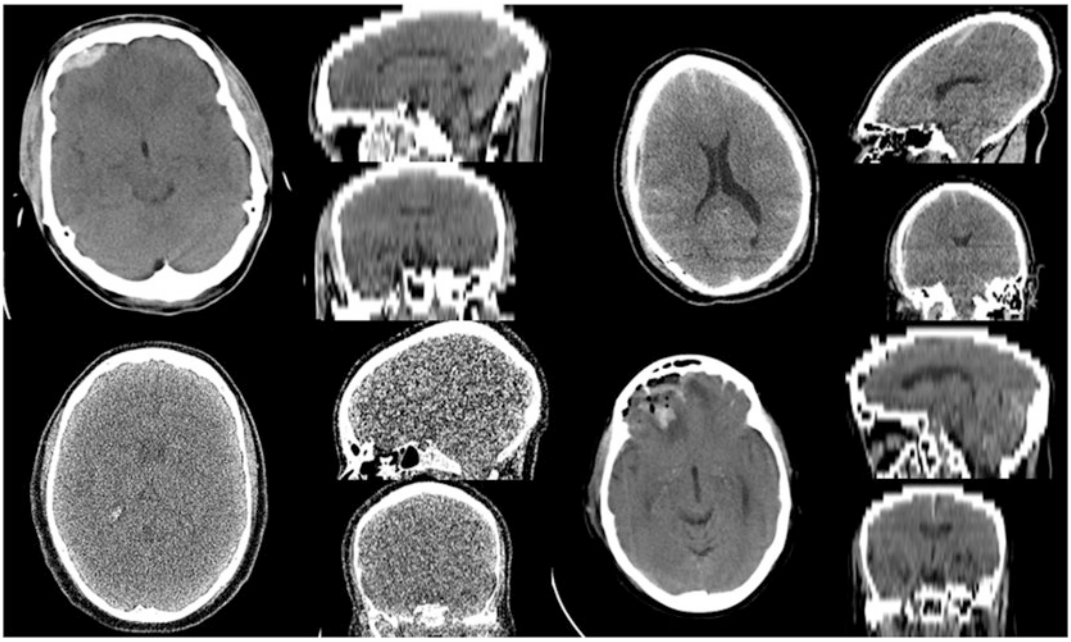


Fig. 6.3 MIDaR level C imaging data contains artefacts, data corruptions and noise

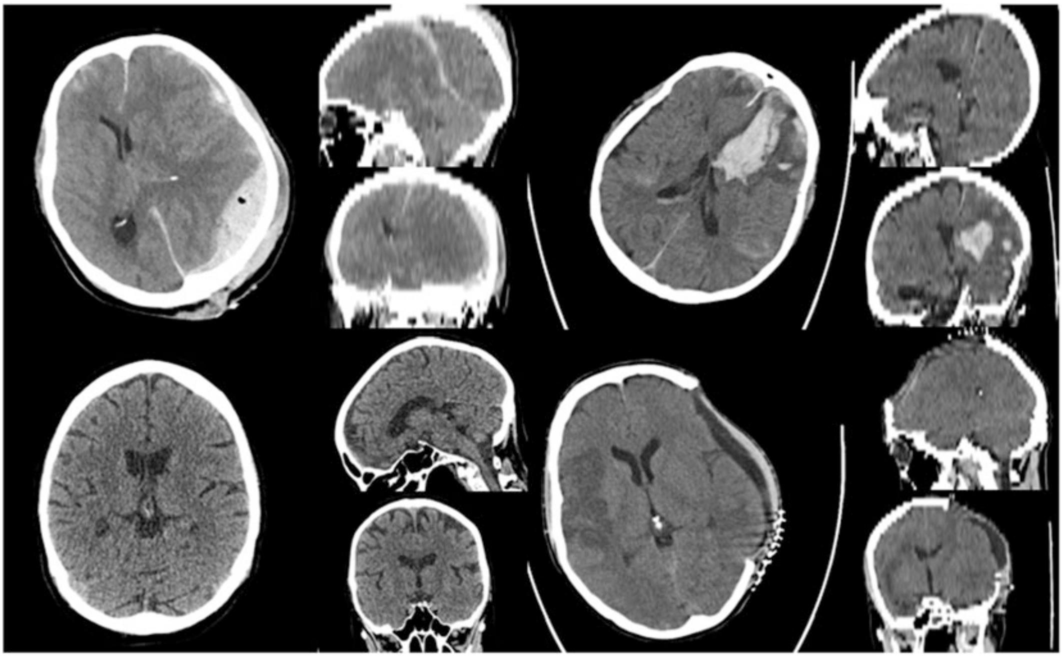


Fig. 6.4 MIDaR level B imaging data is structured and clean, but unlabelled

6.3.4 MIDaR Level A

Level A data is that which is as close to perfect for algorithmic development as possible. It is structured, fully annotated, has minimal noise and, most importantly, is contextually appropriate and ready for a specific machine learning task. An example would be a completely depersonalised, accessible and ethically approved dataset of non-contrast CT head studies, free from image artefact, noise or data corruptions, with patient age, gender, biopsy results (if relevant), blood tests, and diagnosis all structured under the same metatags, combined with expert level hand-drawn segmentations around tumours or other areas of interest. It is vital to note that the volume of Level A data is significantly smaller than the previous levels on the MIDaR scale due to the exacting need for labelled data, often contrary to the need for adequate statistical powering. Researchers may struggle to obtain enough Level A data to provide robust statistical analysis of their models.

Data Labelling Work towards Level A may be the most costly to achieve, depending on the need for an expertly labelled ground truth. Often medical imaging data comes with free text reports only, without annotations. In order to create both strong and weak labels for the images, various data labelling processes must be undertaken. Common techniques include NLP for information extraction [15], expert radiologist manual contouring (Fig. 6.5), derivation of consensus opinions or linkage to existing external clinical gold standard results. Whichever technique, or combination of techniques, is used to annotate and label imaging data, care must be taken that the labelling is performed consistently across the entire dataset, and without bias. This is a significant challenge, and one that remains largely unsolved for medical imaging data at large scales.

Powering It is also important to consider the statistical powering required for validation of machine learning algorithms. Smith and Nichols

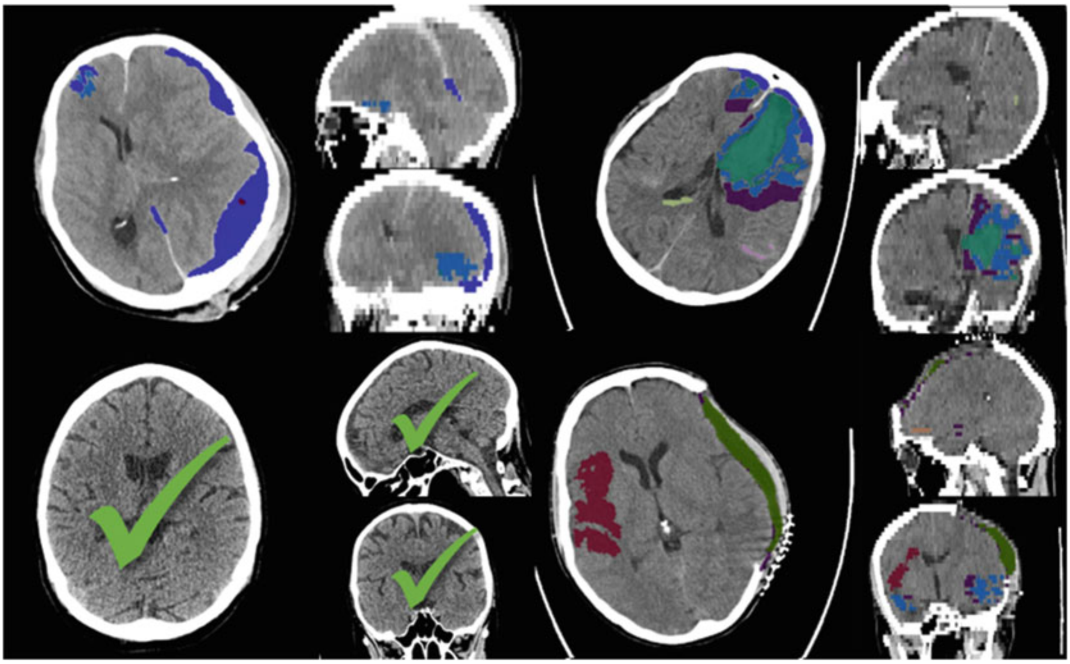


Fig. 6.5 MIDaR level A imaging data is annotated, powered and task ready. Green ticks signify normal scans

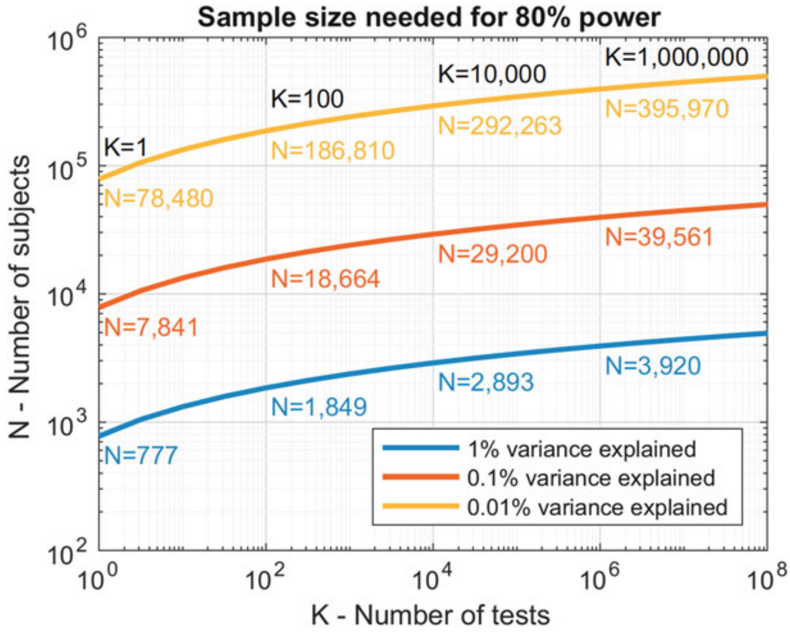


Fig. 6.6 Relationship between sample size and number of variables tested, holding statistical power constant. *Reproduced from* [16]

[16] nicely demonstrated that, roughly, by squaring the number of variables tested (K), the sample size (N) needed to attain 80% power to detect one true association is doubled. The size of this effect is shown in Fig. 6.6 at the percentage of variance at three small values, 1%, 0.1% and 0.01%. In essence they demonstrated that while performing one test requires a large sample size, as sample size increases, the number of tests you can perform increases exponentially. Putting this into practice, machine learning researchers should consider at Level B how many variables they plan on testing in order to ensure adequate volumes of data reach the Level A stage of refinement.

6.4 Summary

The MIDaR scale has been explicitly designed to enable conversations between data providers and researchers regarding the volume and level of data readiness required for machine learning projects. The key features of the MIDaR scale are the four high level categories of data readiness, and the associated decreasing volume of data as each category is reached. It is also worth noting that the 'value' of the data also increases as the scale is climbed; that is to say that Level A data is considered far more useful for machine learning than Level D data, and therefore is considerably more financially valuable.

When considering grant proposals for research, or commercial collaborations, the MIDaR scale will be useful for the design and planning of work packages and activities, especially in Gantt chart format. For instance, by categorising stages of data readiness and assigning their sub-tasks to individuals, the costs and time course for each readiness stage can be accounted for in an accurate and understandable manner. A researcher may approach a hospital and discuss access to their Level D data, then plan for ethics approval, data extraction and access control as part of the refinement to Level C in one single costed work package. They would then be able to further plan their data cleaning processes through to Level B, and finally be able to discuss with domain experts the costs and timings of data labelling for ground truthing to reach level A.

Hospitals and data controllers will also be able to use the MIDaR scale to plan ahead for large-scale research activities, by identifying which

Level D data they wish to refine, and considering how best to convert it to Level A. They may of course also wish to outsource or offer research grants to those who want to take on the various sub-tasks of the data readiness scale to help them on their journey.

Many of the popular open data science challenges (e.g. Kaggle) release small volume Level A datasets to the public, ready for training models in order to solve a particular problem. These competition organisers may also benefit from the MIDaR scale in planning for their next data release.

It is hoped that the MIDaR scale will be used during collaborative academic and business conversations, so that all parties can more easily understand and quickly appraise the relevant stages of data readiness for machine learning in relation to their AI development projects and their associated costs. For data to be refined from Level D to A, interested parties will have to negotiate responsibilities and resources for each task. It may be that neutral third party bodies in each country will be set up to undertake oversight of this work.

We believe that the MIDaR scale could become essential in the design, planning and management of AI medical imaging projects, and significantly increase chances of success.

6.5 Take Home Points

- There is currently no standard methodology for preparing medical imaging data for ML.
- The proposed MIDaR scale incorporates FAIR principles and stages of data readiness into a simple four-point framework:
  - Level D data is abundant, but inaccessible, un-anonymised and immeasurable in terms of quality and often quantity.
  - Level C data is anonymised, ethically cleared for use and access controlled but contains artefacts and noise.
  - Level B data is structured, quality controlled and visualisable, but unlabelled.
  - Level A data is labelled, statistically powered and contextually relevant for ML tasks.
- As data becomes refined towards level A, its value increases but volume decreases.

Acknowledgements With thanks to Hugh Lyshkow, DesAcc Inc. for his invaluable input and insight.

References

1. Sivarajah U, Kamal MM, Irani Z, Weerakkody V. Critical analysis of Big Data challenges and analytical methods. J Bus Res 2017;70:263–286. https://www.sciencedirect.com/science/article/pii/S014829631630488X.
2. Sun C, Shrivastava A, Singh S, Gupta A. Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE international conference on computer vision. Vol 2017-Oct. New York: IEEE; 2017. p. 843–52. ISBN: 9781538610329. https://doi.org/10.1109/ICCV.2017.97. http://ieeexplore.ieee.org/document/8237359/.
3. Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. IEEE Intell Syst. 2009;24(2)8–12. ISSN: 1541-1672. https://doi.org/10.1109/MIS.2009.36. http://ieeexplore.ieee.org/document/4804817/.
4. Gueld MO, Kohnen M, Keysers D, Schubert H, Wein BB, Bredno J, Lehmann TM. Quality of DICOM header information for image categorization. Proc SPIE. 2002;4685:280–7. ISSN: 0277786X. https://doi.org/10.1117/12.467017. http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=880364. http://dx.doi.org/10.1117/12.467017.
5. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, ’t Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016; 3:160018. ISSN: 2052-4463. https://doi.org/10.1038/sdata.2016.18. http://www.ncbi.nlm.nih.gov/pubmed/26978244 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4792175 http://www.nature.com/articles/sdata201618.

6. Kohli MD, Summers RM, Raymond Geis J. Medical image data and datasets in the era of machine learning-whitepaper from the 2016 C-MIMI Meeting Dataset Session. J Digit Imaging. 2017;30 (4):392–9. ISSN: 0897-1889. https://doi.org/10.1007/s10278-017-9976-3. http://www.ncbi.nlm.nih.gov/pubmed/28516233 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5537092 http://link.springer.com/10.1007/s10278-017-9976-3.
7. Lawrence ND. Data readiness levels; 2017. http://arxiv.org/abs/1705.02245.
8. Supplements – DICOM standard. https://www.dicomstandard.org/supplements/.
9. De-identification knowledge base - the cancer imaging archive (TCIA) public access - cancer imaging archive Wiki; 2017. https://wiki.cancerimagingarchive.net/display/Public/De-identification+Knowledge+Base.
10. European Commission - Directorate General for Research and Innovation. Ethics for researchers - Facilitating Research Excellence in FP7. Technical report; 2013. http://ec.europa.eu/research/fp7/index_en.cfm?pg=documents http://ec.europa.eu/research/participants/data/ref/fp7/89888/ethics-for-researchers_en.pdf.
11. Integrated Research Application System; 2018. https://www.myresearchproject.org.uk/.
12. Research Ethics Committees overview - Health Research Authority; 2018. https://www.hra.nhs.uk/about-us/committees-and-services/res-and-recs/research-ethics-committees-overview/.
13. Institutional Review Board; 2018. https://www.niehs.nih.gov/about/boards/irb/index.cfm.
14. Santosh KC, Wendling L. Automated chest X-ray image view classification using force histogram. Singapore: Springer; 2017. p. 333–42. https://doi.org/10.1007/978-981-10-4859-3_30. http://link.springer.com/10.1007/978-981-10-4859-3_30.
15. Pons E, Braun LMM, Myriam Hunink MG, Kors JA. Natural language processing in radiology: a systematic review. Radiology. 2016;279(2):329–43. ISSN: 0033-8419. https://doi.org/10.1148/radiol.16142770. http://pubs.rsna.org/doi/10.1148/radiol.16142770.
16. Smith SM, Nichols TE. Statistical challenges in “big data” human neuroimaging; 2018. ISSN: 10974199. http://www.ncbi.nlm.nih.gov/pubmed/29346749.

7 The Value of Structured Reporting for AI

Daniel Pinto dos Santos

Key Points

1. Structured reporting offers various advantages over conventional narrative reporting (including better data quality for the development of AI systems).
2. NLP could be helpful in extracting data from conventional reports but has some shortcomings if information is missing.
3. Interoperable standards for structured reporting templates are available, and major scientific societies are actively developing and providing templates.
4. AI systems could not only interact with report templates to extract data but also to automatically add and integrate their results to the final report.
5. The radiology report is a key component in various clinical workflows. Structured reports would offer numerous possibilities for AI systems to interact with the report and help improve patient care.

7.1 Introduction

When thinking about a radiologist's daily work, two activities stand out above most others: interpreting an imaging study and communicating the results of the said imaging study to the referring physician.

Clearly over the past decades, there have been tremendous advances with regard to the interpretation of imaging studies. Radiology has moved from the first plain radiographs to spectral computed tomography (CT) and multiparametric magnetic resonance imaging (MRI). Moreover, almost all workflows have been fully digitized, introducing electronic picture archiving and communication systems (PACS) and voice recognition (VR) to boost productivity. However, apart from the benefits of the electronic means of communication facilitating easier transmission of radiological reports, little has changed with regard to how radiologists convey the findings of imaging studies to the referring physician. Compared to one of the earliest known examples of a written radiological report (written as a letter from Dr. William J. Morton in 1896), most of today's reports have retained the same format [1]. While today's reports are generally being divided into subsections such as "Clinical Information," "Findings," and "Impression," they do not follow a particular structure with regard to the imaging findings. While other subspecialties (such as laboratory, endoscopy, and others) have already made the move from narrative reports to somewhat more structured formats, radiology seems to be lagging behind.

D. Pinto dos Santos (✉)
University Hospital of Cologne, Cologne, Germany

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_7

Especially with all the new and emerging technologies in artificial intelligence and computer vision, it could be crucial for further developments to make radiological reports more machine-readable allowing for specific information to be easily extracted. Unfortunately, until today the information contained in prose-like narrative reports is relatively difficult to extract. Of course, information could potentially be extracted using text mining and natural language processing (NLP) techniques, which have certainly made substantial improvements over the past years. However, there are still some relevant challenges to be faced when working with conventional reports. Variation in language and style, hedging and uncertainties, sometimes expressed in very variable terms, might pose difficulties to NLP systems. The information contained in different reports even from the same radiologist on the same clinical questions can also vary considerably leading to inconsistencies on what information can be extracted from the reports in comparable clinical settings.

It therefore seems reasonable to push for standardized and structured reports that not only would allow for information from reports to be handled more easily but would potentially also allow for integration of information from external applications.

7.2 Conventional Radiological Reporting Versus Structured Reporting

Long before the recent technical advances that would favor more structured ways of reporting, various publications have already addressed the fact that conventional radiological reports show a large amount of heterogeneity. Leaving aside the necessary variations in report content due to varying modalities and clinical questions, there is also substantial difference in structure, language, and vocabulary used based on the individual radiologist. An early study by Clinger et al. from 1988 reported that 40% of referring physicians thought that the reports were occasionally confusing [2].

Similar but fortunately a little less marked results were reported almost 20 years later by Bosmans et al. in 2011, showing that only half the respondents of a large survey stated that the language and style of radiology reports are mostly clear [3]. Interestingly, radiologists presented with the same statement were undecided whether or not language and style in reports are mostly clear. Considering that the radiological report is and probably will always remain the primary means of communicating with the referrers, this is a quite remarkable finding.

Consequently, various publications have tried to provide suggestions on how to improve a radiologist's reporting style in conventional prose reports [4, 5]. Nevertheless, formal instruction in radiology reporting is mostly still not provided as part of radiology training. One study from 2004 suggested that residents received only 1 h of training with regard to style and language in radiology reports [6]. Considering this, it is little surprising that most radiologists develop their very personal reporting styles and vocabulary along the years.

Addressing these issues, various studies have been carried out, comparing conventional prose reports to alternative formats such as itemized and structured reporting. There is a relatively large body of evidence supporting that structured reports have substantial benefits over narrative reports. For example, Schwartz et al. were able to show that structured reports had better content and greater clarity compared to conventional reports [7]. Similarly, in a study published by Brook et al., structured reports contained significantly more relevant information and facilitated surgical planning in patients with pancreatic carcinoma [8]. Comparable results could be found in numerous studies and mostly focused on oncological settings but also in other clinical scenarios such as pulmonary embolism [9–13].

With evidence growing that more structured approaches to radiological reporting could be beneficial, various radiological societies have published recommendations advocating for standardized and structured reporting. Among the first large societies to actively promote structured reporting was the American College of Radiology (ACR) in its summary of the 2007 Intersociety Conference [14]. Since then most societies have published recommendations advocating for structured reporting.

It is worth noting that the concept of structured reporting has been introduced to the radiological community as early as 1922, when Preston Hickey recommended the usage of a standard language and format for radiological reports [15]. However, almost a century later, structured reporting has still not found widespread application in clinical routine.

7.3 Technical Implementations of Structured Reporting and IHE MRRT

There are probably two main issues that hinder structured reporting to be adopted in clinical routine. On the one side, some radiologists might argue that they would prefer to simply keep writing narrative reports, because they dislike being forced to follow a predefined structure. Although this might be a difficult issue to overcome, it is probably not the biggest challenge. Much more relevant is probably the technical implementation of structured reporting within the radiologist's workflow.

Since Preston Hickey's first mention of standardized reporting, various attempts of implementation have been made. Of course, in 1922, there were only limited possibilities, but even then, itemized report forms were proposed, e.g., a report form for fractures proposed by Harold Pierce [15]. With the advent of computers in radiology, different software vendors offered solutions to build more standardized and structured reports, ranging from reusable text blocks to itemized lists of findings from which the radiologist could chose to build the final radiological report. Today various solutions are available, but most use proprietary report template formats that prevent different institutions from easily sharing those templates across vendors.

To address this issue, the Radiological Society of North America (RSNA) initiated a Reporting Initiative aimed at providing a vendor neutral standard for structured reporting. In its first development stage, a specific XML structure was proposed. However, this was then replaced by an HTML5-based format, which was also published as trial implementation by Integrating the Healthcare Enterprise (IHE) under the Management of Radiology Report Templates (MRRT) profile [16].

This profile describes how different software parts (named actors) should interact with one another, e.g., when the radiologist wants to query for a specific report template, as well as how these report templates should be built. Although this standard uses HTML5 as markup language for the template files, it is important to consider that the so-called Report Creator (the software the user interacts with to create the report) does not necessarily need to render the template as HTML to the user. The report template file basically describes an HTML form and allows for text areas, input field, and selection boxes. It is worth noting that all of these template elements can be linked to RadLex terms. The RadLex is an ontology of standardized radiological terms created by the RSNA aimed at providing a coding scheme and unified language for radiological reports [17]. This allows for even more reduction of variability, because even though a specific institution could use a different synonym for a particular finding, the associated RadLex code would ensure for consistency with other synonyms of the same finding. This consistency would also still be available if the report template was to be translated to a different language.

The RSNA's website radreport.org already offers a large number of report templates to be used, which to date have been viewed and downloaded over 5.5 million times. The European Society of Radiology (ESR) published a similar collection of report templates in different languages (open.radreport.org), and a memorandum of understanding has been signed to cooperate with the RSNA in developing templates for structured reporting. Some vendors do already provide support for this type of report templates. Mostly vendors of speech recognition software have added the possibility to import such report templates into their systems, which then allow the user to fill text areas and input fields with their dictation. However, these reports although structured to a certain degree are then transferred as plain text to the subsystem that further processes these reports (e.g., HIS/RIS or in some cases PACS). Although this is surely an important first step to promote structured reporting in clinical practice, the true benefits of structured reporting only arise when the data entered by the radiologist are stored in a database that makes the information easily accessible for further uses (e.g., to be used for AI systems). Prototypes of such systems have been published [18] but have not yet been incorporated by major vendors.

There are a few other concepts beyond template-based reporting that should be mentioned in this context. The RSNA recently published the concept of common data elements (CDEs) that define attributes and allowed values for a specific unit of information, e.g., when describing "image quality" (which in this case would be the attribute), only one of the values "adequate," "suboptimal," or "nondiagnostic" should be used [19]. These CDEs do not necessarily need to be used in the context of template-based reporting (as described, e.g., in the IHE MRRT) but could also be in conventional narrative reports to improve uniformity. Another notable project is the Annotation and Image Markup (AIM) which specifies how to store and communicate information about the content of a medical image [20].

These projects and initiatives taken together provide interoperable standards to describe and manage data from radiological studies, from the image-related data to the report content created by the radiologist. Unfortunately, adoption by vendors and subsequently in clinical routine is still very limited so far.

7.4 Information Extraction Using Natural Language Processing

Given the fact that until today structured reporting lacks widespread implementation, there is a massive historical record of narrative reports. It would of course be desirable if the information contained in these reports could be extracted, too. Not only would this information be valuable for retrospective clinical studies but also and especially for the development of AI systems that heavily depend on labeled training data.

Natural language processing (NLP) aims to address this challenge and describes techniques to digest written prose texts and extract relevant information. However, considering that most clinically produced texts have only limited structure and vary in vocabulary depending on the individual physician, these techniques will only be able to extract information to a certain degree. Nonetheless, NLP technologies have seen significant improvements over the last years, most notably to a broader audience with IBM Watson's performance in Jeopardy [21]. Consequently, numerous attempts at extracting information from clinical texts have been made using various techniques, and consecutively many studies have been published examining NLP performance in a medical context [22]. While most studies focused on other clinical text than the radiological report and aimed, e.g., at screening patients for their eligibility to be enrolled in clinical trials [23], there have also been some NLP systems developed specifically to extract information from reports on conventional radiographs or CT scans [24].

Due to the high variability of clinical texts, the performance of traditional rule-based approaches to NLP was not optimal and in various cases failed to generalize when applied to other cases as initially intended or outside the institution where it had been developed [22]. With the introduction of more refined machine learning techniques, there have recently been a number of studies reporting better performances at extracting a variety of information from the radiological report [25–27]. It is however worth noting that most published NLP systems do not aim at extracting detailed information from the report but rather try to categorize the reports into general categories (e.g., containing of not containing specific critical finding or follow-up recommendation).

Considering the incredibly large amount of historical radiological reports available in every institution's systems, these technologies will certainly play crucial roles in extracting valuable information to be reused, e.g., as a set of labels for training of other AI systems. However, these labels need to be treated with some caution as recently seen in the case of CheXNet [28], where a dataset of around 108,000 chest X-ray images [29] was used to train a convolutional neural network for the detection of pneumonia. Although the NLP system used to extract information from the radiological reports performed as expected and provided the labels to be used in training the network, these labels showed relevant inaccuracies [30]. Due to the substantial interobserver variability in terms used to describe a particular finding in the original chest X-rays, the extracted labels had some overlaps that made their suitability as ground truth for an algorithm at least questionable.

While for some tasks the information that is extracted using NLP technologies might be perfectly fine, it is important to consider that great caution is needed when using these data for AI systems that heavily rely on the accuracy of specific labels.

7.5 Information Extraction from Structured Reports

Considering the variability of terms used in conventional narrative reports, it seems obvious that a means of minimizing variability and ensuring completeness of information would be beneficial. Of course, when considering template-based reporting as described in the IHE MRRT, the question arises which information to incorporate in the report that later could be useful. This question is not always easy to answer as it requires some thoughts on which information is clinically relevant to the referring physician and which could be useful in the future to radiologists wanting to mine their data. Then again, a good reporting template should not overwhelm the reporting radiologists, so that inaccuracies arise from poor usability.

Different approaches could be used here, but it seems that consensus-based disease-specific templates would offer the most benefits. Such templates would ensure that reports are composed in accordance with current guidelines and evidence while also being limited to what is clinically relevant in a specific clinical setting. Such approaches have already been advocated in the literature [31–34]. Most notably large scientific societies such as the European Society for Gastrointestinal and Abdominal Radiology (ESGAR) and the Korean Society for Abdominal Radiology (KSAR) have published recommendations for the reporting of MRI in patients with rectal cancer [31, 32]. Also, the Society of Abdominal Radiology (SAR) and the American Pancreatic Association have published a report template for pancreatic ductal adenocarcinoma [33].

While such approaches to the radiological report would certainly greatly reduce variability and improve the completeness and standardization of the information contained, it still does not guarantee the correctness of the information. So, while the data from structured radiological reports might be much easier to use, there remains a certain degree of uncertainty which is inherent to any diagnostic system in which false positives and false negatives are possible. Nevertheless, through reduction of variability with regard to language, style, and content of the report, a much larger amount of data would potentially be accessible for further developments.

However, as pointed above, such report templates would need to be carefully crafted to include all relevant information in a structured way. The IHE MRRT profile also allows for parts of the report to be linked to coding systems such as the RadLex. This would potentially open the possibility of pooling data even from across borders with different languages, as the respective fields in the template would share the same RadLex code.

In a proof of concept, it was shown that data from structured reports can easily be handled and, e.g., be used to calculate epidemiological parameters [35]. In the presented use case, a report template for pulmonary embolism was developed, and over 500 structured reports were generated. As the information from the respective fields in the report template was then stored as discrete data elements in corresponding tables, this data was accessible to all sorts of further analysis and even to third-party applications. This demonstrates that information from structured reports can easily be used in various contexts.

7.6 Integration of External Data into Structured Reports

With all the possibilities of extracting data from structured reports and using this information in the context of AI, it is important to consider that the interaction between AI systems (or any other system) and structured report templates does not necessarily need to be a one-way street.

The automated integration of external data into the report would greatly improve the radiologist's efficiency and accuracy, as these data would not need to be re-dictated or reentered into the respective template fields. It is well-known that this step of either reentering or dictating measurements is prone to errors. In fact, for simple use cases like the integration of data on radiation exposure or applied contrast media, it has already been shown that incorporating data from external sources into structured reports is feasible using dedicated workflows and decreases the number of report addenda [36, 37].

The IHE MRRT profile also supports such integration of external data into the structured report by design. A merge field attribute can be specified that allows for external data to be used as input into the field the attribute is associated with. However, the profile does not (yet) define how this process of incorporating data should be implemented. In our department, we were able to show that this is easily feasible. A mapping table was specified within the reporting platform that referenced both to content of a DICOM SR file and to content of a report template. Subsequently, when a new report for a carotid ultrasound study was created, the platform searched the department's PACS for the DICOM SR file created by the ultrasound machine for the study the radiologist wanted to write the report. It then parsed this file and wrote the data from the DICOM SR to the report template's respective input fields. This workflow allowed to greatly reduce the radiologist's dictation time as only additional findings and final impression were needed to be entered into the respective fields.

Such technical implementation should in principle be feasible with any other third-party application looking to interact with report templates, although it is important to bear in mind that this is at the discretion of the report creator software used to generate the report and is not yet fully standardized.

7.7 Analytics and Clinical Decision Support

The possibilities of making the radiological report, both easily interpretable for AI systems and potentially accessible for AI systems to interact with, could lead to a large number of interesting applications in clinical routine.

When evaluating and treating a patient's condition, ideally all clinical decisions for further management should be made based on evidence and in most cases follow an established treatment pathway. As a simple example, an elderly patient could present himself to the emergency department with symptoms of shortness of breath, moderate tachycardia, and a history of a minor surgery for elective removal of the gallbladder 3 weeks ago. After, e.g., clinically ruling out pneumonia, pulmonary embolism could be suspected as a cause for the patient's symptoms. If the clinician then applied the Wells' score for pulmonary embolism, this patient would fall into the moderate-risk group, scoring three points (1.5 points for heart rate >100/min, 1.5 for surgery within 4 weeks prior) [38]. This in turn would almost certainly lead to the physician ordering a contrast-enhanced CT pulmonary angiography (CTPA) to confirm or rule out pulmonary embolism. Assuming that CTPA then revealed no thrombus in the pulmonary arteries, other causes for the symptoms should be evaluated. However, unfortunately for the patient, the CT scan of the chest also revealed a single incidental nodule of around 7 mm in size. Provided the patient has a history of heavy smoking, according to the Fleischner Society's 2017 guidelines for management of incidental nodules, this nodule should be followed up at 6–12 months and then again at 18–24 months [39].

In this rather simple fictional case alone, there are various ways where AI systems could interact with the radiological workflow and report to support the patient's management. For example, the ordered CTPA scan could be prioritized on the radiographer's and radiologist's worklist given that based on the original publication, the pretest probability for pulmonary embolism in this patient is relatively high with around 16.2% [38]. Furthermore, provided there was a dedicated AI system with good enough performance, such system could then pre-read the CT study and generate a preliminary structured report. Then, the radiologist would read the study and modify the structured report to include the incidental finding of a solitary nodule—or, if a computer-assisted detection system already detected the nodule, confirm the findings. This in turn could then trigger a simple algorithm prompting the referring physician or the radiologist to evaluate the patients' smoking history. Consequently, this would allow for automated suggestion and scheduling of appropriate follow-up for this particular patient based on the current version of the respective guidelines.

It may seem as if this example is of little interest, as the presented case is of relatively low complexity. But especially with regard to follow-up of incidental findings in CT, it was shown that only one third of all recommendations made in the radiology reports are consistent with the respective guidelines [40]. Not to mention that such findings are often not followed up as recommended. A study published by Blagev et al. in 2014 found that no incidental nodules were followed up that were only mentioned in the findings section of the report. And even when specific instructions for follow-up are given, only around a third of these nodules were followed up as appropriate [41].

However, more sophisticated applications for AI systems in this context could also be discussed. Today's guidelines and recommendations are usually established from the evidence obtained by dedicated clinical studies, be they prospective or retrospective. However, it can be questioned if these results do always generalize and translate to clinical routine. For example, considering the presented case of a patient with suspected pulmonary embolism, a prospective validation study on the value of the Wells' score would suggest a much lower pretest probability of only around 3% [42]. Furthermore, other studies suggest that D-dimer testing, potentially with age-adjusted cutoff values, should also be considered in calculating pretest probability [43, 44]. Also, there is evidence suggesting that CTPA studies are frequently not ordered in accordance with evidence-based guidelines [45, 46], that pulmonary embolism might be overdiagnosed [47], and that smaller emboli might not even need treatment [48, 49].

Similar examples could be constructed for oncological use cases as well. For example, AI systems could help to extract information from radiological reports and compare them with the histopathological findings. A similar approach has already been published using a more simple method where a dashboard was created that matched RadLex terms to terms used in the pathological report [50]. Results from such algorithms could then be used to support not only the radiologist to avoid misinterpretations or discrepancies but also to, e.g., guide the referring physician which treatment option was most beneficial to a specific patient with a similar constellation of findings.

For all of these use cases, AI systems could make a significant contribution to improving patient management and lowering healthcare costs. Conducting rigorous clinical trials should not and will certainly not be replaced, but if the information in radiological reports was more easily accessible, AI systems could use this data from clinical routine to continuously monitor, validate, and potentially refine the performance of such guidelines.

7.8 Outlook

Various publications have demonstrated the benefits of structured reporting. Not only does it improve the quality of radiological reports, but also referring physicians and radiologists alike tend to prefer it compared to conventional narrative reporting. From a more technical point of view, it seems obvious that it would also greatly help to make data from radiological reports reusable and accessible for other software. The potential applications are numerous.

Data from radiological reports could not only be better used to develop AI systems, but such systems could also interact with the reports. Data could be integrated to decrease the radiologists' workload and improve accuracy. It could be constantly monitored and be used to trigger messages to other clinicians, initiate other workflows based on findings from a report such as proactively scheduling and suggesting imaging protocols for follow-up (e.g., for incidental findings), or automatically translate reports to more lay language and thus also improve communication of the findings to the patient.

In a publication by Bosmans et al., the authors deemed structured reporting the "fusion reactor" for radiology [51]. It is clear that as the contribution of the radiologist to a patient's management and means of communicating with the referring physicians, the radiological report is a key component where data from various sources, be it imaging findings, laboratory results, or patient history, converge. It could therefore be a central component from which AI systems extract but also integrate data to generate new knowledge and support clinical workflows.

Unfortunately support for interoperable report templates and usage of such templates in clinical routine are still lacking. Nevertheless, radiologists should push for structured reporting to be implemented in their daily practice as it could be a cornerstone for many potential improvements.

References

1. Langlotz CP. The radiology report. 2015.
2. Clinger NJ, Hunter TB, Hillman BJ. Radiology reporting: attitudes of referring physicians. Radiology. 1988;169(3):825–6.
3. Bosmans JML, Weyler JJ, De Schepper AM, Parizel PM. The radiology report as seen by radiologists and referring clinicians: results of the COVER and ROVER surveys. Radiology. 2011;259(1):184–95.
4. Hall FM. Language of the radiology report: primer for residents and wayward radiologists. Am J Roentgenol. 2000 Nov;175(5):1239–42.
5. Ridley LJ. Guide to the radiology report. Australas Radiol. 2002;46(4):366–9.
6. Sistrom C, Lanier L, Mancuso A. Reporting instruction for radiology residents. Acad Radiol. 2004;11(1):76–84.
7. Schwartz LH, Panicek DM, Berk AR, Li Y, Hricak H. Improving communication of diagnostic radiology findings through structured reporting. Radiology. 2011;260(1):174–81.
8. Brook OR, Brook A, Vollmer CM, Kent TS, Sanchez N, Pedrosa I. Structured reporting of multiphasic CT for pancreatic cancer: potential effect on staging and surgical planning. Radiology. 2015;274(2):464–72.
9. Flusberg M, Ganeles J, Ekinci T, Goldberg-Stein S, Paroder V, Kobi M, et al. Impact of a structured report template on the quality of CT and MRI reports for hepatocellular carcinoma diagnosis. J Am Coll Radiol. 2017;14(9):1206–11.
10. Sahni VA, Silveira PC, Sainani NI, Khorasani R. Impact of a structured report template on the quality of MRI reports for rectal cancer staging. Am J Roentgenol. 2015;205(3):584–8.
11. Sabel BO, Plum JL, Kneidinger N, Leuschner G, Koletzko L, Raziorrouh B, et al. Structured reporting of CT examinations in acute pulmonary embolism. J Cardiovasc Comput Tomogr. 2017;11:188–95.
12. Dickerson E, Davenport MS, Syed F, Stuve O, Cohen JA, Rinker JR, et al. Effect of template reporting of brain MRIs for multiple sclerosis on report thoroughness and neurologist-rated quality: results of a prospective quality improvement project. J Am Coll Radiol. 2016;14:371–379.e1.
13. Evans LR, Fitzgerald MC, Varma D, Mitra B. A novel approach to improving the interpretation of CT brain in trauma. Injury. 2017;49:56–61.
14. Dunnick NR, Langlotz CP. The radiology report of the future: a summary of the 2007 Intersociety Conference. J Am Coll Radiol. 2008;5:626–9.
15. Hickey P. Standardization of Roentgen-ray reports. Am J Roentgenol. 1922;9:422–5.
16. IHE Radiology Technical Committee. IHE radiology technical framework supplement management of radiology report templates (MRRT). 2017. p.1–51.
17. Langlotz CP. RadLex: a new method for indexing online educational materials. Radiographics. 2006;26(6):1595–7.

18. Pinto dos Santos D, Klos G, Kloeckner R, Oberle R, Dueber C, Mildenberger P. Development of an IHE MRRT-compliant open-source web-based reporting platform. Eur Radiol. 2017;27(1):424–30.
19. Rubin DL, Kahn CE. Common data elements in radiology. Radiology. 2016;283:837–44.
20. Channin DS, Mongkolwat P, Kleper V, Rubin DL. The annotation and image mark-up project. Radiology. 2009;253(3):590–2.
21. Tesauro G, Gondek DC, Lenchner J, Fan J, Prager JM. Analysis of Watson's strategies for playing Jeopardy! J Artif Intell Res. 2013;47:205–51.
22. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform. 2017;73:14–29.
23. Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, et al. Automated clinical trial eligibility prescreening. J Am Med Inform Assoc. 2015;22(1):166–78.
24. Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al. Natural language processing technologies in radiology research and clinical applications. Radiographics. 2016;36(1):176–91.
25. Hassanpour S, Langlotz CP. Information extraction from multi-institutional radiology reports. Artif Intell Med. 2016;66:29–39.
26. Gerstmair A, Daumke P, Simon K, Langer M, Kotter E. Intelligent image retrieval based on radiology reports. Eur Radiol. 2012;22(12):2750–8.
27. Zech J, Pain M, Titano J, Badgeley M, Schefflein J, Su A, et al. Natural language-based machine learning models for the annotation of clinical radiology reports. Radiology. 2018;287:570–80.
28. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. 2017. cs.CV, arXiv.org.
29. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers R. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. https://arxiv.org/abs/1705.02315. Accessed 12 Aug 2018.
30. Oakden-Rayner L. CheXNet: an in-depth review. 2018. https://lukeoakdenrayner.wordpress.com/2018/01/24/chexnet-an-in-depth-review/, Accessed 12 Aug 2018.
31. Beets-Tan RGH, Lambregts DMJ, Maas M, Bipat S, Barbaro B, Curvo-Semedo L, et al. Magnetic resonance imaging for clinical management of rectal cancer: updated recommendations from the 2016 European Society of Gastrointestinal and Abdominal Radiology (ESGAR) consensus meeting. Eur Radiol. 2017;23(Suppl 1):2522–11.
32. KSAR Study Group for Rectal Cancer. Essential items for structured reporting of rectal cancer MRI: 2016 consensus recommendation from the korean society of abdominal radiology. Korean J Radiol. 2017;18(1):132–51.
33. Al-Hawary MM, Francis IR, Chari ST, Fishman EK, Hough DM, Lu DS, et al. Pancreatic ductal adenocarcinoma radiology reporting template: consensus statement of the Society of Abdominal Radiology and the American Pancreatic Association. Radiology. 2014;270(1):248–60.
34. Anderson TJT, Lu N, Brook OR. Disease-specific report templates for your practice. J Am Coll Radiol. 2017;14(8):1055–7.
35. Daniel PDS, Sonja S, Gordon A, Aline M-K, Christoph D, Peter M, et al. A proof of concept for epidemiological research using structured reporting with pulmonary embolism as a use case. Br J Radiol. 2018;91:20170564.
36. Goldberg-Stein S, Gutman D, Kaplun O, Wang D, Negassa A, Scheinfeld MH. Autopopulation of intravenous contrast type and dose in structured report templates decreases report addenda. J Am Coll Radiol. 2017;14(5):659–61.
37. Lee M-C, Chuang K-S, Hsu T-C, Lee C-D. Enhancement of structured reporting – an integration reporting module with radiation dose collection supporting. J Med Syst. 2016;40(11):852.
38. Wells PS, Anderson DR, Rodger M, Stiell I, Dreyer JF, Barnes D, et al. Excluding pulmonary embolism at the bedside without diagnostic imaging: management of patients with suspected pulmonary embolism presenting to the emergency department by using a simple clinical model and d-dimer. Ann Intern Med. 2001;135(2):98–107.
39. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung ANC, Mayo JR, et al. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. Radiology. 2017;284(1):228–43.
40. Lacson R, Prevedello LM, Andriole KP, Gill R, Lenoci-Edwards J, Roy C, et al. Factors associated with radiologists' adherence to Fleischner Society guidelines for management of pulmonary nodules. J Am Coll Radiol. 2012;9(7):468–73.
41. Blagev DP, Lloyd JF, Conner K, Dickerson J, Adams D, Stevens SM, et al. Follow-up of incidental pulmonary nodules and the radiology report. J Am Coll Radiol. 2016;13(2 Suppl):R18–24.
42. Wolf SJ, McCubbin TR, Feldhaus KM, Faragher JP, Adcock DM. Prospective validation of wells criteria in the evaluation of patients with suspected pulmonary embolism. Ann Emerg Med. 2004;44(5):503–10.
43. Righini M, Van Es J, Exter Den PL, Roy P-M, Verschuren F, Ghuysen A, et al. Age-adjusted D-dimer cutoff levels to rule out pulmonary embolism: the ADJUST-PE study. JAMA. 2014;311(11):1117–24.
44. Char S, Yoon H-C. Improving appropriate use of pulmonary computed tomography angiography by increasing the serum D-dimer threshold and assessing clinical probability. Perm J. 2014;18(4):10–5.
45. Raja AS, Ip IK, Dunne RM, Schuur JD, Mills AM, Khorasani R. Effects of performance feedback

reports on adherence to evidence-based guidelines in use of CT for evaluation of pulmonary embolism in the emergency department: a randomized trial. Am J Roentgenol. 2015;205(5):1–5.

46. Raja AS, Ip IK, Prevedello LM, Sodickson AD, Farkas C, Zane RD, et al. Effect of computerized clinical decision support on the use and yield of CT pulmonary angiography in the emergency department. Radiology. 2012;262(2):468–74.
47. Hutchinson BD, Navin P, Marom EM, Truong MT, Bruzzi JF. Overdiagnosis of pulmonary embolism by pulmonary CT angiography. Am J Roentgenol. 2015;205(2):271–7.
48. Yoo HH, Queluz TH, Dib El R. Anticoagulant treatment for subsegmental pulmonary embolism. Cochrane Database Syst Rev. 2016;126(4):e266.
49. Bariteau A, Stewart LK, Emmett TW, Kline JA. Systematic review and meta-analysis of outcomes of patients with subsegmental pulmonary embolism with and without anticoagulation treatment. Acad Emerg Med. 2018;25(1):CD010222.
50. Kelahan LC, Kalaria AD, Filice RW. PathBot: a radiology-pathology correlation dashboard. J Digit Imaging. 2017;30(6):681–6.
51. Bosmans JML, Neri E, Ratib O, Kahn CE. Structured reporting: a fusion reactor hungry for fuel. Insights Imaging. 2015;6(1):129–32.

Artificial Intelligence in Medicine: Validation and Study Design

8

Luke Oakden-Rayner and Lyle John Palmer

Artificial intelligence (AI) applied to medicine is expected to have a significant impact on clinical practice [1]. Companies and academic groups worldwide have recognised the potential of technologies such as deep learning to enhance healthcare, and many research teams are now racing to produce AI systems to augment, or even to replace, doctors.

To take one specific area of medicine as an example, the sudden explosion of interest and investment into AI applied to medical image analysis ($152 million in 2017, up from $80 million in 2016 [2]) has far outstripped the clinical, bioethical and legal best practice framework necessary to implement AI in clinical settings. Indeed, at the time of writing, the US Food and Drug Administration has yet to provide guidance on *exactly* how they intend to assess and regulate these technologies.

The limited number of expert clinicians with meaningful experience and skills in AI has led to research teams that are predominantly or entirely made up of computer scientists, engineers and developers. These groups are rarely trained to design or evaluate their systems in a "medical" way, which risks suboptimal outcomes for patients as well as business failures when the systems do not perform as expected.

In developed nations such as the USA, it is unlikely that truly "unsafe" technologies will reach clinics and cause harm to patients due to a strong system of regulation. However, in common with clinical trials [3], there have been some noteworthy relocations of commercial medical AI research to the developing nations [4, 5], presumably (at least in part) to minimise both cost and regulatory burden. The concomitant risk of substandard research and hence direct harm to patients cannot be easily dismissed.

In this exciting, if not feverish, environment, it is more important than ever that we understand how to assess new technologies that may directly impact human health and, in the worst case scenarios, could lead to harm or even death.

8.1 The Validation of AI Technologies in Medicine

When we assess any change in medical practice that has the potential to impact human health, we want to know the answers to two fundamental questions regarding the validity of the change:

L. Oakden-Rayner (✉) · L. J. Palmer
School of Public Health, The University of Adelaide, Adelaide, Australia

Australian Institute of Machine Learning, Adelaide, Australia
e-mail: lyle.palmer@adelaide.edu.au

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_8

1. *Is it safe?*
2. *Is it effective*?

Safety is almost never absolute; few aspects of clinical medicine are completely free from risk of harm. Therefore, safety is defined as an "acceptable" risk of harm to a patient, usually compared to current practice. What an "acceptable" risk is will vary depending on the problem at hand and how much risk is associated with current methods. For instance, a higher level of risk of harm associated with an intervention for terminal cancer may be acceptable. Determination of acceptable risk is complex, generally involving government regulatory bodies and a range of experts, including clinicians, statisticians, health economists and possibly others.

Efficacy or *performance* (how well a system works) likewise depend upon the purpose of the AI system. In general, we can think of the efficacy as "how well does this system live up to its claims?" If we claim that a system is useful because it "saves lives", then how many lives does it save, and compared to what standard? If it will "save money", what is the dollar value of the savings? If it is "more accurate", then how do we measure that and what is the difference? These are different questions, which need to be evaluated in different ways.

When considering safety and efficacy, we need to recognise that AI applied to human *medicine* is different from most other forms of technology. In other domains, performance is often valued above all else, and the risks of a new technology are sometimes treated as secondary. This is exemplified in the unofficial motto of many Silicon Valley software companies—*"move fast and break things"*. In contrast to software companies, the official motto of many doctors (the Hippocratic Oath) begins with *"first, do no harm"*. The risk to life and health in medical research requires us to put safety first. In fact, human drug trials are legally required to *prove* drug safety before being *allowed* to test performance in humans.

Because any change in medical practice carries risk, medical AI system failures can have profound negative impacts on the health and lives of patients. Before we start considering how to validly assess performance, we need to first consider safety in more detail.

8.2 Safety in Medical AI

A key issue related to the question of medical AI safety is the notion of *autonomy*. For instance, AI systems can perform a range of tasks in medical imaging. These can be simple, such as image processing tasks that humans find tedious and mechanical (e.g. measuring the size of an organ), or complex and cognitive, such as diagnosing a disease or even predicting what will happen to a patient in the future. In this framework, *risk increases with increasing complexity*. This is because, as a task increases in complexity from the perspective of humans, human aptitude to correctly judge the decisions of an AI system generally undergoes a concomitant decline. The most complex medical problems are often those where there is uncertainty regarding best practice and a dearth of objective evidence. These are also the exact kind of problems where AI systems may ultimately provide the largest benefit. To take an extreme example, imagine an AI model that claimed to predict individual patient risk of heart attack in the next 5 years with 80% accuracy. This is potentially very valuable clinical information; however, there is no doctor in the world that could reliably evaluate such a claim prima facie; it is simply not something doctors routinely do or indeed know how to do.

There is a threshold (the dotted line in Fig. 8.1), where AI systems transition from operating with doctors "in the loop" (i.e. *augmenting* human clinical practice) to being able to operate independently of human experts (i.e. *replacing* human clinical practice). A data processing system that performs a measurement is simply providing a piece of information to a doctor to inform their decisions, much as a blood test does today. Decisions made on these results are entirely the responsibility of the doctor. Similarly, efficiency tools like triage systems can reorder workflows but do not change the fact that all medical decisions are still made by a human doctor.

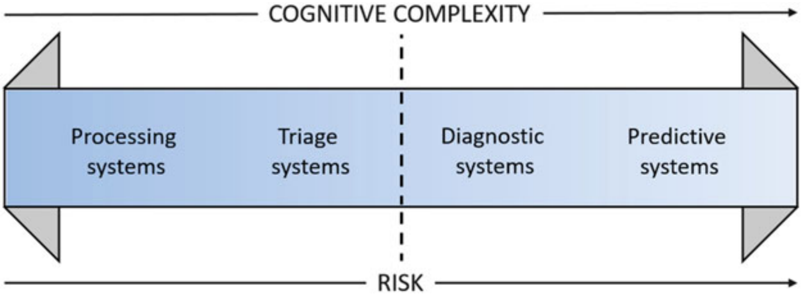


Fig. 8.1 AI risk through the lens of autonomy. As the tasks performed by AI become more complex from the perspective of humans, the level of risk increases

Diagnostic and predictive systems that can perform at or above the level of a human expert offer the possibility of *removing* humans from the loop; they can potentially lead to treatment decisions without human input. Such a situation has recently arisen in a number of medical AI systems, including an automated diagnostic system recently approved by the US Food and Drug Administration (FDA) [6]. A company called IDx has created a medical AI system that can automatically assess a photograph of the back of the eye for signs of diabetic eye disease. The FDA approval states "[the system] provides a screening decision without the need for a clinician to also interpret the image or results, which makes it usable by health care providers who may not normally be involved in eye care". This system produces both a diagnosis and a simple treatment plan—it decides which patients need specialist review and which do not. This is the equivalent of taking the primary care doctor out of the decision-making loop; the family doctors who will use the system do not have the specialist knowledge required to check the results, nor are they expected to. What are the risks associated with such a system? If the system overdiagnoses eye disease (i.e. too many false positives) and hence produces too many referrals, it will turn people into ophthalmic patients unnecessarily and increase the cost of healthcare. If the system underdiagnoses eye disease (i.e. too many false negatives), then undertreatment and preventable blindness may result.

How can we know these systems are safe before they are applied to patients? Unlike pharmaceuticals, medical devices and systems do not have to be tested in clinical conditions prior to approval. In general, showing the system performs as well as a current clinical method in a controlled experiment is acceptable, and it is assumed that these models will be safe in the real world. In medical AI, where it is unlikely that risks are negligible, especially if we are removing doctors from the decision-making process, this assumption is yet to be tested.

The rest of this chapter will consider how best to answer these questions. How can we show that a system is safe? How can we test performance? What is convincing evidence? How reliable do these results need to be? These questions are complex, but the medical community has spent many centuries honing the techniques required to answer them. These questions are evaluated using the guidelines and epidemiological methods developed for *clinical studies*.

8.3 Assessing Model Efficacy Using Clinical Studies

Epidemiological studies of people in a clinical setting are associated with a well-understood and accepted set of methods for performing experiments that allow us to answer key questions to some degree of certainty *while* avoiding undue risk of harm to patients.

The first point to make explicit is that clinical studies have nothing to do with *designing* or *training* models for medical AI. The training set, model architecture, model hyperparameters and so on are irrelevant. This is completely different to most machine learning research, where these factors are at least as important as the results. In medical AI studies, we can essentially treat an

AI model as a black box, in the same way that the physics of X-ray production is not relevant to a study comparing mammography to breast tomosynthesis. The purpose of a clinical study is limited to the assessment of the effects of a change in clinical practice.

Clinical studies come in many shapes and sizes for answering different questions, including observational studies such as surveys, case-control studies, cohort studies and experimental designs such as randomised controlled trials. All of the clinical studies used in medical AI to date have been *observational studies*. Specifically, almost all of them are *retrospective cohort studies* [7]. Cohort studies are a widely used study design in epidemiology, as they allow the assessment of associations between multiple exposures on the one hand and multiple outcomes on the other hand. A retrospective cohort design means that historical data on exposures and outcomes on a group of individuals with a disease of interest is collected retrospectively. Data collected from historical records usually includes relevant events for each individual, including the nature and time of exposure to a factor (such as an imaging study), the latent period prior to disease diagnosis and the time of any subsequent occurrence of the outcome. These data can then be used to assess the safety and performance of an AI system by performing a *"what-if"* experiment. Since we know the ultimate outcome of these patients in terms of their disease, we can ask: "if we had evaluated the patient with an AI system at an earlier time point, how would it have performed?" Such experiments are also known as *counterfactual*, where we investigate not what has happened, but what might have happened under differing conditions.

While it is possible to use other types of clinical studies to validate AI systems, this chapter will focus on retrospective cohort studies. At the end of the chapter, some possible limitations to this study design will be addressed.

The precise methods used in constructing retrospective cohort studies vary depending on the exact change in practice we want to evaluate but generally boil down to a few key concepts:

a. Start with a salient *clinical question* you want to answer.
b. Design a method to use to find the "correct" answer to the question, i.e. to create a reliable *ground truth* to measure the AI system against.
c. Identify the patient group we want to apply the system to—the *target population*.
d. Gather the patients we will test the system on—the *cohort*.
e. Perform the measurements we will use to judge safety and efficacy—*metrics*.
f. Pre-commit to a specific method to assess the performance of our system—the *analysis*.

These key points form the backbone of what is termed "study design". While a complete discussion of the topic is beyond the scope of this book (but can be found in a good epidemiology textbook such as Rothman et al. [8]), the most important thing to understand is that these points need to be considered and pre-committed to *a priori*, before any experiments are attempted. Failure to do so can invalidate any results obtained, to the extent that medical randomised trials and even systematic reviews are expected to preregister their study design. Any deviation from this design in the reported study must be thoroughly justified. While this sort of process is not yet commonplace in medical AI research, it is worth understanding why it is important.

In practice, points (a)–(d) above are often moot, as it will often be the case that there are data from a local (or public) extant research cohort available to medical AI researchers. The relevant design choices will have already been made by other medical researchers, often for unrelated (i.e. non-AI) purposes. In this setting, understanding the choices that created the data will be necessary to assess the validity of the study design. Alternatively, when a cohort must be constructed de novo, it is important that the checklist above guides the process.

Each of these points is examined in more detail below.

8.3.1 The Clinical Question

The most important element of study design is your clinical question. This is the purpose of the study, the "why?" of the AI system. In most medical AI studies, the questions are quite simple and are defined by the system you have built: "can our system perform better at a given clinical task than the current best practice?"

The question has two key elements: the task and the comparison.

The *task* is the question at hand. Measuring the size of an anatomical structure, triaging urgent cases and diagnosing a specific disease are all tasks. The range of medical tasks where AI may have a role is vast, and the question of how to select a worthwhile task is beyond the scope of this chapter. Suffice to say, almost any part of medical or paramedical practice can potentially be improved by AI. This means the task can be whatever makes sense for your clinical goal or business agenda.

The *comparison* is what we measure our AI system against. To determine if it is safe *enough* or performs *well enough*, we need a yardstick to define "enough" by. If your system measures the size of the heart, how well is it currently measured? If you are trying to diagnose patients with a disease, which ones actually have it, and how accurately do we currently detect them? An important concept here is that of a *gold standard*: an unequivocal measure of outcome that is widely agreed upon by the clinical community. This might be biochemistry, radiology, surgery or pathology confirming an outcome such as cancer. This is discussed further in Sect. 8.3.2.

In many tasks we will want to directly compare our model performance against the performance of human experts. In other cases we will compare our results against a test which is currently used in practice. In some rare cases (this is most often seen in predictive tasks, which are not a large part of modern medical practice), it may be that the task is entirely novel, and there is no yardstick to compare the model to.

If we compare against human experts, we will need to specify how many we intend to test. Most of the time, this decision is informed more by cost and availability of experts rather than any study design choice, but other factors relevant to this decision will be discussed later in the chapter (see Sects. 8.3.4 and 8.3.5).

Often, the largest problem we face in designing a medical AI experiment is determining the most relevant and accurate ground truth that we will use to compare our models to current practice.

8.3.2 The Ground Truth

If we want to measure the performance of our system (or current practice), we need a set of patients where the outcome of interest is known with certainty (this answer is often called the *label* for each case). We then use this ground truth to compare our system's answers against what we know to be true.

Ideally this ground truth should be perfect, where every single case is unequivocally characterised. Unfortunately, in medicine this is rarely possible; very few outcomes, diagnoses or treatments can be identified without some level of error. There are several reasons for this, including machine measurement error; bias due to selection, measurement or confounding; the use of subjective definitions; missing data, measurement errors, interpretation or reporting errors; and stochastic variation (chance).

Most diseases, pathologies and outcomes have *subjective definitions*. This means that they are open to some degree of interpretation, from the highly subjective such as the diagnosis of certain mental illnesses to the purely objective such as whether patients have died, which allows no scope for human interpretation. Often, diagnoses will fall somewhere in the middle. For instance, the diagnosis of stroke utilises objective criteria, but a clinical exam still plays a significant role in the diagnosis, and the diagnosis of stroke by any given method is subject to a degree of variability.

Missing data is fairly common in medical research, where patients or observations which are intended to be included in the study are not available for some reason. For example, patients can fail to return for follow-up appointments or withdraw from a study, move away or even pass away during the study period. For conditions where the ground truth is defined by the development of a diagnosis or outcome within a certain period of time (which can be the case in both retrospective and prospective trials), losing patients to follow-up will result in an inaccurate ground truth.

In the context of retrospective longitudinal cohort studies, differential loss to follow-up may affect the pool of patients available to be recruited into the cohort and can substantially bias study results. Such loss can arise from the "healthy survivor effect", where those most likely to be available to be recruited are those who are most healthy.

Similar bias can occur due to medical management heterogeneity. Patients with the same diseases or presenting complaints may undergo different tests, so only some of them will have the data required for enrolment in the study. Like the "healthy survivor effect", this will bias the study results if this inconsistency is not randomly distributed (i.e. being given a different test actually predicts your outcome).

Missing data is also present with patients who never present to medical care at all; if a model is supposed to work in all patients with stroke, it can never be trained or tested on patients who either never realised they had a stroke or chose not to go to hospital when they had symptoms. This can still bias the dataset but is much harder to recognise; how do you know a patient you have no information on is absent? This sort of missing data is usually only detected when population statistics are inconsistent, for example, when a study population contains less patients with a certain disease than would be expected.

Errors also commonly impact on the quality of the ground truth. For example, transcription errors in radiology reports can result in mislabelling of cases. A common error is that a negative word such as "no" or "none" is not transcribed, inverting the meaning of the report. We also have to consider if there are errors in how the ground truth is obtained. Because many datasets in medical AI are very large, it is common to use an algorithmic method to identify positive and negative cases. The algorithm can make errors, even when the original data is flawless. An example is in the case of a large public chest X-ray dataset, where automated text mining was used to identify cases with various imaging appearances [9]. The team estimated that the accuracy of this method was around 90%, meaning at least 10% of the labels were incorrect.

An inaccurate ground truth can lead to both negative and positive bias in the results of medical AI experiments. If the AI system is *more* likely to get the correct answer than the comparison *because* of the inaccurate ground truth, we will *over*estimate the true performance of our model. Conversely, if the AI system is *less* likely to get the correct answer than the comparison *because* of the inaccurate ground truth, we will *under*estimate the true performance of our model.

The first solution to these problems is to use criteria that are as objective as possible. Measurements made by machines are typically more objective than assessments that require human interpretation. Patient outcomes (such as whether the patient has died or experienced surgical complications) are also usually more objective, although they are also more likely to be missing—due to the need to access historical records for the patients in order to detect the outcomes.

If you must use human interpretation as your ground truth, it is common to develop the consensus opinion of multiple experts rather than use the interpretations of individuals (which is rarely used because it is considered unreliable). Consensus opinions are typically more stable and precise, for example, when screening mammograms are reported by two radiologists instead of one, it has been shown to increase sensitivity by around 10% [10].

The only way to control the effects of missing data is to understand your dataset. A thorough exploration of the relevant patient population is necessary to identify where data may be missing and to determine how best to manage any related biases.

Finally, the most important method of controlling ground truth error is to always review the ground truth when you can. For instance, in the case of medical image AI, the images and case labels should be reviewed by an expert radiologist wherever possible. In many datasets, particularly if the ground truth was created with an algorithm, it is common to find cases where the visual appearance is clearly inconsistent with the label. It is not unusual to find cases labelled "normal" but demonstrating glaringly obvious pathology. If the dataset is too large to check all cases, a random subset should be reviewed, so the accuracy of the ground truth can be estimated. It is generally good practice to report this estimate when presenting the results of your experiments.

8.3.3 The Target Population

The target population is the group of patients upon whom your system is designed to work. The purpose of a medical AI study is to estimate performance of the AI system in this group of patients. It is the group you will draw your test set from (see Sect. 8.3.4 cohort) and will usually be the source of your training data as well (Fig. 8.2). It is possible that the training data comes from other sources such as publicly available datasets, but it is important to be clear that a medical AI study can only make conclusions about patients in the target population.

The question of *why* you might want to target one particular population over another is highly complex and depends heavily on the eventual goal of your system. A screening system should be applicable to the community broadly, whereas a system to detect severe trauma need only work in patients who have suffered a traumatic event, perhaps in the setting of a hospital emergency department. There is no good rule of thumb to use; you must simply try to understand the task at hand and which patients the problem applies to. Consultation and collaboration with clinicians expert in the task being tested are key here.

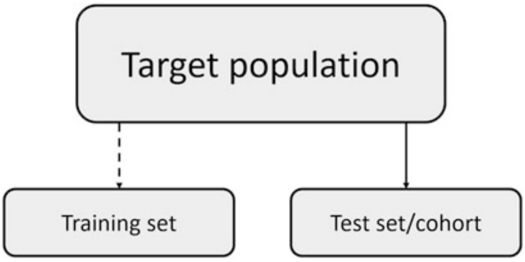


Fig. 8.2 The test set/cohort is drawn from the target population, the group that the study is intended to demonstrate the performance of the system in. The training set *may* be drawn from the same population, but this is not always the case

In medicine, we have learned from long experience that a wide range of population factors will determine how well a system works in practice [11]. Factors such as genetics/ancestry, sex, age, socioeconomic status and many others can alter baseline risk of disease, affect treatment response and impact disease development and progression.

Identifying the factors that can affect performance in your population can be tricky, since you do not actually have access to the future patients you want your system to apply to. There are two ways to go about this:

1. The ideal way is when the dataset you use to train your model is very large, drawn either sequentially or randomly from your target population. For example, in Gulshan et al. [12] the training set was 128,175 retinal photographs from three hospitals in India. In this setting, you should be able to assume that the training set is sufficiently similar to the target population (i.e. patients from the eye clinics in those three hospitals) and just describe the characteristics of the training set. It is important to note here that *perfectly* representative samples are almost impossible to find or construct in the real world, and hence clinical and epidemiological research is almost always done on samples that are not fully representative.
2. It will therefore most often be the case that your training dataset is smaller and different in some ways from the target population. In this case, you must rely on previous epidemiological research in the field. Most major clinical tasks have been studied in the past, and you should be able to find previous studies that describe the characteristics of your population.

> For instance, it is well understood from previous research that adult patients diagnosed with obstructive sleep apnoea will tend to be middle-aged, male and overweight or obese [13]. Alternatively, local administrative health data (such as from a government data linkage system or HMO/insurance company registry) may be available to define some of the key characteristics of your target population.

As a general rule of thumb, you should always consider the *big three* demographic factors:

- *Race/ancestry*
- *Age*
- *Sex*

While there are many other important factors, it is reasonable to consider these three the minimum to characterise a population. Occasionally it is possible to ignore one or more of these (particularly ancestry), but it would require justification. Whether you will need to address other types of variation will depend on your task.

The classic example in radiology of the problems caused by failing to appropriately consider population effects is seen in prenatal ultrasound. We measure the size of the foetus to determine if it is growing well and compare the size against a standard baseline. The outcome is: healthy babies with non-white ancestry are much more likely to be identified as developing abnormally [14], typically with smaller bones among Asian populations and larger bones among African-American populations [15]. In each of these populations, the baseline is different. Asian children are smaller on average than European children, who are in turn smaller on average than African-American children. Unfortunately, the baseline widely used in radiology was developed in a predominantly European population, so our results are less accurate in other racial groups. The presence of subgroups within our target population with different characteristics and baselines is known as *population stratification* (discussed further in Sect. 8.3.4).

There is nothing wrong with designing a system to work with a limited population (in this case, a measurement system for people of European descent), but if you market your system as working in all patients, then you are likely to run into real problems. Even outside of medicine, it is a major risk to ignore the potentially biasing effects of population stratification. Google was rightly heavily criticised for an image recognition system that incorrectly identified people with dark skin tones as "gorillas" [16]. More recently, Google has taken special care to train their models with population stratification in mind. Talking about a new photo/video analysis system, a Google team recently said: "we also made sure our dataset represented a wide range of ethnicities, genders, and ages" [17]. It is no surprise to medical researchers that Google has focused on the same big three demographic factors that we have always found to be almost universally important.

Understanding our target population can help us avoid these problems and let us predict situations where our models are more likely to fail. There is also another important reason to do so; without understanding our target population, we cannot select an appropriate cohort, and without a cohort, we will be unable to test our models.

8.3.4 The Cohort

The cohort is the set of people the system is tested on, often called the *test set* in the machine learning literature. For most problems that we want AI to solve, this will be a small subset of your target population.

The problem we face is that we need our experiments to say something meaningful about how the system will perform with real patients, that is, about how *generalizable* our results are to the target population. If we use too few patients or they are too different from our target population, then our results will be meaningless in practice. There are two major elements of cohort selection:

- *Which cases need to be part of the cohort to produce results that can reliably be extrapolated to the much larger population?*

- *How many cases do you need to use to convincingly show an effect?*

In general, a random, large sample of the population will work well. It should be similar to the population and needs to contain “enough” patients. Unfortunately, every patient that we use for testing our models is a patient that we cannot use to train our models. Since performance increases as a direct function of the size of the training dataset, there is a trade-off between achieving a strong system and being able to *prove* it is strong.

Many teams building medical AI appear to favour the former, selecting cohorts/test sets of the bare minimum size that they have estimated able to prove their system is safe and effective (these methods will be discussed later in this section, as well as in Sect. 8.3.6 on Analysis). While this is an attractive approach given the technologist’s mindset of *“performance at any cost”*, it leaves very little room for error. Even a single incorrect assumption in the design will invalidate the results of the entire study if the sample size is borderline, which wastes all of the effort of performing the study in the first place.

Which cases to include relates to the issue of *sampling bias*. This is when the AI system results are unreliable because the sample of patients in the cohort is too different from the target population, as in the examples of applying European-derived foetal measurements to Asian women or applying image recognition to people with different skin tones. In general the solution is to randomly select cases from the target population, with the goal of ending up with a cohort that is similar in essential characteristics to the target population.

It is not enough to randomise though, we must also confirm that the cohort reflects the target population. After the randomisation process, it is always worth comparing the characteristics of the cohort and the population (as in Fig. 8.2). If the average age, proportions of sex and ethnicity, prevalence of disease or any other important factors are not similar, then it is likely that your cohort is *biased*, simply by chance. In this setting, it is worth re-drawing the cohort from the population.

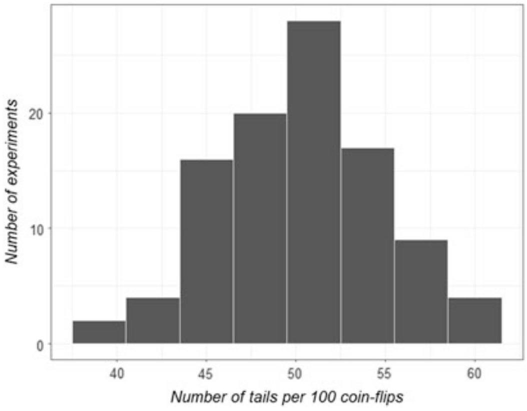


Fig. 8.3 A histogram showing the results of 100 experiments, where each experiment is a fair coin flipped 100 times. While the true probability of the coin landing on heads or tails is 50%, we see a wide range of results due to random chance

How many cases you need is a question of *statistical power*. The key concept to understand is that experimental results are somewhat random. If you run an experiment 100 times, the test sets will vary enough to produce different results each time. This is usually explained with a thought experiment using coin flips, as shown in Fig. 8.3.

When we do a medical AI study, we typically only perform a single experiment. In the coin example, there is nothing that prevents the first experiment you do returning a result of 35% tails, and this is also true for medical studies. There is always a chance that even a well performed experiment can produce misleading results.

The factors that determine how reliable your results will be are the *effect size* and the *sample size*. The stronger the effect (i.e. the difference between the model and the baseline), the easier it is to detect for a given sample size. Concomitantly, the larger your sample size, the smaller the true effects that you will be able to detect.

There are mathematical formulae that you can use to estimate how big your cohort will need to be to answer your question, but a good rule of thumb is that if you have over 100 cases (i.e. patients with the condition of interest), you

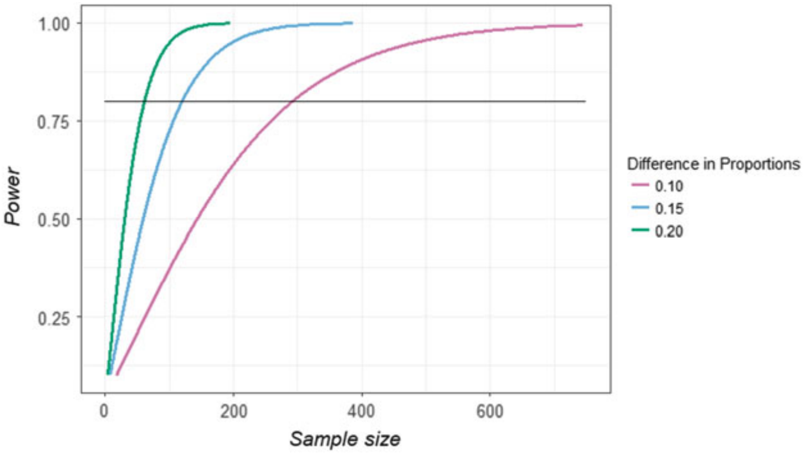


Fig. 8.4 Power curves showing the required sample size to identify a difference in proportions (e.g. in sensitivity or specificity) with α = 0.05 and using a two-sided test of proportions. The lower proportion is set at 0.70, and the black line reflects the 0.8 level of power usually considered the minimum appropriate in clinical studies

should have enough power to find a modest effect. This is *not* the final cohort size you will require, as there are many complexities that will vastly inflate the number of cases required (see below and Sect. 8.3.6 on Analysis). As we will see below, this rule of thumb often results in the need for cohort sizes in the thousands.

While the specifics of the methods in Fig. 8.4 are unimportant here, these numbers provide an intuition for how the required sample size increases as the effect size decreases. Again, this type of analysis assumes there are no other factors that increase the required sample size (which is almost never true).

Two of these factors are relevant when designing a cohort: (1) if the effect size you are searching for is small (i.e. less than 10%) and (2) if your cohort is stratified. In both settings the number of cases you will need can increase dramatically.

Stratification means that there are subgroups (strata) in your cohort, in which your model performance may be systematically different. The obvious example in medical AI studies is positive and negative cases (healthy people and people with a disease). Unless your model has exactly the same sensitivity and specificity, by definition it will be more likely to make errors in one of these groups than the other.

To deal with this, in classification experiments the rule of thumb applies to each class (i.e. positive and negative). You will need at least 100 cases of each of your classes to be able to reliably detect a modest effect. If you have two classes (e.g. patients with and without fractures), then you will need a minimum of 200 cases. For multi-class experiments (e.g. in trying to identify patients with mild, moderate and severe eye disease), you may need more.

Similarly, if you want to describe the performance within other subgroups of your cohort, then each subgroup will require a well-powered experimental design. An example could be that you want to assess how well a system can detect lung cancer on chest scans in smokers and non-smokers. In this case the rule of thumb would suggest you should gather at least 400 cases, 100 positives and 100 negatives for both smokers and non-smokers.

The final issue to consider when you are designing your cohort is *prevalence*—the frequency of occurrence of the disease or condition in the target population. Most diseases are rare, often occurring in less than 1% of the population. This means that if you need 100 cases with the disease, you will need 9900 disease-free cases. This is not a problem for testing your model; the cohort can simply be 10,000 cases in size, but if you want human doctors to perform the task as a comparison, it would be nearly impossible to get them to review that many cases. A solution is to provide the humans with an *enriched* test set, where the prevalence is much higher. The most common approach is to produce a test set with 50% prevalence (i.e. 100 positive cases and 100 negative cases in this example), which you can test both the doctors and the model on and compare their results.

That said, it is important to assess the model at clinical prevalence as well, so the 10,000 case test set will also be used. This will be explained in more detail in Sect. 8.3.5 on Metrics.

One last issue we need to consider occurs in the specific case where you are using a human comparison in your experiments (e.g. comparing the model to the performance of human doctors). In this setting the size of your doctor group is also relevant. While a large patient cohort will improve your power to detect small effects, it will be worthless if you use too few doctors. This is because the doctors will differ in their interpretations as well, so we have a further layer of possible variability. In mammography, for example, radiologists only agree with each other's interpretations around 78% of the time, and they only agree with *themselves* around 84% of the time [18]. While many studies are performed with only one or two radiologists, this is probably inadequate to accurately estimate human performance; the chance of randomly observing an outlier (like a score of 35% tails in our coin-flip experiment) is high.

We will look more at the size of human observer groups in Sect. 8.3.6, the analysis.

While cohort selection may appear to be complex, the underlying principle is quite simple. A large, randomly drawn sample from the target population will be adequate for most situations, as long as there are enough cases from the smallest stratum. While there is a tension between cohort size and training set size in many AI studies, the risk of producing an invalid study must be carefully considered. Allowing for a conservative margin of error by increasing the size of your cohort is a good way to avoid wasting a large amount of time and money.

8.3.5 Metrics

There are many possible ways to present results for a medical AI study and many ways to measure performance. These performance measures are called metrics, and the overarching principle of selecting which metrics to use is that you need to identify and report the different *ways* that your system makes mistakes.

If we want to diagnose a disease, for example, we can make two sorts of errors. We can overcall the disease in healthy people or under-call the disease in sick people. We call these false positives and false negatives, and the most commonly used metrics in medical practice reflect these; the true positive rate or *sensitivity* and the true negative rate or *specificity* are measures of these two types of error, as shown in Fig. 8.5.

Similarly, a model which predicts how long a patient is going to live (a regression task where sensitivity and specificity are not relevant) can erroneously overestimate or underestimate this value. While it is common in machine learning to treat these errors as equivalent and to report directionless metrics like the log-loss or mean-squared-error, the direction of the error makes a big difference in medical practice. Imagine being told you have a month to live but having a year. Conversely, imagine being told you have a year to live but having a month. These outcomes are significantly different to patients and their families, so it is important to understand (and report) what sort of mistakes any AI system is making.

Fig. 8.5 A confusion matrix extended to show sensitivity and specificity, probably the most commonly used metrics in medical research

	Case positive	**Case negative**
Predicted positive	True positive	False positive
Predicted negative	False negative	True negative
	Sensitivity = $\frac{\text{True positives}}{\text{Case positives}}$	*Specificity* = $\frac{\text{True negatives}}{\text{Case negatives}}$

The other important factor to consider is the *prevalence* of the condition—what proportion of the population actually have it? Sensitivity and specificity are prevalence-invariant metrics, meaning that they will remain the same whether the prevalence is 50% or 5%. This does not adequately describe the function of the model, because the number of each type of error in practice will change dramatically with prevalence.

Here is a simple example, where an AI model for detecting cancer has a sensitivity of 90% and a specificity of 90%. If we assume a prevalence of 1%, then we see the model will generate 1 false negative per 1000 cases and 99 false positives. This is very different than the impression given by sensitivity and specificity which at face value suggest that the false-positive and false-negative rates are balanced. To better reflect the high number of false positives that we will actually see in practice, we should use a prevalence-dependent metric such as the *positive predictive value* (also known as the *precision*), shown in Fig. 8.6.

While the specificity is 90% in the above example, the PPV is only 8.3%. This is a much better reflection of the very high number of false positives the model will produce in practice.

Finally, a common approach when assessing performance of a medical AI system in diagnostic tasks is to compare the model against the performance of humans at the same task. Unfortunately this can be difficult to demonstrate. This is because we have a single model, but we require multiple doctors to be tested so we can appreciate the (often large) variation among doctors (see Sect. 8.3.6 on The Analysis).

If we want to compare one versus many like this, we need to be very careful about how we judge human performance. The obvious approach would be to use average human sensitivity and specificity, comparing this to the sensitivity and specificity of the model at a specific *operating point*, but this actually results in quite a large underestimate of human performance. This is because all decision-makers *trade-off* sensitivity and specificity to make decisions, and this trade-off is not linear but rather a curve. This curve is called the *ROC curve*.

Since the trade-off is curved, the average of sensitivity and specificity will always exist inside the curve (the pink dot in Fig. 8.7). Points below the curve are fundamentally worse decision-makers than points on the curve, and the average

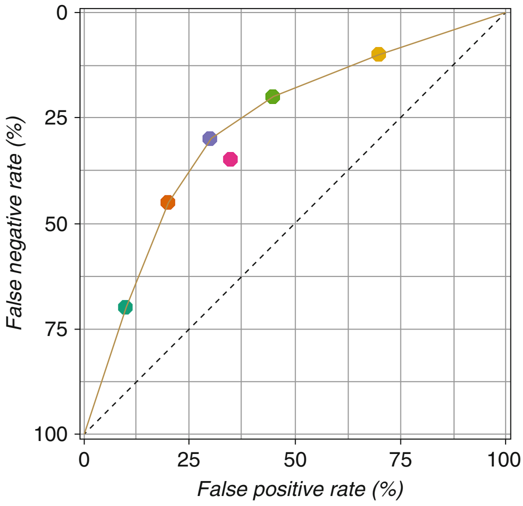


Fig. 8.7 The receiver operating characteristic (ROC) curve. Decision-makers of equivalent quality exist on a curve, and the further up and left the curve is, the better the decision-makers. The pink dot is the *average* sensitivity and specificity of the five decision-makers on this curve

Fig. 8.6 A confusion matrix extended to include the positive predictive value (or precision), which is probably the most commonly used prevalence-variant metric in medical research

	Case positive	**Case negative**	
Predicted positive	True positive	False positive	*PPV* = $\frac{\text{True positives}}{\text{Predicted positives}}$
Predicted negative	False negative	True negative	
	Sensitivity = $\frac{\text{True positives}}{\text{Case positives}}$	*Specificity* = $\frac{\text{True negatives}}{\text{Case negatives}}$	

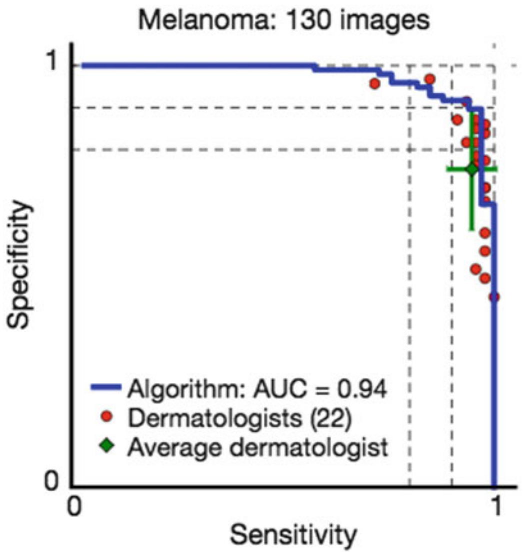


Fig. 8.8 A ROC curve (flipped left-right in this example) showing the performance of an AI model in blue and 22 dermatologists in orange. In this case, the task is determining if a skin lesion is malignant

of sensitivity and specificity gets *worse* as the decision-makers improve (and the curve becomes more convex).

In general, it is better to *show* this curve and the human experts on it that it is to try to boil it down into a number that can summarise it. This way you can appreciate how close the doctors are to the performance of your model.

In the real-world example in Fig. 8.8 from Esteva et al. [19], the dermatologists are very close or even above the ROC curve, but the average is inferior (below and to the left) of the hypothetical curve that the doctors form. This means that the average point is an inferior decision-maker to all doctors (because doctors operating in this region of ROC space all have better sensitivity *and* specificity) and is not a fair metric when the purpose of the study is to compare human and AI performance.

It is possible to directly compare a group of doctors to a single model using ROC analysis by calculating the area under the ROC curve (the AUC) for each group, but the complexity of this topic is beyond the scope of this chapter. A good starting point might be Hanley and McNeil [20], which discusses the ROC curves that can be created in situations where the doctors use a multipoint scoring system (called a Likert scale). Shiraishi et al. [21] provide further insight into how ROC analysis and AUC have been used previously in the radiology literature.

As a good rule of thumb, it is better to over-describe your results than under-describe them. Present several metrics so your readers can understand what sorts of errors the model makes. The US FDA suggests the use of multiple metrics in medical AI studies [22], supporting this rule of thumb. For detection and classification tasks, which make up the majority of medical AI projects, at the very least, sensitivity, specificity and PPV should be reported. For any detection or classification tasks where there is a head-to-head comparison with doctors, presenting these results on a ROC curve is also a good idea, and AUC is commonly used as a primary comparison metric when it can be calculated.

8.3.6 The Analysis

The final element when designing a medical study is to define how you will analyse your results. This typically means identifying which statistical tests you will apply. This goal is to determine how reliable the results are and how likely it is that chance played a role in the results.

There are really two approaches to statistical analysis in medical research, both of which are useful. These are:

- *Estimating metric uncertainty* using confidence intervals.
- *Null hypothesis significance testing* using *P*-values.

These two approaches are intimately related to each other and largely provide the same evidence. However, in general it is good practice to apply both methods in your analysis, which allows the readers to better understand the statistical validity of your results at a glance.

Both methods rely on the same general concept: there is some inherent variability in performance which we can quantify, and by making some simple assumptions about the shape of this variability, we can estimate how likely the results are to arise by chance.

This brings us to our first subtlety; what *type* of variation are we interested in?

Consider the case where we have a dozen doctors who all view the same chest X-rays, and we want to compare how well they detect lung cancer to the performance of our model. Each doctor will have different results, a different level of sensitivity and specificity, for example. So there is *variation across decision-makers*. We also know from thinking about cohort selection that the data used impacts performance, and if a different test set was used, the sensitivity and specificity of each doctor would change. This is *variation across data*. So which of these components of variation do we measure?

The answer comes back to our clinical question (task). In this case, we might decide that the key question is: *"can our model detect possible lung cancer on chest X-rays with higher specificity and equal or higher sensitivity than doctors?"*

Once trained, AI models are static entities when they are applied in practice; they will give the same answer every time they see the same image. As such, variation across models (different decision-makers) is not what we want to measure (even though it is possible to do so [23]). Instead, we care about how the model will perform when it sees new data which may be slightly different from our test set, the variation across data.

Doctors are not a static, homogenous group of identical decision-makers; as mentioned earlier they disagree with each other and with themselves frequently. Performance among doctors can vary widely, as shown in the earlier ROC curve with dermatologists from Esteva et al. Since this effect is likely to be much greater than the variation across data (given a large enough cohort), the variation across decision-makers may be more important here. If there is a need to estimate both types of variation, several methods have been described in the literature [24] although are not often used in practice.

Variation among doctors is a strong reason to use as many doctors as you can for your experiments. Our sample size discussion before applies equally to the number of doctors—if you only test with two or three (which is currently fairly common in the literature), it is reasonably likely that your estimates of doctor performance will be significantly biased, and results in studies like this should be viewed with some scepticism.

We can in fact estimate how different numbers of doctors can influence the sample size needed to achieve statistically reliable results. In Table 8.1, this relationship is demonstrated in several common medical AI scenarios.

This table is unlikely to be directly relevant to most problems due to the quite narrow underlying assumptions, but it provides a useful intuition about how effect size, sample size and the number of doctors/observers interact. In particular, it should be noted how rapidly the required sample size increases as the number of doctors/observers decreases. It is highly unlikely that studies with only two or three doctors will be adequately powered to detect AUC differences below 0.1 in magnitude, even with a cohort containing thousands of cases. We should also make clear that the required sample size increases as interobserver variability increases.

Variation across data can be estimated in many ways, but the most common method is probably the *bootstrap* [26]. This method resamples the

Table 8.1 Estimate of sample size required to detect a suspect difference in AUC, with α = 0.05 and β = 0.80, adapted from [25]. These estimates assume an AUC of around 0.75 and a moderate degree of interobserver variability (±0.025). Fifty percent prevalence is assumed, and as a rough guide if the prevalence is 20%, then the required sample size will be around double the presented estimates

Difference in AUC	Four observers	Six observers	Ten observers
0.05	Not possible	3769	201
0.10	291	78	32
0.15	77	30	20

test data many times, producing an estimate of performance variation over many subsets. This is simple to perform, with easy-to-use implementations in most programming languages, and it provides good estimates of uncertainty. These can be turned into *confidence intervals*, which are a range of values around the performance estimate that cover a defined level of uncertainty. For example, the most commonly used 95% confidence intervals simply show the range of values within two standard deviations of the performance estimate.

To measure variation across decision-makers, simple calculations can be applied to the set of performance results from each of the doctors to calculate a standard deviation. Again, for a 95% confidence interval, just show two standard deviations above and below the performance value.

Now that you have a measure of variance for both the model and the doctors, you can perform *null hypothesis significance testing*. The exact method of doing so will vary depending on your task, but in general you are trying to determine if the difference between the model and the doctors is likely to have occurred by chance. A *P*-value of less than 0.05 means that there is less than a 1 in 20 chance of detecting a difference as big or greater than the one in your results by chance alone, assuming there is actually no difference (i.e. under the null hypothesis). As you may appreciate from this definition, *P*-values and null hypothesis tests can be very complex and confusing and should be applied with care [27–29].

In general, confidence intervals are considered easier to interpret (albeit with their own complexities), but since it is currently a de facto requirement in medical research to supply *P*-values, it is usually a good idea to present both.

One last factor to bear in mind which commonly affects the interpretation and validity of the analysis in medical AI studies is *multiple hypothesis testing*. This is the idea that if you run multiple experiments instead of just one, you are more likely to find spurious results (otherwise known as *false discoveries*). If we have a 1 in 10 chance of getting a result of 40% tails or less in 100 flips of a *fair* coin, then if you run the experiment twice, you now have very close to a 2 in 10 chance. If you run the experiment 20 times, you now have around a 9 in 10 chance of finding at least one result of 40% tails or less.

This does not matter much when you perform the same experiment multiple times (because you can see how many of the results are significant), but the same rules apply if you perform many different experiments. Imagine you are testing a medical AI system that is trained on the NIH chest X-ray dataset [9] (which now identifies 14 different image findings per case). If you test a set of radiologists against your model in each of the 14 categories, the chance of your model spuriously performing better than the humans in one of the categories (with a *P*-value of just below 0.05) is actually about 50%. You would need to find superhuman performance with a *P*-value of just below 0.05 in four or more categories before the probability of finding similar results simply by chance is around 5%.

The obvious solution to this is to present the results for every test you perform. If you run 14 experiments comparing radiologists to an AI system, then you show all 14 results. If only one result is positive, then readers can be sceptical about the findings. This approach does rely on the statistical maturity of the audience, however, and can be misleading because the *P*-values on display will still say “less than 0.05” in each experiment, which to many readers will suggest a low chance of the findings reflecting a false discovery.

A better approach is to perform a statistical correction for multiple hypothesis testing. The simplest is to do a *Bonferroni correction*, which means that we divide the *P*-value threshold we use to determine significance by the number of tests performed. In this case, 0.05/14, which equals 0.0036. With a *P*-value threshold of this level, the probability of one of the experiments demonstrating spurious significant results is around 5%. There are a variety of other solutions including replication of results in independent test sets/cohorts, post hoc correction of *P*-values using the false discovery rate or similar methods and the empirical estimation of *P*-values (by simulation).

There are three other ways that multiple hypotheses testing commonly affects medical AI studies: the use of public datasets, the use of hand-crafted features and data dredging.

Public datasets present an interesting challenge. It does not make any difference statistically whether a set of experiments are performed by one team or multiple teams; every additional experiment increases the chance of finding and publishing false discoveries. For public datasets and competitions, there are often hundreds of teams who have published on the same cohort. It is reasonable to assume that this significantly increases the global false discovery rate, but there has been little research into this effect. No research discipline controls for this type of multiple hypothesis testing at this stage, so all we can do is recognise that there may be an issue and be more sceptical around results on heavily used datasets.

The issue of multiple hypotheses when using hand-crafted features is much more clearly understood, with well-defined solutions. While we typically talk about deep learning models when we are discussing medical imaging AI, there is still a vibrant research community working on older methods. In general these approaches measure an image using a set of image features, such as how bright it is or how much of a certain image texture is present. The numbers are then fed into a simple statistical model, such as a multiple linear regression system. The important lesson from other fields dealing with high dimensional data such as genomics is that when this is done, every single feature should be treated as a separate experiment. It is not uncommon to use thousands of features in these models, which vastly inflates the risk of false discoveries unless this is controlled. A Bonferroni correction or similar method should always be performed in this sort of research, if statistical validity is an important goal.

Finally, there is the issue of data dredging. This is unfortunately very common in science more broadly and is thought to be partly responsible for why so much medical research is not reproducible [30]. Data dredging is where a set of experiments are run, and the question being asked is only decided *after* the results have been seen. Essentially, the team picks the result that looks best. This vastly increases the chances of a false discovery.

While this form of academic dishonesty seems extreme at first glance, if we consider how AI research is performed in public datasets such as the NIH chest X-ray dataset or the ImageNet dataset [31] in machine learning more broadly, where teams tinker with their models until they work better than anything that has come before, we should recognise that this practice is widespread.

The solution to data dredging is pre-commitment. You define your question prior to performing any experiments and do not deviate from that design. If the results are negative that is just the outcome of the experiment.

In medical research where the results are likely to impact patient care, pre-commitment is a mandatory requirement. Teams are expected to publish their study design online prior to even collecting the data, and any deviation from this plan will be challenged at the time of publication. This is not an expectation in medical AI research at this stage, but we should certainly question any results which do not explicitly state the number of tests performed during experimentation and confirm that measuring the primary outcome was a pre-commitment prior to performing the study.

8.4 An Example of Study Design

To bring this all together, let us look at an example. Imagine that we have the clinical problem that we want to be able to determine which patients with chest X-rays have lung cancer.

Clinical Question "can our model detect possible cancer in the lungs on chest X-rays more accurately than five thoracic subspecialist radiologists?"

In this case we choose to use five subspecialists because this is the group of doctors that were locally available, and subspecialists should reflect the upper range of human performance.

You can see that this formulation is very incomplete right now, because we have not worked through the rest of the experimental design. We simply started with the task and the comparison.

Ground Truth "primary or secondary cancer in the lungs proven on biopsy, at the time of study or within 1 year"

This is likely to be a reliable and accurate ground truth, compared to something like "lesions that appeared to be cancer on follow-up CT scan" or "lesions likely to be cancer based on the consensus of a panel of thoracic specialists".

However, there are several possible sources of error in this ground truth. Patients with cancer who were not detected at the time of the original chest X-ray and who did not receive follow-up within 1 year will be missed positive cases. Likewise, patients with cancer who have negative biopsies (because the sensitivity is not 100%) will be missed. We might look at other research and our own data to estimate the false-negative error rate at below 5%.

Given the requirement of biopsy confirmation, the chance of false positives caused by misdiagnosis will be very low. However, if a patient who did not have cancer at the time of the imaging develops a new malignancy in the following year, this would create a false positive under our current formulation. It may be necessary to make this part of the rule more complex to minimise such errors.

Target Population "All patients who have chest X-rays in our hospital"

While we could have narrowed the scope of the target population (and task) to *"all patients having chest X-rays to investigate possible lung cancer"*, the task we are interested in here is to be able to diagnose lung cancer in a chest X-ray taken for any reason. This target population will be able to produce results that can be extrapolated to the hospital it was produced in, as well as other hospitals with a similar patient group. To understand this further, we will need to characterise our population. In particular, it will be required for us to note the distribution of age, sex, ethnicity and smoking history in the population. Smoking history is particularly important because it influences not only cancer risk but also what types of cancer are likely to arise (which will affect what they look like and how well our model will detect them).

We need to recognise that these results may be less reliable if the model will be applied in a community radiology clinic, since this population is likely to be significantly different from that of a hospital. Similarly, generalising the results to a different country with a different population and a different healthcare system should only be done with caution and may not be possible at all.

We should also understand that we have made our task more difficult by using this population compared to, for instance, an outpatient group from an oncology clinic, as the ability to visually detect cancer in patients with severe lung disease (such as pneumonia, fibrosis, pneumothorax and so on) is much more limited. Patients with these conditions will definitely occur within this population.

Cohort "A random sample of the population, with similar demographic characteristics"

Random sampling with a check to ensure a similar distribution of patient characteristics will prevent any significant sampling bias. The question then is how big should our cohort be?

Our rule of thumb says we need at least 100 cases of cancer in the test set. This is going to be very difficult to achieve, because the rate of cancer in the lungs on chest X-ray is very low in the general hospital population. If it is below 0.5% per study, then we will need 20,000 X-rays just to get enough data for the test set. We might need 10 times more than this to build a training set that is likely to produce a high-performance model.

For the head-to-head testing with doctors, we will also randomly select 100 cases without cancer. To allow for a broader set of metrics to test the performance of the AI system, we will randomly select a much larger set of negative cases to test the system at clinical prevalence. Since prevalence is 0.5%, we will select 19,900 negative cases.

Given the size of the cohort and the number of thoracic radiologists we will test, we estimate that it is plausible to detect a difference as small as 0.1 in a metric like sensitivity (with $\alpha = 0.05$ and $\beta = 0.80$). This means that a study of this size *will not* be informative if the model is worse than the radiologists by a difference of less than 0.1, even if reducing performance by 0.05 may be highly relevant clinically. In this study we must acknowledge this limitation and clearly state in any publications that further experiments may be required to exclude smaller differences.

Metrics "The primary outcome will be the sensitivity of the model compared to the average sensitivity of the doctors, at the points closest to the highest achieved human specificity, and average human specificity. We will also present the AUC and precision (PPV) of the model at the selected operating point in the second test set".

Since this is a classification study, ROC-AUC would be a good metric to use but is difficult to calculate for the human group because this task does not lend itself to using a Likert scale (we could ask the doctors to rate the risk of malignancy on a five-point scale, but this is not something they are experienced at doing in clinical practice, which will bias the results against the radiologists).

Since the task is cancer detection, we will use sensitivity as our primary metric. This is a clinical choice; we made the decision that false negatives (missing cancers) are worse than false positives (overdiagnosis). We present results at two operating points, acknowledging that average human specificity is likely to underestimate human performance and peak human specificity may be an overestimate.

Presenting both prevalence-variant and invariant metrics in the second test set will demonstrate to readers the types of errors the system makes in a clinical population.

Analysis "We will use bootstrapped confidence intervals for the model and the exact Clopper-Pearson method to produce confidence intervals for the average human sensitivity. We will perform null hypothesis significance testing with McNemar's chi-square test".

In this case, there is little risk of multiple hypothesis testing, since we only have two hypotheses. We can pre-commit to the study as described in the final study design, and in doing so, we will have made the results of the analysis as reliable as they can be, within the given margins of error in the analysis.

Final Design We can draw it all together now. "Our model will demonstrate a higher sensitivity than the average of a team of five thoracic radiologists at both average and peak human specificity operating points, at the task of detecting malignancy in the lungs on chest X-ray, proven on biopsy at the time of study or within 1 year. We pre-commit to testing the above metrics only once on the test data, as our primary outcome. We will present further metrics on a second larger test set at clinical prevalence, without the human comparison".

8.5 Assessing Safety in Medical AI

At the start of this chapter, it was stated that safety is more important than performance in medical studies. This may seem strange, considering that all of the following material has mostly been focused on performance testing. There is a very good reason for this. However, we do not currently assess safety very well in contemporary medical AI studies or even at the regulatory level.

The current method of assessing safety is simply to compare a system against a method currently used in practice, tested in a retrospective cohort study. A classic example is to take a cohort of patients with and without a disease and have both doctors and an AI system assess the images. If the AI system performs as well as the doctors on a given set of metrics, this is assumed to mean that the system is safe.

There are two major problems here: that the model may perform differently in practice than in a controlled experiment and that the model may perform differently in a different clinical setting. A single retrospective cohort study cannot assess either of these risks.

Radiologists are well aware of the risks of retrospective cohort studies. The first computer-aided diagnostic (CAD) device for screening mammography was approved by the FDA in 1998. Breast CAD was shown in modestly sized multi-reader multicentre studies to improve cancer detection rates by between 2 and 10%, but these systems have not only failed to achieve these results in practice in the decades since but may have in fact *reduced* cancer detection rates and increased false positives [32]. One explanation suggested in the literature is that radiologists with different levels of experience used CAD in different ways. While the more experienced radiologists (such as those in the early studies) were unlikely to be swayed by the system, less experienced readers could be lulled into a false sense of security. Others found that breast CAD underperformed in certain forms of cancer (non-calcified lesions), and this stratification had not been taken into account in the earlier studies.

These sort of unexpected problems that only arise as technologies are applied in clinics are exactly what we need to be wary of with medical AI. If an AI system replaces one small part of a complex chain of clinical care, how does every other part of that chain react? Will experts provide adequate oversight of the system, or when they are overloaded with work, will they place too much trust in a model that appears to perform well almost all of the time? Does having a machine supply you a diagnosis lead to cognitive biases like satisfaction syndrome, where the radiologist may be less likely to find a second pathology simply because they have already got an answer? Will these systems that perform well in retrospective studies also work well in clinical practice?

In the case of breast CAD, the answer has been no. The price paid for the use of breast CAD has been detecting less cancer, performing unnecessary painful and invasive breast biopsies and spending an additional $400 million per year in the USA alone.

We know from other areas of machine learning research that unexpected problems arising from external forces are very real. Almost every self-driving car company has now abandoned plans to release “level 3” technology, where a human driver must be ready to take control at any time if the car does something wrong. In 2015 Google reported they had found in testing that “people trust technology very quickly once they see it works. As a result, it’s difficult for them to dip in and out of the task of driving when they are encouraged to switch off and relax” [33]. This is called “the hand-off problem”, and while they did not mention any accidents caused by this behaviour at the time, it is not just a hypothetical risk. There have been two self-driving car-related fatalities in recent years, and driver inattention while operating level 3 autonomous systems was implicated in both cases.

Thus far, we have no idea if similar problems will occur with modern medical AI, as we have yet to see medical AI systems applied to real patients almost anywhere. Of the few systems that are being used in clinics, we have not seen any results that purport to demonstrate safety “in the wild”.

In statistics this is known as the problem of causal inference; we want to know how patient outcomes change when the AI system is applied to them. The way to demonstrate this in medical research is by using a different type of clinical study design—a prospective randomised control trial (RCT).

In an RCT, we randomly assign actual patients to either a treatment arm (i.e. applying the AI system in their care) or a control arm (managing them without the AI system). By doing this, not only can we detect if there is a difference in clinical outcomes for the patients (which our earlier metrics cannot show), but we can actually infer causation; we can show the AI system is responsible for these changes. Ideally, an RCT would be performed in several clinical settings, also demonstrating how well the model generalises to other patient groups.

Thus far, no RCT has been performed to test a medical AI system, and regulatory bodies do not currently require companies to perform them prior to approval of their AI systems. Instead they accept the results of retrospective studies and require post-market testing, meaning the companies have to continue to report results once the systems are in clinical use.

A second option which is less powerful than an RCT but that could be reassuring regarding performance and safety is to perform an external validation of the results, i.e. replication. This would involve applying the model to one or more independent cohorts. This is very common in medical research, with external validations often performed in other countries to ensure there is a distinct patient group.

If we consider the spectrum of AI risk presented in Fig. 8.1, regulatory approval until now has been limited to systems to the left of the dotted line with only one exception. The IDx system which the FDA approved to assess the health of the retina in the eye in the hands of non-expert health professionals.

There is certainly a decent argument to be made that systems that carry low risk of harm should not need to be tested in an RCT or even an external validation but instead are very similar to previous medical devices. The regulation of these devices in the past has shown that this approach is usually safe, even if examples such as breast CAD may point to weaknesses in this system. But as we see more and more systems at the high-risk end of the spectrum emerge, it is certainly timely to consider the lessons medical research has taught about assumptions of safety and whether we need to seriously consider the role of RCTs and external validation in medical AI studies.

8.6 Take-Home Points

This chapter has covered the basics of how to test the efficacy and safety of medical AI systems in a way that will be acceptable to medical professionals and will provide some degree of confidence in the results. Let us synthesise everything here into a checklist for the design, reporting and assessment of medical AI studies.

Checklist

- Clinical question: Identify what the task is and what comparison you will use.
- Ground truth: Plan how you will discover the ground-truth labels for your data, ideally using objective, error-free methods with a minimal amount of missing data. Check the cases manually (or a random subset of them) to ensure the ground truth is as accurate as you expect it to be. You should report an estimate of ground-truth accuracy in any publications.
- Target population: Identify the target population based on the role you want the system to perform and then characterise that population. Is your training set a good proxy for the population, or do you need to reference other research that characterises the patients in this task? Describe age, sex and ethnicity at minimum (if available).
- Cohort: Ideally you will use a large random subset of the population. Aim for at least 100 cases in your smallest stratum. Given stratification, low disease prevalence, small effect sizes and (if relevant) the humans the number of observers, your final sample size will often need to be thousands of cases. Where possible, perform a power calculation that includes both estimates of sample variation and human expert variation to ensure that your cohort size is adequate.

- Metrics: While these are task dependent, you should try to make clear what *kind* of errors your model makes. For example, in classification tasks, use metrics that explain the false positives, false negatives and the role of prevalence. A good starting set of metrics to consider is sensitivity (recall), specificity and positive predictive value (precision). AUC is a great overall metric to use in classification/diagnostic tasks.
- Analyses: Pre-commit to measuring a single quantity where possible. Do not use your test set for any similar experiments until this has been tested. Provide estimates of uncertainty such as a *P*-value or confidence interval (preferably both).

  If you are testing multiple hypotheses in your experiments (e.g. in research using hand-crafted image features in a radiomics framework), present all of the results rather than cherry-picking the good ones. Consider applying a statistical control for multiple hypotheses, such as the Bonferroni correction.
- Safety: Consider the level of risk for when the model is applied to actual patients? If this risk is high, does an external validation or RCT need to be performed prior to marketing?

References

1. Giger ML. Machine learning in medical imaging. J Am Coll Radiol. 2018;15:512–20.
2. Harris S. Record year for investment in medical imaging AI companies. 2017. <https://www.signifyresearch.net/medical-imaging/record-year-investment-medical-imaging-ai-companies/>
3. Petryna A. When experiments travel: clinical trials and the global search for human subjects. Princeton, NJ: Princeton University Press; 2009.
4. Simonite T. Google's AI doctor gets ready to go to work in India. 2017. <https://www.wired.com/2017/06/googles-ai-eye-doctor-gets-ready-go-work-india/>
5. Enlitic. Enlitic to partner with Paiyipai to deploy deep learning in health check centers across China. 2017. <https://www.prnewswire.com/news-releases/enlitic-to-partner-with-paiyipai-to-deploy-deep-learning-in-health-check-centers-across-china-300433790.html>
6. U.S. Food and Drug Administration. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. 2018. <https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm604357.htm>
7. Euser AM, Zoccali C, Jager KJ, Dekker FW. Cohort studies: prospective versus retrospective. Nephron Clin Pract. 2009;113:c214–7.
8. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. Philadelphia, PA: Wolters Kluwer Health; 2008.
9. Wang X, et al. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). IEEE; 2017. p. 3462–3471.
10. Anderson E, Muir B, Walsh J, Kirkpatrick A. The efficacy of double reading mammograms in breast screening. Clin Radiol. 1994;49:248–51.
11. Manrai AK, Patel CJ, Ioannidis JP. In the era of precision medicine and big data, who is normal? JAMA. 2018;319:1981–2.
12. Gulshan V, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316:2402–10.
13. Punjabi NM. The epidemiology of adult obstructive sleep apnea. Proc Am Thorac Soc. 2008;5:136–43.
14. Ogasawara KK. Variation in fetal ultrasound biometry based on differences in fetal ethnicity. Am J Obstet Gynecol. 2009;200:676. e671–4.
15. Shipp TD, Bromley B, Mascola M, Benacerraf B. Variation in fetal femur length with respect to maternal race. J Ultrasound Med. 2001;20:141–4.
16. BBC News. Google apologises for Photos app's racist blunder. 2015. <http://www.bbc.com/news/technology-33347866>
17. Agarwala A. Automatic photography with google clips. 2018. <https://ai.googleblog.com/2018/05/automatic-photography-with-google-clips.html>
18. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med. 1994;331:1493–9.
19. Esteva A, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115–8.
20. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29–36.
21. Shiraishi J, Pesce LL, Metz CE, Doi K. Experimental design and data analysis in receiver operating characteristic studies: lessons learned from reports in radiology from 1997 to 2006. Radiology. 2009;253:822–30.
22. U.S. Food and Drug Administration. Software as a medical device: clinical evaluation. 2017. <https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm524904.pdf>
23. Gal Y, Ghahramani Z. In: International conference on machine learning. 2016. p. 1050–1059.
24. Obuchowski NA, et al. Multireader, multicase receiver operating characteristic analysis:: an

empirical comparison of five methods. Acad Radiol. 2004;11:980–95.

25. Obuchowski NA. Sample size tables for receiver operating characteristic studies. Am J Roentgenol. 2000;175:603–8.
26. Efron B. Bootstrap methods: another look at the jackknife. In: Kotz S, Johnson NL, editors. Breakthroughs in statistics. New York, NY: Springer; 1992. p. 569–93.
27. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat. 2016;70:129–33.
28. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle P value generates irreproducible results. Nat Methods. 2015;12:179.
29. Ioannidis JP. The proposal to lower P value thresholds to .005. JAMA. 2018;319:1429–30.
30. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2:e124.
31. Deng J, et al. In: IEEE conference on computer vision and pattern recognition, 2009. CVPR 2009. IEEE; 2009. p. 248–255.
32. Kohli A, Jha S. Why CAD failed in mammography. J Am Coll Radiol. 2018;15:535–7.
33. Google. Google self-driving car project monthly report. 2015. <https://static.googleusercontent.com/media/www.google.com/en//selfdrivingcar/files/reports/report-1015.pdf>

Part IV

Big Data in Medicine

Enterprise Imaging

9

Peter Mildenberger

9.1 Introduction

The first picture archiving and communication systems (PACS) have been implemented in radiology more than 20 years ago. This has been supported by introducing the DICOM standard in 1993. Today, all imaging modalities in radiology are digital. Film-based reading of studies is out of date. Large radiological departments produce several terabytes of image data every year.

Along with the introduction of PACS in radiology, the interest in digital image management started in several other clinical professions. Cardiology has been one of the first departments outside radiology using the similar DICOM image objects as radiology, esp. for angiography. The introduction in DICOM for imaging outside radiology is obvious with the publication of the supplement 15 “Visible Light Image Object” in 1998. Other examples for the adoption of DICOM are dental medicine or ophthalmology, which required new image objects too. The most advanced and challenging profession is probably pathology, because the new scanners for whole slide imaging are fast enough to support regular workflow in pathology but also produce a much higher data volume than other clinical applications [1]. It is expected that pathology departments will end up with several petabytes per year.

In the early phase of PACS, it has been usual to have separate solutions for different clinical applications based on departmental image acquisition and storage, e.g. for ultrasound studies, cardiological imaging, ophthalmology, etc. This approach had several disadvantages, of course the costs for maintaining several systems in parallel are much higher compared with centralized management, but mainly the need for communication and access to the information inside one hospital or across different enterprises required a harmonization of image management with common storage of image data and universal image viewing capabilities. Actual image management systems are able to handle DICOM and non-DICOM data (e.g. PDF, ECG, movies) in a neutral form to different applications (also called Vendor Neutral Archive or VNA); this means that a central, universal archive is serving for different, specialized applications and should be accessible by a universal viewer [2].

Therefore, it is very much accepted and part of IT strategies to have a harmonized image management for all or most of the different image sources in healthcare institutions. This will become more relevant with integration of mobile devices and/or patient-generated images

P. Mildenberger (✉)
Department of Radiology, University Medical Center of Johannes Gutenberg-Universität Mainz, Mainz, Germany
e-mail: mildenbe@uni-mainz.de

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_9

from outside the hospital itself, which should be feeded into the imaging environment.

Such repositories could be understood as "Big Data". In regard to artificial intelligence applications, it sounds very reasonable that these data are a real value for research and the development of new self-learning algorithms. However, most of these data are unstructured, even not or eventually wrong classified, e.g. using scanning protocols from different body region than the region examined.

9.2 Basic Principles of Enterprise Imaging (EI)

In healthcare provider institutions today, it is expected to have access to almost all clinical information as part of an electronic health record (EHR). Linked with such an EHR, there should be also a solution for accessing clinical imaging and multimedia data through one integrated platform. This lesson has been learned while implementing such IT systems in the beginning as separated departmental solutions but finding that there are similar technical implementations and clinical needs. The improved request for clinical conferences or multidisciplinary team meetings fosters the harmonization of systems and provision of integrated image access to different imaging sources. Of course, there are different requirements in different clinical scenarios. Workflow is one of these challenges, because the well-known and standardized radiological workflow is not fulfilling the need in other professions, in which a more encounter-based image capturing is usual, e.g. clinical photography. Another part of this discussion is the protection of patient rights, granting access rights, etc.; this part of regulation is even more sensitive and relevant for image exchange across hospital boundaries.

In 2016, a joint initiative of HIMSS and SIIM published a series of collaborative white papers on enterprise imaging. Based on this, it is evident that a successful enterprise imaging program should include a strategy with consistent workflows to "optimal capture, index, manage, store, distribute, view, exchange, and analyse all clinical imaging and multimedia content to enhance the electronic health record" [2]. This HIMSS-SIIM collaborative workgroup has identified several key elements for a successful enterprise imaging program, which are as follows:

- Governance
- Enterprise imaging strategy
- Enterprise imaging platform (infrastructure)
- Clinical images and multimedia content
- EHR enterprise viewer
- Image exchange services
- Image analytics

An appropriate and effective governance for enterprise imaging programs should guarantee the involvement of all relevant stakeholders from the different clinical departments, administration and IT experts. As part of this governance, regulation for shared handling of personnel, infrastructure and knowledge is crucial, especially when there has been no culture for communication between departments in place in place before.

The implementation of enterprise imaging should be based on an IT strategy, which is accepted by the leadership and in line with the governance. Such an enterprise imaging program might start with a very limited number of clinical partners but should have the strategic perspective and needs from all potential partners in mind. This is especially relevant for the decision-making in the storage and viewing capabilities, e.g. for new imaging objects or handling of special objects like multiframe or 4D—studies or ECG measurements. As part of switching from decentralized silos to a common platform, restructuring of IT infrastructure will be necessary. This could require restructuring of IT support too. As this could be a difficult discussion, a strong governance is very helpful for such a process [2, 3].

9.3 Enterprise Imaging Platform

An enterprise imaging platform should provide different functionalities. There is the central core with storage, interfacing imaging devices and providing worklist services, if not done by EHR. Interoperability with the EHR requires standards-based recommendation based on ADT messages (Admit Discharge Transfer, a core HL7 message type), orders, providing results back to the EHR, storing images through the EHR and communication for viewing imaging studies and handling metadata. On the other side, the enterprise imaging platform has to serve all different image sources, which could be DICOM-based (as most of the devices in hospitals today) or non-DICOM-based, which is relevant, e.g. for mobile applications or video data. Additionally, image exchange with external partners is more and more important; this includes media-based solutions (e.g. CD import) or health information exchange (HIE) or telemedicine solutions. Usual standards are DICOM, HL7 and web services. For cross-enterprise image exchange, the IHE-based family of XDS profiles (Cross-Enterprise Document Exchange) is very common. Mobile- and web-based applications are available based on FHIR (Fast Healthcare Interoperability Resources), a new standard series by HL7. Besides the imaging functionalities, there are critical factors as reliable and efficient retrieval services. Secure access control and auditing is another requirement, as providing solutions for disaster recovery and business continuity in case of maintenance or technical failures. Such a central repository could be built on an existing PACS, if such a solution fulfils the requirements for an enterprise-wide application [2, 4].

Compared with the traditional PACS workflow, there are new and different requirements in an enterprise imaging environment. Traditionally, imaging is mostly used to solve differential diagnosis in a diagnostic process, which is independent from the profession. Implementing enterprise imaging and enabling all different devices add new and different workflows to the traditional, DICOM-based PACS processes. Especially mobile devices will be used to capture status of a clinical finding, other documents of for quality assurance. This might be very often encounter based and not DICOM images. In the context with storing and accessing such studies, there are new challenges, because the categorization will not work by information like modality, name of imaging protocol or image-object definition. Based on imaging source, imaging type and operational workflow, the HIMSS-SIIM workgroup proposes four broad categories: diagnostic imaging, procedural imaging, evidence imaging and image-based clinical reports [2]. In this concept, the **procedural imaging** is relevant for documentation of therapeutic procedures, which might be percutaneous (e.g. stent placement) or surgical processes. **Evidence imaging** is dedicated to the documentation of clinical findings, e.g. size and location of polyp during endoscopy, types of injuries on visible light images, or secondary capture of a 3D-reconstruction. Image-based clinical reports are defined as a concurrent delivery of images and textual information, which provides links between imaging findings and textual reports or schematic diagrams [2].

The deployment of an EI platform in a hospital will impact several workflows in many departments. While radiology and cardiology do have created automated workflows for acquisition, storage and image distribution, this is often not adopted in other clinical professions with different, mostly lower number of imaging studies. This could lead to inefficient processes and limit the throughput and also acceptance of EI solutions. Conventional workflows, as in radiology or cardiology, are based on orders, which drive automatization of workflows including unique study identifiers or worklists for technicians and for reporting [5]. In many other clinical scenarios, there are no orders and an encounter-based workflow is in place, e.g. in evidence imaging as photo documentation. In such situations, correct patient identification must be secured. By using a DICOM worklist, there is automatically the metadata set to identify patient, imaging procedure, etc. This is missing in encounter-based workflows. Different approaches are known to solve this task. Some institutions are using workflow reminders, with the intention to capture patient data before the first and after the last clinical image. This is depending on human interaction, and it could fail. Another approach is sticking barcode on every image. This can be effective and prevent misidentification, but—as for ultrasound images with names burned in the scan—it might impact further handling for other purposes, e.g. research or teaching application, if the patient identification is clearly accessible on every image and cannot be easily anonymised. There are newer developments providing digital cameras with support for DICOM worklist access and DICOM image objects, this might be a solution in different scenarios. Other technical requirements for workflow support include the ability to perform standard measurements or colour standardization. Reliable patient positioning could be relevant for follow-up studies to provide the ability to find changes in finding, e.g. size of dermatological finding or orthopaedic joint flexibility. Landmarks, time points and location of patient positioning should be part of an imaging record [5, 6].

Workflow support for reporting might be another new aspect in some clinical departments. While radiology, cardiology or pathology are well trained to make a report for any imaging study, this might be different elsewhere. It is known that several imaging studies are stored but without annotations or reports so far. This could have different reasons, like behaviour or missing tools for efficient reporting in EHR. Otherwise, it is crucial for communication in shared decision-making for medical treatment to have all information available, including images and textual information. This is even more relevant in the context of research or image analytics within AI applications, because images without appropriate information could not be used efficiently for self-learning algorithms. Efficient workflow support should enable links between information from the EHR and imaging studies in the EI platform in a bidirectional way, which means that an access to imaging should be provided from the EHR and also an easy access to reports should be available through the enterprise viewer, and also this viewer should grant access to imaging from different departments for one medical scenario, e.g. displaying diagnostic imaging, intraoperative photo documentation and pathology slides side by side [5].

In an EI environment, metadata are very relevant for structuring the information and efficient access to imaging studies, e.g. showing all studies from one department or all studies for one body part. Metadata are well structured and defined for DICOM imaging objects [7], including study descriptions, information on body part examined or departmental information. The DICOM standard provides a catalogue for body parts, however this is limited and potentially to specific of not enough specific in other imaging domains. Also, there are workflow issues, which, e.g. imaging of several body regions (e.g. chest and abdomen in one scan), might be classified just with one region or misclassification based on the application of study protocols from other body parts, e.g. MR sequences for brain study used in abdominal imaging without correction/adoption of the body parts. It is recommended to have an ontology in place, which should allow synonyms as also provides a relational structure with parent and children terms. In radiology, there is RadLex as an ontology and RadLex Playbook for procedure description available, both initiated and provided by RSNA. Aggregation of information across institutions requires mapping of procedure descriptions on a standard description; ACR has addressed this in the context of the ACR dose registry using the RadLex Playbook descriptions [5, 8].

Enterprise Imaging should provide a standardized access to all imaging studies integrated in the EHR. There are many reasons to do so. A simple one is improved usability with single-sign on through the EHR, similarly this could be used for granting access rights to the imaging studies, because this is already defined and regulated as part of the EHR. An EHR-integrated universal viewer should handle different imaging data, which could be simple photo documentation up to functional imaging in MRI or whole-slide pathology scans, because one standard interface will improve the acceptance and usability for the different user groups in an enterprise. This spectrum of imaging data is very broad, including

DICOM objects, point-of-care documentation and sometimes also scanned documents. Therefore, there is an increasing interest to fulfil the different requirements with a universal multipurpose application, called "Enterprise Viewer" [9]. The HIMSS-SIIM workgroup proposes the following definition of an enterprise viewer: "thin-client or zero-client application used on any off-the-shelf device to distribute, display, and manipulate multi-specialty image, video, audio, and scanned documents stored in separate centralized archives through, or standalone from, the EHR." [9]. This group divides the users into four main groups:

- Users performing diagnostic interpretation needing most advanced image manipulation and reporting capabilities
- Surgical subspecialists using image manipulation for planning procedures
- General providers and non-providers needing access to basic image viewing tools
- External users such as patients or referrers

Enterprise image viewers could solve several clinical use cases, e.g. access to different archives, integrated view on studies from different clinical professions, collaboration providing teleconference functionalities and diagnostic interpretation without a dedicated PACS. Beyond in-house use cases, such solutions could support access for on-call physicians, patient-portals, referring physician access or also educational purposes [9].

Technical considerations for the deployment of enterprise viewers include optimized installation and support resources, because the classical structure of PACS environments with dedicated workstations, which requires installation, updates, maintenance, etc., would not work in an enterprise with probably several thousands of viewing devices. Several other technical aspects are relevant for enterprise viewers. Viewers should run on different devices, independent from the operating system or kind of display (e.g. workstation, tablet, smartphone) and without transfer of full-DICOM files. Therefore, most enterprise viewers are HTML5 based and do have rendering servers and transfer DICOM images into other formats, which might allow faster access and appropriate rendering depending on the display size and zoom factor. Security aspects should be encountered too, e.g. secure connection to rendering servers and if mobile devices get lost. To prevent data leaks, enterprise viewers should support built-in encryption and lifecycle management policies with immediate deletion of image content after its presentation. Access to the enterprise imaging platform could be directly from the enterprise viewer or through the EHR. Direct access requires full management of users and role-based access rights. Therefore, in most cases an EHR-based access would allow similar usability with built-in access permissions. Audit records for user activities could be handled using the IHE ATNA (Audit Trail and Node Authentication) profile [9].

Health information exchange is getting more and more relevant in modern healthcare systems, which rely on information exchange especially for imaging studies. This exchange is necessary from provider to another, from provider to patient and from patient to patient. Also patient portals with the option to download and transfer images for second opinions are used or in wider deployment. An EI platform should be prepared to support all these kind of interfacing inside and outside the hospital walls. Online cross-enterprise image transfer (and also document and other data formats) is mostly based on the concepts of IHE with XDS and XDS-I. In Europe, there are some countries building national solutions based on these IHE profiles, e.g. Austria, Switzerland and Luxembourg. Centralizing such exchange capabilities in a hospital, which would eliminate CD/DVD import and export, could enhance the data security and data privacy protection and also reduce the training and failure rates [2].

EI platforms will provide large data collections; however, the opportunities for the application of business and clinical applications are still limited or immature. Of course, there are a lot of metadata available, but definitions of value sets and standardization of these data are still in an early phase. But such annotations and semantic interoperability will be required for more advanced analyses on imaging utilization or the application of deep learning algorithms. This will certainly change in the future, due to the ongoing developments in artificial intelligence, which will play an increasingly important role in healthcare [2, 10].

9.4 Standards and Technology for an Enterprise Imaging Platform and Image Sharing Across Enterprises

An enterprise platform should support several standards beyond DICOM. This is relevant due to the inclusion of photography, movies, waveform documents, etc. There is a challenge in supporting different vendors, different standards and different interfaces—but interoperability is a prerequisite for a successful EI platform. Due to different use cases and workflows, it could be necessary to support different standards for similar tasks, e.g. photo documentation with dedicated DICOM devices or mobile devices only supporting JPEG or PNG formats. The central core component is—according to the HIMSS-SIIM concept—an Enterprise Image Repository (EIR), which supports DICOM and non-DICOM imaging objects including an index of the images and content of the archive, along with the metadata of these imaging studies [11]. An enterprise viewer can be an integrated part of this core component, provided by the same vendor, but could be in principle an independent solution interfacing with the image repository.

Additionally, there should be standard interfaces for all different imaging devices/formats and also services like worklist provider, auditing, image exchange, media import, etc. Standard interfaces include DICOM, DICOMweb and XDS-I.b, which would support many imaging sources across an enterprise. Management of scanned documents in EHR environments, e.g. forms, consents and external reports, is often handled via an enterprise content management (ECM) system. Such systems are sometimes already in place, so it will be necessary to discuss an upgrade from ECM to Enterprise Image Repository (EIR). This might be a difficult decision, usually ECM is not prepared to handle the amount of images in efficient manner and guarantee acceptable access performance. It might be easier to include the handling of scanned documents into an EIR. Different use cases are described in the HIMSS-SIIM concept [11]:

- Native DICOM devices with worklist support are almost in use in radiology, cardiology, and dental medicine or ophthalmology or ultrasound machines. Image objects are well defined, and metadata are available within these imaging sources.
- Endoscopy, microscopy and digital cameras are part of the “visible light” group. Typically, such imaging sources are not supporting DICOM directly but could be interfaced with dedicated software. Accurate patient demographics can be obtained by an order-based workflow with worklist or an encounter-based workflow using DICOM query services, HL7 message or by using IHE PIX and PDQ profiles.
- Capturing of medical photos or videos is probably the most challenging use case for technical support to guarantee correct integration of patient demographics, creating orders and links for accessing such studies from the EHR, etc. Functionality and usability depend on the import software; an easy way to solve such use cases is “wrapping” these non-DICOM images while importing into an encapsulated DICOM object. As an alternative, handling as a “document” according to the IHE XDS concept could be implemented.
- Image management for importing external images or sharing imaging studies with external partners should be addressed by the enterprise imaging platform too. This is including media support based on IHE PDI (Portable Data for Imaging) and Import Reconciliation Workflow (IRWF) to adopt internal patient demographics and orders.

In the context of handling all these different formats, it is evident that in principle, most of these non-DICOM data could be easily handled as DICOM encapsulated objects, which would solve several issues like handling of metadata, integration with XDS-based image exchange solutions, etc. A valid alternative is to encapsulate non-DICOM data in a HL7 CDA format, as it is used in IHE XDS-SD (Scanned Document). A very detailed review on different technical aspects is provided in the HIMSS-SIIM paper on "Technical Challenges of Enterprise Imaging" [11].

Any enterprise imaging platform should be able to support improved collaboration and patient care with external physicians and patients themselves. This is relevant to increase patient empowerment and patient satisfaction and to improve cost efficiency by avoiding unnecessary repeated imaging. Several use cases are common, e.g. transfer to trauma centres in emergency cases; telemedicine, esp. teleradiology or telepathology, to allow access to specialist in rural areas; remote wound management monitoring; second opinion consultation, etc.

In the past, proprietary solutions for point-to-point communication or legacy web-based solutions have been introduced to solve such requirements. With growing networks of partners, interoperability of systems became a key factor. It is obvious that this is addressed in a very successful approach worldwide by the IHE XDS (Cross-Enterprise Document Exchange) concept, which includes a whole family of profiles supporting mapping and accessing of patient demographics, patient consent, scanned documents, imaging studies, lab results, medication, etc. The principle concept of IHE consists of a so-called affinity domain, which represents one hospital or group of partners, which will provide a registry of documents and imaging studies available for exchange. The registry contains links to the decentralized data repositories. In this context, the enterprise imaging repository is usually an IHE imaging document source, while the enterprise viewer could (and should be able to) act as an IHE XDS-I document consumer. There are profiles available to connect different affinity domains and allow cross-community access and also profiles for advanced handling of patient consent, which are relevant to grant access for external users. Support for efficient access with mobile devices is enabled by newer IHE profiles like MHD and MHD-I (Mobile access to Health Documents); these profiles are using newer web standards such as FHIR or DICOMweb, a RESTful standard. Of course, such networks will raise many more issues to be solved, like responsibilities, policies for storing external data in local PACS or provide secondary readings for externals imaging studies [10, 12–14].

9.5 Legal Aspects

It is obvious that information exchange and communication is a key factor in modern healthcare involving experts from different faculties. Several legal aspects have to be discussed while creating the EI governance.

As healthcare information are sensitive data, data protection and privacy should be considered too. Patient privacy and access control are relevant for several data, e.g. information on psychiatric therapy or plastic surgery and images related to sexual assault or child abuse. It is recommended to have one leading system managing access control, which could be based on general items, e.g. access permission for a specific department or on a case-by-case basis, which would require a reason why a user needs access to the imaging study. This type of information has to be recorded and audited later on.

Such access restrictions have to be managed also for image sharing with external partners, and in such communication a stepwise approach with additional permission/consent regulations might be necessary. According to European regulations, the patient consent is mandatory for image exchange, and patients could select data and providers to be included or excluded in communication [2, 15].

Usually it is possible to use anonymised data in research and education. It is important to keep in mind that for several imaging data, anonymization alone is not good enough. For example, sharing CT or MRI data of brain examinations could probably allow full 3D reconstruction of the face and reidentification of a person.

Several other legal or regulatory aspects should be encountered for an EI project. The duration of storing images in an archive is mostly regulated for radiology, but other fields do have a broad spectrum of workflows. Also, the selection of images is a sensitive factor, e.g. rejection of mistakes, selection of distinct images in ultrasound and storing full video versus selection of sequences. Image quality is another example. Several years ago, there has been an intensive discussion on using compression, especially to save storage costs. The need for "lossy" compression has been decreased with the availability of high-quality storage at relatively low costs over time. The European Society of Radiology (ESR) has published a consensus paper several years ago, describing in which use cases "lossy" or irreversible compression might be used, e.g. long-term archiving or teleradiology, and in which not [16]. File format could be a critical issue, if specific definitions would be used, for which no maintenance over time or broad support by viewers is guaranteed. Wrapping such non-DICOM images into a DICOM format could solve such issues. An institution should define the integration of mobile devices into an EI platform. Mobile devices could support the capturing of images, as also the access to an EI platform or communication on reports. Using individual, personal devices should be carefully examined and implemented, due to potential data privacy risks. The HIMSS-SIIM workgroup has listed a lot of features, which should be supported by a mobile device application including, e.g. query worklists, scan barcodes, no local storage of patient demographics, etc. [5].

9.6 Enterprise Imaging in the Context of Artificial Intelligence

In an ideal world, enterprise imaging platforms would collect many different imaging studies with annotations and reports across different medical faculties and should be linked with EHR and other data, e.g. lab results, patient-generated information, etc. But, this is not the reality today, even not in advanced healthcare institutions.

There is an increasing interest and need to facilitate interoperability of different data repositories for different reasons, improving patient care and quality for clinical reasons, and the further development and application of artificial intelligence tools are another one.

The number of medical image analyses is growing exponentially, as are the tools available for this purpose, based upon machine learning (ML) techniques [17–21].

Some tools are dedicated to extract features out of the imaging studies itself; this process is described under the term radiomics. For further improvements of such research activities, it is mandatory to have a huge number of imaging studies linked with further information on the imaging itself and other clinical data. However there are different drawbacks regarding ML or DL applications for medical imaging today, and some examples here fore are as follows:

- Usually there is no annotation in the imaging studies, indicating which images and in what area these images are affected.
- There is limited standardization in description of anatomical regions.
- There is limited standardization in procedure descriptions, even in the same institution, there could be different descriptions using different content or synonyms for similar procedures.
- Imaging protocols differ markedly for similar clinical indications.
- Imaging reports do not provide coding for findings or are not well structured and provide narrative text only.
- Imaging is stored without metadata and/or reports at all, etc.

Annotation of findings has been addressed by the RSNA in a project called "The Annotation and Image Mark-up Project" providing a standardized semantic interoperability model for annotations and markup. Different information, like author, time stamp and image(s), represent the finding. Information could be kept in semantically precise and computationally accessible manner using the RadLex terminology [22].

A further step has been implemented with the inauguration of “QIBA”(Quantitative Imaging Biomarker Alliance) addressing the development of hardware and software to achieve accurate and reproducible results from imaging methods, including acquisition protocols, data analyses, display methods and reporting structures [23].

Standardization of procedure descriptions is a key factor to build huge data collections for machine learning algorithms; mapping to terms from the RadLex Playbook is one accepted approach. Adoption and acceptance of this approach are still somewhat limited.

Machine learning (ML), deep learning (DL) and convolutional neural networks (CNN) are different methods for creating computer-based algorithms that are trained with data to improve their accuracy. Actually, this is almost done as supervised learning based on external information, which could be linked with the image features generated by these ML, DL or CNN tools [18, 21].

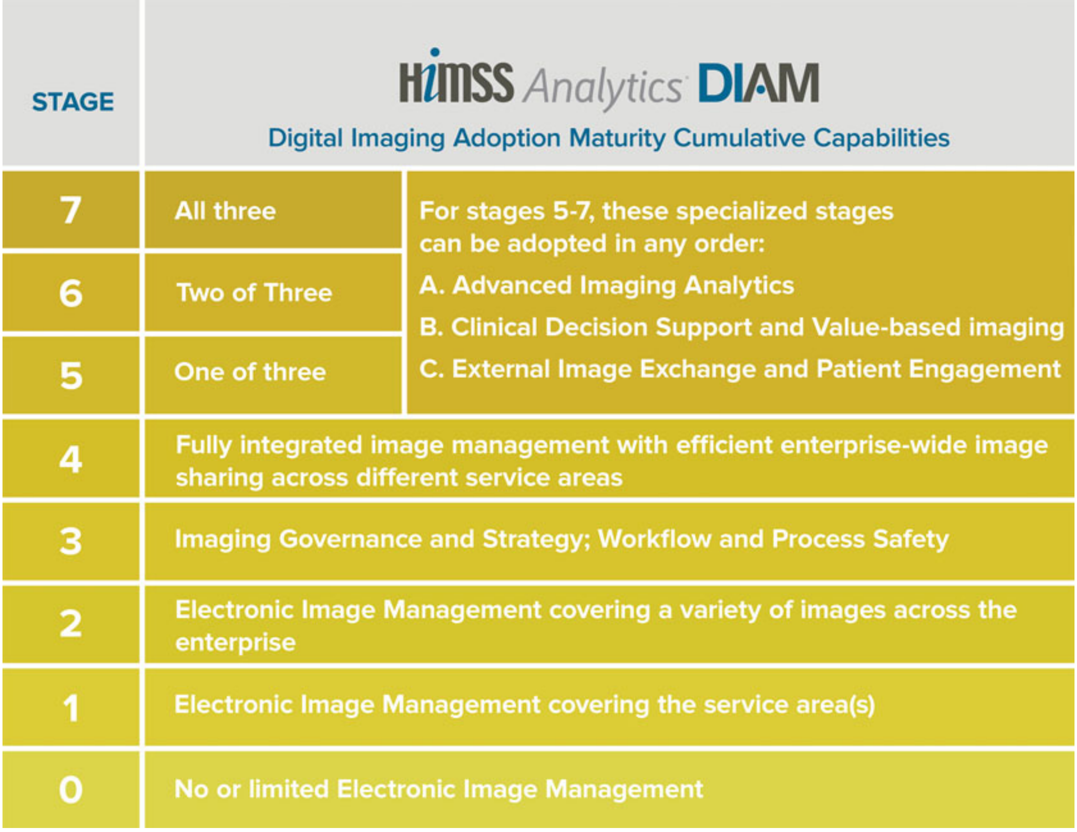
For an efficient automatization of such processes, it is important to have enough labels of different findings, categories, etc. Implementation of a process for manual annotation of the images would be very difficult, because this would require a lot of manual interactions in thousands of imaging studies. Therefore, it is even more important to have reliable metadata and reports. In regard of reports, the discussion and interest on structured reporting have been increased over the last years. IHE is providing a dedicated profile for the “Management of Radiology Report Templates” (MRRT); a joint RSNA and ESR “Template Library Advisory Panel” (TLAP) has been established, and tools to use such templates as integrated solutions in an enterprise IT environment are available today. Within the MRRT profile, it is possible and recommended to have coded information, almost RadLex based [18, 24].

For the development of algorithms, but even more for the introduction of AI in clinical workflows, such a standardization of interfaces, but also of semantic interoperability, is very relevant [17, 25]. An efficient integration of AI should allow automatization of processing, so that at the point and time of reporting, the AI results are already available and thus able to augment the radiologist’s interpretation of imaging studies. Optimisation of the development, testing and implementation of AI in imaging management systems will require appropriate solutions, which provide reliable data, preferably from different institutions and qualified test data. Other technologies such as speech recognition, cloud-based communication and bidirectional interaction could be further aspects in future reporting scenario [26]. The American College of Radiology (ACR) has initiated the Data Science Institute (ACR DSI), and in Germany a cross-institutional cooperation has been created with a platform called HIGH-med [27].

The maturity of enterprise imaging could be monitored based on a new tool developed by HIMSS, ESR, SIIM and EuSOMII called “Digital Imaging Adoption Model” (DIAM). DIAM is already available for radiology, presented in 2016 by ESR and HIMSS [28]. DIAM is based on several stages including many different aspects such as governance and workflow and process security but also specialized applications such as advanced image analytics, clinical decision support, image exchange, etc. (Fig. 9.1). It is obvious that advanced image processing, including AI tools, will be part of state-of-the art enterprise imaging platforms soon.

9.7 Take-Home Points

- Enterprise imaging will replace further decentralized PACS silos, enabling new opportunities for collaboration across different faculties.
- Development of AI tools could benefit a lot from the variety of training data, which could



STAGE	HIMSS Analytics DIAM Digital Imaging Adoption Maturity Cumulative Capabilities	
7	All three	For stages 5-7, these specialized stages can be adopted in any order: A. Advanced Imaging Analytics B. Clinical Decision Support and Value-based imaging C. External Image Exchange and Patient Engagement
6	Two of Three	
5	One of three	
4	Fully integrated image management with efficient enterprise-wide image sharing across different service areas	
3	Imaging Governance and Strategy; Workflow and Process Safety	
2	Electronic Image Management covering a variety of images across the enterprise	
1	Electronic Image Management covering the service area(s)	
0	No or limited Electronic Image Management	

Fig. 9.1 Stage description for enterprise imaging maturity model (HIMSS analytics DIAM for enterprise imaging)

be available with such enterprise imaging platforms.

- Clinical integration will require automatization of workflows with optimized integration of user interfaces.
- The maturity of an enterprise imaging platform can be monitored by DIAM.

References

1. Hartman D, Pantanowitz L, McHugh J, Piccoli A, OLeary M, Lauro G. Enterprise implementation of digital pathology: feasibility, challenges, and opportunities. J Digit Imaging. 2017;30(5):555–60.
2. Roth CJ, Lannum LM, Persons KR. A foundation for enterprise imaging: HIMSS-SIIM collaborative white paper. J Digit Imaging. 2016;29(5):530–8.
3. Roth CJ, Lannum LM, Joseph CL. Enterprise imaging governance: HIMSS-SIIM collaborative white paper. J Digit Imaging. 2016;29(5):539–46.
4. Aryanto KYE, Wetering R, Broekema A, Ooijen PA, Oudkerk M. Impact of cross-enterprise data sharing on portable media with decentralised upload of DICOM data into PACS. Insights Imaging. 2014;5(1):157–64.
5. Towbin AJ, Roth CJ, Bronkalla M, Cram D. Workflow challenges of enterprise imaging: HIMSS-SIIM collaborative white paper. J Digit Imaging. 2016;29(5):574–82.
6. Cram D, Roth CJ, Towbin AJ. Orders- versus encounters-based image capture: implications pre- and post-procedure workflow, technical and build capabilities, resulting, analytics and revenue capture: HIMSS-SIIM collaborative white paper. J Digit Imaging. 2016;29(5):559–66.
7. DICOM. Standard. Available from: https://www.dicomstandard.org/current/.
8. Bhargavan-Chatfield M, Morin RL. The ACR computed tomography dose index registry: the 5 million examination update. J Am Coll Radiol. 2013;10(12):980–3.
9. Roth CJ, Lannum LM, Dennison DK, Towbin AJ. The current state and path forward for enterprise

image viewing: HIMSS-SIIM collaborative white paper. J Digit Imaging. 2016;29(5):567–73.

10. Li S, Liu Y, Yuan Y, Li J, Wei L, Wang Y, et al. Implementation of enterprise imaging strategy at a Chinese Tertiary Hospital. J Digit Imaging. 2018;31:534–42.
11. Clunie DA, Dennison DK, Cram D, Persons KR, Bronkalla MD, Primo HR. Technical challenges of enterprise imaging: HIMSS-SIIM collaborative white paper. J Digit Imaging. 2016;29(5):583–614.
12. Vreeland A, Persons KR, Primo H, Bishop M, Garriott KM, Doyle MK, et al. Considerations for exchanging and sharing medical images for improved collaboration and patient care: HIMSS-SIIM collaborative white paper. J Digit Imaging. 2016;29(5):547–58.
13. Schwind F, Münch H, Schröter A, Brandner R, Kutscha U, Brandner A, et al. Long-term experience with setup and implementation of an IHE-based image management and distribution system in intersectoral clinical routine. Int J Comput Assist Radiol Surg. 2018; https://doi.org/10.1007/s11548-018-1819-2.
14. Liu S, Zhou B, Xie G, Mei J, Liu H, Liu C, et al. Beyond regional health information exchange in China: a practical and industrial-strength approach. AMIA Ann Symp Proc. 2011;2011:824–33.
15. Balthazar P, Harri P, Prater A, Safdar NM. Protecting your patients' interests in the era of big data, artificial intelligence, and predictive analytics. J Am Coll Radiol. 2018;15(3, Part B):580–6.
16. European Society of R. Usability of irreversible image compression in radiological imaging. A position paper by the European Society of Radiology (ESR). Insights Imaging. 2011;2(2):103–15.
17. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. Radiology. 2016;278(2):563–77.
18. Kruskal JB, Berkowitz S, Geis JR, Kim W, Nagy P, Dreyer K. Big data and machine learning—strategies for driving this bus: a summary of the 2016 intersociety summer conference. J Am Coll Radiol. 2017;14(6):811–7.
19. Tang A, Tam R, Cadrin-Chênevert A, Guest W, Chong J, Barfett J, et al. Canadian Association of Radiologists white paper on artificial intelligence in radiology. Can Assoc Radiol J. 2018;69(2):120–35.
20. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, et al. Current applications and future impact of machine learning in radiology. Radiology. 2018;288(2):318–28.
21. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018;9:611–29.
22. Channin DS, Mongkolwat P, Kleper V, Rubin DL. The annotation and image mark-up project. Radiology. 2009;253(3):590–2.
23. Obuchowski NA, Reeves AP, Huang EP, et al. Quantitative imaging biomarkers: a review of statistical methods for computer algorithm comparisons. Stat Methods Med Res. 2014;24:68–106.
24. Pinto Dos Santos D, Klos G, Kloeckner R, Oberle R, Dueber C, Mildenberger P. Development of an IHE MRRT-compliant open-source web-based reporting platform. Eur Radiol. 2017;27(1):424–30.
25. Charles E, Kahn J. From images to actions: opportunities for artificial intelligence in radiology. Radiology. 2017;285(3):719–20.
26. Dreyer KJ, Dreyer JL. Imaging informatics: lead, follow, or become irrelevant. J Am Coll Radiol. 2013;10(6):394–6.
27. Haarbrandt B, Schreiweis B, Rey S, Sax U, Scheithauer S, Rienhoff O, et al. HiGHmed – an open platform approach to enhance care and research across institutional boundaries. Methods Inf Med. 2018;57(S 01):e66–81.
28. Studzinski J. Bestimmung des Reifegrades der IT-gestützten klinischen Bildgebung und Befundung mit dem Digital Imaging Adoption Model (Evaluating the maturity of IT-supported clinical imaging and diagnosis using the Digital Imaging Adoption Model: are your clinical imaging processes ready for the digital era?). Der Radiologe. 2017;57(6):466–469.

Imaging Biomarkers and Imaging Biobanks

10

Angel Alberich-Bayarri, Emanuele Neri, and Luis Martí-Bonmatí

10.1 Introduction

The main barriers limiting the widespread use of quantitative imaging biomarkers in clinical practice lie in the lack of standardization regarding their implementation on those aspects related to technical acquisition, analysis processing, and clinical validation. These developments have multiple consecutive steps, ranging from the proof of concept and mechanism, the hallmark definition, the optimization of image acquisition protocols, the source images, the analytical methodology, the type of measurements to the structured report. The overall pipeline has to provide additional value and support to radiologists in the process of diagnosis and assessment [1]. To enable the use of quantitative imaging biomarkers in both clinical and research settings, a whole process has to be established, including display methods, image analysis guidelines, and acquisition of quantitative data and parametric images. A consensus-based multidisciplinary approach seems the best practice to achieve success.

Based on the recommendations of the Quantitative Imaging Biomarkers Alliance (QIBA), supported by the Radiological Society of North America (RSNA), and the European Imaging Biomarkers Alliance (EIBALL), which is sustained by the European Society of Radiology (ESR), a standard methodology for the development, validation, and integration of image analysis methods for the extraction of biomarkers and radiomic data, as well as for their potential implementation in clinical practice, is being applied to an increasing extent with the aim of reducing variability across centers. All analytical methods developed must comply with critical requirements such as conceptual consistency, technical performance validation (precision and accuracy assessment), clinical endpoint validation, and meaningful appropriateness. Additionally, the continuous technological advances and improvements in medical imaging hardware and software require regular reassessment of the accuracy of quantitative evaluation of medical images, radiomic features, and regular updates of the standardization requirements.

A. Alberich-Bayarri (✉)
Biomedical Imaging Research Group, La Fe Health Research Institute and Polytechnics and University Hospital, Valencia, Spain

Quantitative Imaging Biomarkers in Medicine (QUIBIM S.L.), Valencia, Spain
e-mail: alberich_ang@gva.es

E. Neri
Diagnostic Radiology 3, Department of Translational Research, University of Pisa, Ospedale S. Chiara, Pisa, Italy

L. Martí-Bonmatí
Biomedical Imaging Research Group, La Fe Health Research Institute and Polytechnics and University Hospital, Valencia, Spain

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_10

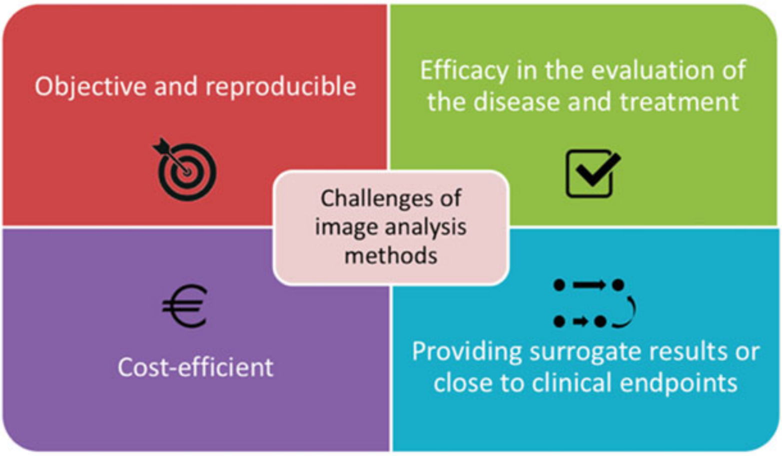

Objective and reproducible
Efficacy in the evaluation of the disease and treatment
Challenges of image analysis methods
Cost-efficient
Providing surrogate results or close to clinical endpoints

Fig. 10.1 Challenges for the adoption of quantitative image analysis methods within clinical routine

Besides this, there is still a risk of persistent heterogeneity of image quality through time, due to differences in technologies implemented by vendors and protocols used across centers. Therefore, standardization of image quality to be used for the analysis of different imaging biomarkers will not be feasible. Although the need for further standardization of all processes in the imaging biomarkers pipeline already started more than 10 years ago, with the intention of increasing their usage in large clinical trials and facilitating their integration in clinical practice, the solution has not arrived yet. The use of artificial intelligence (AI)-based approaches, like the one presented in Chap. 5, could be a disruptive way of changing this trend, by making it possible for complex and deep neural networks to learn from the lack of homogeneity in the collected data.

There are several challenges to be met in the adoption of advanced image analysis methods in clinical routine (Fig. 10.1). Imaging biomarkers not only have to be objective and reproducible, as mentioned earlier, but they also have to show a clear efficacy in the detection and diagnosis of the disease or in the evaluation of treatment response. This diagnostic efficacy must be confirmed by a clear relationship between the biomarkers and the expected clinical endpoints, allowing them to act as surrogate indicators of relevant clinical outcomes such as the prediction of treatment response, progression-free survival, overall survival, and other. Finally, the methodology must be cost-efficient in order to achieve clinical integration and further expansion of its utility.

In this chapter, the general methodology for the development, validation, and implementation of imaging biomarkers is presented. The approach consists of a systematic methodology that allows to obtain features of high precision and accuracy in the imaging biomarker results, making their integration in automated pipelines feasible, for the generation of massive amounts of radiomic data to be stored in imaging biobanks.

10.2 Stepwise Development

In order to brush up an established methodology for the extraction of imaging biomarkers, a summary of the stepwise methodology for radiomic development will be introduced to the reader [2].

The path to the development and implementation of imaging biomarkers involves a number of consecutive phases (Fig. 10.2), starting from the definition of the proof of concept and finalizing with the creation of a structured report including quantitative data. The final step in the development of an imaging biomarker also involves the validation of its relationship with the objective reality to which it's surrogated, either structural, physiological or clinical, and the monitoring of its overall feasibility in multicenter clinical studies. Biomarkers need to follow all phases

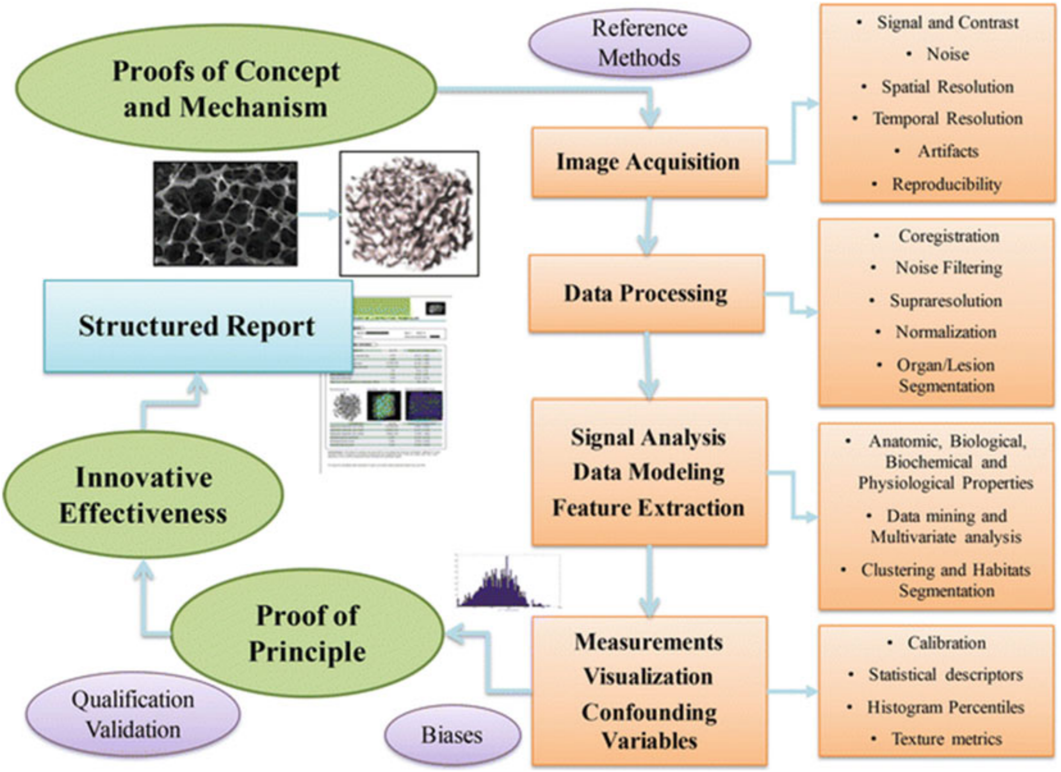


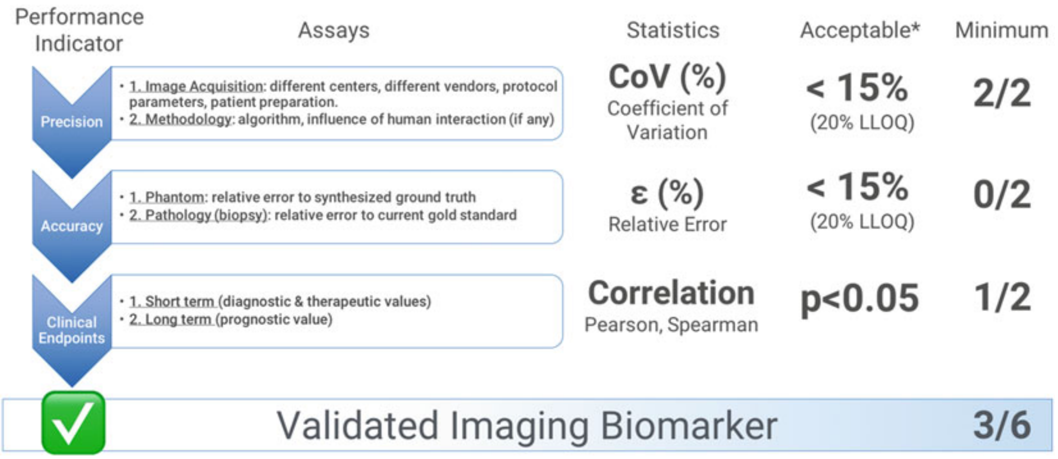
Fig. 10.2 Stepwise development of imaging biomarkers [3]. The stepwise workflow starts from the definition of proof of concept and mechanism, where the clinical need is defined and the relevant information to be measured by the imaging biomarker is determined. The workflow continues with the technical development of an image analysis pipeline, from the image acquisition protocol definition till the generation of quantitative measures. Finally, the extracted measurements are evaluated in a proof of principle within a control population and finally in patients to check for the innovation effectiveness. Finally, a quantitative structured report is generated for the integration of the new imaging biomarker in clinical routine

of development, validation, and implementation before they can be clinically approved [3].

Radiomics solutions should be structured in this stepwise approach, in order to foster standardization of methodologies. Integration of an imaging biomarker into clinical practice needs conceptual consistency, technical reproducibility, adequate accuracy, and meaningful appropriateness. This strategy should permeate all quantitative radiology solutions, from the user interfaces to the source codes of the algorithms. By implementing this methodology, images analysis researchers should be able to reckon the limitations and uncertainties due to limitations in any of the steps involved, such as improper quality of source images (acquisition), uncorrected bias of intensity distribution (processing), oversimplification of mathematical models (analysis), or not statistics not representative for the whole distribution of values (measurements). For example, the reproducibility and feasibility of the implementation of a methodology for radiomic analysis will change dramatically if the segmentation process (tissue or organ) is performed in a manual, semiautomated way or in a completely automated manner supported by artificial intelligence (AI) and convolutional neural networks (CNN).

10.3 Validation

There is no current international consensus on how to validate imaging biomarkers. Our process proposal for validation of imaging biomarkers


Performance Indicator
Assays
Statistics
Acceptable*
Minimum
Precision
• 1. Image Acquisition: different centers, different vendors, protocol parameters, patient preparation.
• 2. Methodology: algorithm, influence of human interaction (if any)
CoV (%)
Coefficient of Variation
< 15%
(20% LLOQ)
2/2
Accuracy
• 1. Phantom: relative error to synthesized ground truth
• 2. Pathology (biopsy): relative error to current gold standard
ε (%)
Relative Error
< 15%
(20% LLOQ)
0/2
Clinical Endpoints
• 1. Short term (diagnostic & therapeutic values)
• 2. Long term (prognostic value)
Correlation
Pearson, Spearman
p<0.05
1/2
✓
Validated Imaging Biomarker
3/6

*European Medicines Agency. *Guideline on bioanalytical method validation*. 21 July 2011

Fig. 10.3 Imaging biomarkers validation pipeline

(Fig. 10.3) considers three steps, taking into account the possible different influences that might introduce uncertainty in the measurements. This pipeline is inspired by the guidelines for the evaluation of bioanalytical methods from the European Medicines Agency (EMA) [4]. The biomarkers are validated in terms of their precision, accuracy, and clinical relationship.

The technical validation of the imaging biomarkers will determine both the precision and accuracy of the measurements, as well as their margin of confidence.

Unlike accuracy, precision can be evaluated for all imaging biomarkers. Obtaining a high precision rate is considered mandatory for the imaging biomarker validation. For precision evaluation, the coefficients of variation (CoV) of the biomarker, obtained repeatedly with the variation of different factors, are calculated. The variable factors can be related either to the image acquisition or to the methodology. In order to evaluate the influence of the image acquisition in the variability of measurements, the imaging biomarkers ideally should be calculated by testing the following variable conditions with the same subjects:

– Imaging center
– Equipment
– Vendors
– Acquisition parameters
– Patient preparation

For the evaluation of the influence of the methodology in the obtained measurements, it is recognized that imaging biomarkers should be calculated with the same subjects and acquisition protocols while changing the following conditions:

– Operator (intra-operator variability, inter-operator variability)
– Processing algorithm

The higher CoV for all the experiments (with varying acquisition characteristics, with varying the operator) should be below 15%. However, in cases with reduced image quality, which can be considered as the *lowest limit of quantification (LLOQ)*, the 15% CoV threshold can be extended to 20% (Fig. 10.3).

The accuracy of the method can be evaluated by comparing the obtained results with a reference pattern in which the real biomarker value is known. The reference pattern can be based on information extracted from a pathological sample after biopsy or from synthetic phantoms (physical or digital reference objects) with different compounds and known properties that emulate the characteristics of the biological tissue. For accuracy evaluation, the relative error of the imaging biomarker compared to the real value from the gold standard must be calculated. The relative error should be below 15% and in *lowest limit of quantification* conditions below 20%.

In some cases, there is no reference pattern available, either because the synthesis of a stable phantom is a complex process or because the considered reference pattern also has a high variability and a coarser category-based analysis than the continuous numerical domain of imaging biomarkers (e.g., steatosis grades in pathology vs. proton density fat fraction quantification from MR). The lack of knowledge in accuracy can be compensated by surpassing the clinical sensitivity and specificity of the calculated imaging biomarker (i.e., we do not know how accurate we are, but we know that the specific imaging biomarker is related to some disease hallmarks).

The main purpose of the clinical validation is to show the relationship between the extracted imaging biomarker and the disease clinical endpoints. The imaging biomarker can be evaluated either as a short-term (assessing detection, diagnosis, and evaluation of treatment response) or long-term (prognostic patient status) measurement. The type and degree of relationship between the imaging biomarkers and clinical variables have to be analyzed based upon sensitivity, specificity, statistical differences between clinical groups, and correlation studies.

10.4 Imaging Biobanks

A biobank is a collection, a repository of all types of human biological samples, such as blood, tissues, cells, or DNA, and/or related data such as associated clinical and research data, as well as biomolecular resources, including model- and microorganisms that might contribute to the understanding of the physiology and diseases of humans. In Europe, the widest network of biobanks is represented by the BBMRI-ERIC (Biobanking and BioMolecular resources Research Infrastructure) (http://bbmri-eric.eu).

In 2014, the European Society of Radiology established an Imaging Biobanks Working Group of the Research Committee, with the intention of defining the concept and scope of imaging biobanks, exploring their existence, and providing guidelines for the implementation of imaging biobanks into the already existing biobanks. The WG defined imaging biobanks as "*organised databases of medical images, and associated imaging biomarkers (radiology and beyond), shared among multiple researchers, linked to other biorepositories*" and suggested that biobanks (which only focus on the collection of genotype-based data) should simultaneously create a system to collect clinically related or phenotype-based data. The basis of this assumption was that modern radiology and nuclear medicine can also provide multiple imaging biomarkers of the same patient, using quantitative data derived from all sources of digital imaging, such as CT, MRI, PET, SPECT, US, X-ray, etc. [5]. These imaging biomarkers can also be classified in different types, depending on their function. Such imaging biomarkers, which express the phenotype, should therefore be part of the multiple biomarkers included in biobanks.

As an example, we have the following biomarkers for the clinical scenario of oncology [5]:

- *Predictive biomarker*: used as a tool to predict the progression and recurrence of disease.
- *Diagnostic biomarker*: used as a diagnostic tool for the identification of patients with disease.
- *Morphologic biomarker*: A biomarker measuring the size or shape of a macroscopic structure in the body
- *Staging biomarker*: used as a tool for classification of the extent of disease.
- *Monitoring biomarker*: used as a tool for monitoring the disease progression and its response to treatment.

Even more, the core content of imaging biobanks should not only exist out of images, but should also include any other data (biomarkers) that may be extracted from images, through computational analysis. All these data should then be linked to other omics, such as genomic profiling, metabolomics, proteomics, lab values, and clinical information [6].

Two types of imaging biobanks can be defined:

1. *Population-based biobanks*: developed to collect data from the general population, in such case the aim of the data collection is to identify risk factors in the development of disease, to develop prediction models for the stratification of individual risk, or to identify markers for early detection of disease.
2. *Disease-oriented biobanks*: developed to collect multi-omics data from oncologic patients or patients affected by neurodegenerative disease, in order to generate digital models of patients. Such models will be used to predict the risk or prognosis of cancer or degenerative diseases and to tailor treatments on the basis of the individual responsivity to therapies. On the basis of the imaging biomarkers that are currently available, cancer of the breast, lung, colorectum, and prostate seem the most suitable entities for developing disease-oriented imaging biobanks, but further applications are expected (neurological tumor such as neuroblastoma, glioblastoma, rare tumors, etc.)

Imaging Biomarkers and Biobanks in Artificial Intelligence

The paradigm shift of working in local environments with limited databases to big infrastructures like imaging biobanks (millions of studies) or federations of imaging biobanks (reaching hundreds of millions of studies) requires the integration of automated image processing techniques for fast analysis of pooled data to extract clinically relevant biomarkers and to combine them with other information such as genetic profiling.

Imaging biomarkers alone will not suffice, and they must be considered in conjunction with other biologic data for a personalized assessment of the disease [7].

As a practical example, it would be of interest to automatically detect whether a certain alteration in the radiomics signature through imaging biomarkers is present in subjects with a given mutation like BRCA. The same can be applied to the relationships between radiomics characteristics and other disease hallmarks in neurodegeneration, diffuse liver diseases, respiratory diseases, osteoarthritis, among many others. Dataset management in imaging biobanks should be able to work longitudinally with different time points along the disease course [8].

However, for these applications, standard statistics analysis methods and tools cannot be applied due to the difficulty in handling large volumes of data. For these applications, the use of advanced visual analytics solutions that help to rapidly extract patterns and relationships between variables is a must (see Fig. 10.4).

Software for imaging biobanks should allow the management of source medical images; the results of associated and labelled clinical, genetic, and laboratory tests (either in the same database or linked); the extracted imaging biomarkers as radiomic features; and data mining environments with visual analytics solutions that simplify the extraction of variable patterns in a huge number of registries.

In this setting it is predictable that the application of machine learning tools will be beneficial. The analysis, stratification among patients, and cross-correlation among patients and diseases, of a huge number of omics data contained in biobanks, are an exercise that cannot be performed by the human brain alone; therefore, a computer-assisted process is needed.

Machine learning could be fundamental for completing human-supervised tasks in a fast way, such as *image acquisition*, *segmentation*, extraction of imaging biomarkers, collection of data in biobanks, data processing, and extraction of meaningful information for the purpose of the biobank.

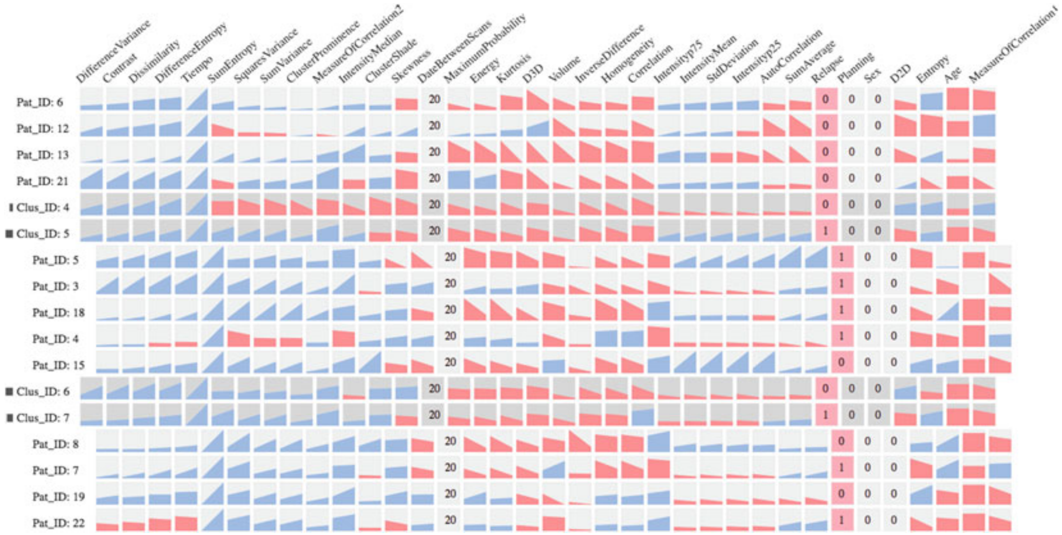


Fig. 10.4 Clustering of patients by the longitudinal evolution in different radiomics features at diagnosis and aftertreatment in rectal cancer. The vertical axis shows patients and the clusters extracted in a non-supervised manner. The horizontal axis includes both radiomic features and the clinical variable of relapse, which is binary (0, no relapse; 1, relapse). The slope in each box represents an increase (blue) or decrease (red) in the imaging biomarker through the different time points

10.5 Conclusion

Imaging biomarkers are an essential component of imaging biobanks. The interpretation of biomarkers stored in the biobanks requires the analysis of big data, which is only possible with the aid of advanced bioinformatic tools. As a matter of fact, medical informatics is already playing a key role in this new era of personalized medicine, by offering IT tools for computer-aided diagnosis and detection, image and signal analysis, extraction of biomarkers, and more recently machine learning, which means that these tools have to adopt a cognitive process mimicking some aspects of human thinking and learning through the progressive acquisition of knowledge.

10.6 Take-Home Points

- Before clinical approval, AI algorithms and biomarkers have to follow all phases of development, validation, and implementation.
- The technical validation of AI algorithms that generate imaging biomarkers will determine both the precision and accuracy of the measurements, as well as their margin of confidence.
- The imaging biomarkers generated by AI can be evaluated either as a short-term (assessing detection, diagnosis, and evaluation of treatment response) or long-term (prognostic patient status) measurement tool.
- Imaging biobanks do not only consist of images but any other data (biomarkers) that can be extracted from them through computational analysis; all these data should then be linked to other omics, such as genomic profiling, metabolomics, proteomics, lab values, and clinical information.

References

1. Martí Bonmatí L, Alberich-Bayarri A, García-Martí G, Sanz Requena R, Pérez Castillo C, Carot Sierra JM, Manjón Herrera JV. Imaging biomarkers, quantitative imaging, and bioengineering. Radiologia. 2012;54:269–78.

2. European Society of Radiology (ESR). ESR statement on the stepwise development of imaging biomarkers. Insights Imaging. 2013;4:147–52.
3. Martí-Bonmatí L. Introduction to the stepwise development of imaging biomarkers. In: Martí-Bonmatí L, Alberich-Bayarri A, editors. Imaging biomarkers. Development and clinical integration, vol. 2; 2017. p. 27. isbn:9783319435046.
4. European Medicines Agency. Guidelines on bioanalytical methods validation. 21 July 2011. EMEA/CHMP/EWP/192217/2009 Rev. 1 Corr. 2.
5. O'Connor JP, Aboagye EO, Adams JE, et al. Consensus statement. imaging biomarkers roadmap for cancer studies. Nat Rev Clin Oncol. 2017;14(3):169–86. https://doi.org/10.1038/nrclinonc.2016.162.
6. European Society of Radiology (ESR). ESR position paper on imaging biobanks. Insights Imaging. 2015;6:403–10.
7. Alberich-Bayarri A, Hernández-Navarro R, Ruiz-Martínez E, García-Castro F, García-Juan D, Martí-Bonmatí L. Development of imaging biomarkers and generation of big data. Radiol Med. 2017;122:444–8.
8. Neri E, Regge D. Imaging biobanks in oncology: European perspective. Future Oncol. 2017;13:433–41.

Part V

Practical Use Cases of A.I. in Radiology

11 Applications of AI Beyond Image Interpretation

José M. Morey, Nora M. Haney, and Woojin Kim

The use of deep learning has led to rapid advancement in machine's ability to perform image analysis, such as detection of tuberculosis on chest X-ray, lung nodule and interstitial lung disease on chest CT, pulmonary embolism on CT angiography, breast mass on mammography, intracranial hemorrhage on head CT, brain tumor on MRI, and many others [1–3]. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is an annual competition that began in 2010 where participants evaluate their algorithms for image classification and object detection [4]. When a deep convolutional neural network was used in 2012, there was a dramatic decrease in the image classification error rate. By 2015, the algorithms began to perform at levels exceeding human ability at certain image classification tasks. Along with rapid advancement in deep learning, there has been increasing media attention, research publications, and startups focused on using AI to identify findings within medical images [3]. This affects fields in medicine that rely on images, such as radiology, pathology, ophthalmology, dermatology, and neurology [5–10]. As a result, it is easy to have a tunnel vision when it comes to the potential of AI in radiology. The purpose of this chapter is to widen the field of view to demonstrate many other areas within radiology where AI can benefit beyond image interpretation [11].

Borrowing a concept from Michael Porter's book, *Competitive Advantage: Creating and Sustaining Superior Performance*, Boland et al. described something called the "imaging value chain" to highlight discrete steps within radiology workflow where they can provide value to the patient [12]. The original imaging value chain had outlined the following components: patient, referring physician, appropriateness determination and patient scheduling, imaging protocol optimization, imagine modality operations, interpretation and reporting, and communication, interconnected by data mining and business intelligence. We have modified the steps to illustrate an AI imaging value chain (Fig. 11.1). The subsequent sections will describe each of the components in greater detail.

J. M. Morey
Singularity University, Moffett Field, CA, USA

Liberty Biosecurity, Arlington, VA, USA

Hyperloop Transportation Technologies, Culver City, CA, USA

NASA iTech, Hampton, VA, USA

Eastern Virginia Medical School, Norfolk, VA, USA

University of Virginia, Charlottesville, VA, USA

N. M. Haney
Johns Hopkins Hospital, Baltimore, MD, USA

W. Kim (✉)
Nuance Communications, Burlington, MA, USA

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_11

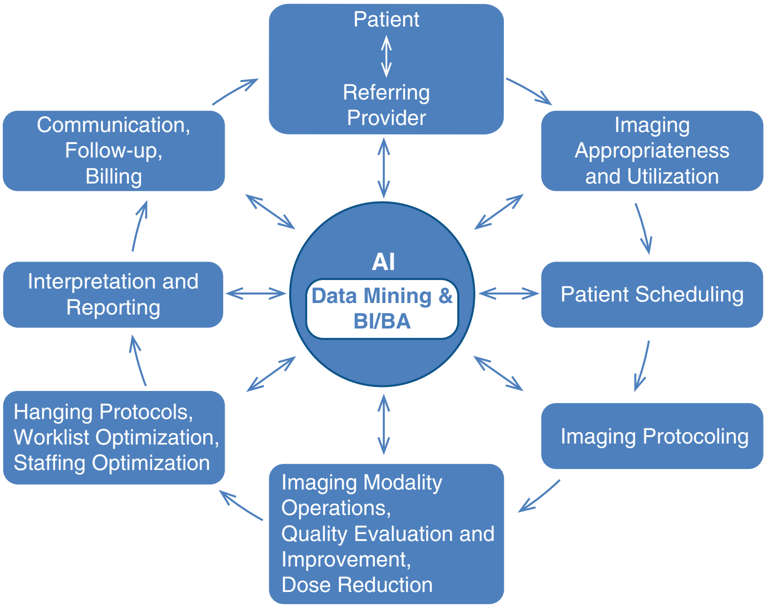

Patient
Referring Provider
Imaging Appropriateness and Utilization
Patient Scheduling
Imaging Protocoling
Imaging Modality Operations, Quality Evaluation and Improvement, Dose Reduction
Hanging Protocols, Worklist Optimization, Staffing Optimization
Interpretation and Reporting
Communication, Follow-up, Billing
AI
Data Mining & BI/BA

Fig. 11.1 Schematic representation of the AI imaging value chain based on the imaging value chain described by Boland et al. [12] (BI: Business Intelligence, BA: Business Analytics)

11.1 Imaging Appropriateness and Utilization

Imaging utilization has been increasing over the years. Clinical decision support (CDS) is designed to help referring providers make better-informed decisions when ordering imaging exams. Improvements in CDS, which reduce inappropriate utilization of imaging resources, can enhance the quality of patient care and reduce healthcare cost [13–15]. Machine learning has the potential to aid in the selection of the most appropriate imaging tests and predict trends in imaging utilization [16, 17]. For example, Hassanpour et al. used support vector machine classifier to analyze radiology reports to predict patients who are likely to be high utilizers of imaging services with an accuracy of 94%, showing the possibility of using machine learning to curb imaging use [18].

11.2 Patient Scheduling

Optimization of scheduling can enhance the patient experience and thus positively impact patient satisfaction. AI can improve scheduling of imaging exams. For example, many facilities use fixed block lengths when scheduling patients for MRI exams. However, MRI exams can be variable in length, which can result in a significant amount of time where the MRI scanner is idle. In the United States and Canada, it is estimated that the scanners are operational only 70% of the time. Muelly et al. used AI to optimize the amount of time allotted to patients undergoing MRI [19, 20]. The inputs were based on the exam protocol, patient demographics, contrast usage, and the historical average of unplanned sequence repeats per exam protocol. They were able to demonstrate such scheduling optimization can increase the number of exams scanned per day while reducing mean patient wait times. Additionally, if patient preferences can be met regarding the choice of physicians and preferred time slot, patient satisfaction can be further improved [21].

One of the common challenges in healthcare is patient no-show visits (NSVs). For the patients, NSVs can potentially delay care, which may result in adverse outcomes. For the healthcare practices, NSVs can lead to financial losses. Radiology is no exception. Previous studies using data analytics and machine learning have shown certain sociodemographic factors and past behaviors can be predictors of patient NSVs and can assist in the creation of an effective overbooking approach to maximize capacity and decrease the cost of NSVs [22–37]. Further, machine learning techniques can be used to create solutions to improve the NSV rates [28]. For example, Harvey et al. showed how NSVs in radiology could be predicted using regression modeling [27].

11.3 Imaging Protocoling

Once the appropriate imaging exam has been ordered and scheduled, radiologists and radiology trainees often spend time protocoling exams in advance or modifying imaging protocols at the time of the exam to ensure the exam answers the clinical question. This task typically involves reviewing indication or reason for the exam as well as the medical history of the patient within the electronic medical record (EMR), which includes recent office visit notes, prior hospital admissions and discharge summaries, past surgical history including surgical operative notes, medication list, allergies, and relevant lab values and pathology results. Also, they review prior imaging exams and radiology reports, including how they were protocoled in the past. Depending on the complexity of the case, this can be a time-consuming process. AI has the potential to protocol these cases in advance with the radiologist supervision, which can save time while minimizing variations and errors [11]. Indeed, AI has shown promising results in determining proper protocols for MRI exams [38–42].

Some authors, however, have highlighted the downside of “black-box” AI algorithms when it comes to implementing these in real life. Trivedi et al. used IBM Watson to automate determination of the need for intravenous gadolinium-based contrast in musculoskeletal MRI exams [38]. The authors had difficulty troubleshooting some of the errors made, including a “critical error” IBM Watson made when it decided to give contrast to a patient with an end-stage renal disease. Despite limitations, the use of AI in protocoling has the potential to improve efficiency, decrease error rates, and be incorporated into the clinical workflow of the ordering providers to provide CDS when ordering imaging exam.

11.4 Image Quality Improvement and Acquisition Time Reduction in MRI

Emergency scenarios, such as stroke, require rapid acquisition protocols to detect life-threatening pathology and deliver the appropriate treatment. MRIs are the more time-intensive modality compared to other imaging modalities. The patients also need to be still for the procedure. Thus, there have been a number of researches in applying AI to improve image quality while reducing scanning times. Under-sampled MRI data, including compressed sensing MRI (CS-MRI) that reconstructs MR images from very few Fourier k-space measurements to allow for rapid acquisition times, can be reconstructed using deep learning to create corrected images that have image quality similar to standard MRI reconstruction with fully sampled data while reducing acquisition times. The reduction in acquisition times can improve the patient experience through reduction or elimination of multiple breath-holds during MRI exams, enhance patient throughput, reduce motion artifacts, and allow for imaging of moving organs, such as the heart [43–50].

Gong et al. used a deep learning method to reduce gadolinium dose in contrast-enhanced brain MRI by training deep learning model to approximate full-dose 3D T1-weighted inversion-recovery prepped fast-spoiled-gradient-echo (IR-FSPGR) images from pre-contrast and 10% low-dose images [51]. Their deep learning model showed gadolinium dose can be potentially reduced tenfold without significant image quality degradation.

11.5 Image Quality Improvement and Radiation Dose Reduction

Improving image quality typically comes at the expense of increased radiation dosing. However, the widespread and increasing use of CT has raised concerns about potential radiation risks, and this has motivated the development of various dose optimization techniques. CT manufacturers, for example, have developed iterative reconstruction techniques to reduce noise, which in turn provide artifact reduction and radiation dose savings [52, 53].

The new research is showing how AI can be used to lessen the CT dose even further. Similar to creating "super-resolution" images, also known as single image super-resolution, where a photorealistic high-resolution image can be generated from a low-resolution image by using deep learning [54–60], researchers are applying deep learning to improve the quality of the ultra-low-dose CT images [61–65]. A multicenter survey study used an artificial neural network (ANN) to reconstruct CT images from low-dose acquisitions. When the survey respondents were asked to compare the image quality of the deep learning-based post-processed low-dose CT images to those using standard reconstructions, they found the deep learning-based post-processed images to be comparable to what they would expect to see from higher-dose CT exams. In many cases, their ANN turned a nondiagnostic exam to a diagnostic study, where the survey respondents found 91% of the images to be diagnostic compared to only 28% of the images when they were created using the standard reconstruction method. These techniques show promising results to lower the radiation dose further with improved reduction of motion artifacts without potentially compromising diagnostic image quality [66].

In positron-emission tomography (PET) imaging, there is concern over the risk of radiation exposure due to the use of radioactive tracer. Similar to low-dose CT exams, lowering the dose can result in low signal-to-noise ratio (SNR) with degradation of image quality, affecting one's ability to make the diagnosis. Xu et al. proposed a deep learning method using an encoder-decoder residual deep network to reconstruct low-dose PET images to a standard-dose quality using only 0.5% of the standard radiotracer dose. Such improvements have the potential to reduce exam times, decrease radiation exposure, lower costs, and alleviate shortages in radiotracers. With such reduction in dose, it also raises the possibility of using PET as a screening tool [67].

11.6 Image Transformation

AI can be used to simulate images that are of the different sequence on MRI or with features of different modalities. For example, one can use deep learning methods to perform MRI image-to-image translation, such as from T1 to T2, from T1 to T2-FLAIR, from T2 to T2-FLAIR, and vice versa [68]. Liu et al. developed a deep learning approach for MR imaging-based attenuation correction (MRAC) by generating discrete-valued pseudo-CT scans from a single high-spatial-resolution diagnostic-quality 3D MR image in brain PET/MR imaging, which resulted in reduced PET reconstruction error relative to current MRAC approaches [69]. Creating such pseudo-CT (also known as synthetic CT) images using MR images can also be used for radiation therapy [70, 71]. Finally, some researchers have used deep learning methods to predict PET images from CT and MRI. Ben-Cohen et al. combined a fully convolutional network (FCN) with a conditional generative adversarial network (GAN) to generate simulated PET images from CT [72], and Li et al. generated synthetic PET images using MR images in patients with Alzheimer's disease [73].

11.7 Image Quality Evaluation

Evaluations for image quality are typically done through visual inspection. Occasionally, suboptimal images are not recognized until long after the exam has been completed, decreasing the radiologist's ability to make a confident diagnosis resulting in requiring additional imaging to be done through repeat scan, which can lead to increased cost and decreased patient satisfaction. Deep learning technology can potentially automate the detection of suboptimal quality during the exam acquisition, with examples of research in MRI [19]. For example, Esses et al. used a CNN algorithm to screen for nondiagnostic T2-weighted images of the liver on MRI that demonstrated high negative predictive value [74].

11.8 Hanging Protocols

Using the Digital Imaging and Communications in Medicine (DICOM) metadata, typical picture archiving and communication system (PACS) viewers can display radiology exams in specific, predetermined layout according to hanging protocols, also known as default display protocols (DDP) [75]. The hanging protocols allow a specific set of images to be presented consistently. By minimizing the manual adjustments a radiologist has to make each time, hanging protocols can increase the efficiency of the radiologist [20]. Their importance grows with increasing complexity of the exam being reviewed. For example, an MRI exam can have multiple series of pulse sequences and imaging parameters in various anatomic planes, which gets more complicated when there is a need to make a comparison with prior exam(s). A good hanging protocol has to take into account image order, orientation, modality, anatomy/body part, image plane, window level, pulse sequence, etc. Even elements like the percentage each image takes up in screen real estate can be incorporated into hanging protocols.

While there has been work like the RSNA® RadLex® Playbook™, there is no standardization of naming convention of these metadata [75–78]. Also, manual entry of these study and series descriptions can lead to errors and inconsistencies. As a result, it is difficult for many PACS viewers to consistently display hanging protocols, leaving the radiologists making frequent adjustments. To illustrate its potential importance, imagine if you had to readjust your seat, rearview mirror, and wing mirrors every 15 min while driving. Machine learning and deep learning technology can be used to identify these elements within an image and adapt appropriately. Furthermore, it can learn from user behavior to adjust the hanging protocols to individual needs.

11.9 Reporting

AI in reporting and documentation can improve differentials and diagnoses. Duda et al. developed a Bayesian network interface for assisting radiology interpretation and education [79]. Others have developed expert-based Bayesian network to provide neuroradiology support in the diagnosis of brain tumors [80] and spine pathology [66].

There are many evidence-based guidelines within radiology, such as those provided by the ACRassist™ [81]. However, it may be difficult for the radiologists to remember all the guidelines in detail. AI has the potential to provide decision support for the radiologists by detecting and presenting whenever a particular guideline is applicable based on the radiologist's interpretation at the time of reporting, including appropriate follow-up recommendations.

Advancements are being made in the field of radiology to establish standards that can be applied to and automated within reporting, making it easier to share and reuse data [11, 82]. The reusability of imaging datasets and associated reports is becoming increasingly important as quality datasets are necessary to create and validate AI products [75]. Automated methods using natural language processing (NLP) can be used to identify findings in radiology reports, which can be used to generate a large labeled corpus that can be used for deep learning applications [83].

Once the radiology report has been created, NLP, a form of AI, helps computers comprehend and interpret human language [84, 85]. Pons et al. studied peer-reviewed papers on the subject and have grouped them in five broad categories that represent different relevant purposes: diagnostic surveillance, cohort building for epidemiologic studies, query-based case retrieval, quality assessment of radiologic practice, and clinical support services [86]. Diagnostic surveillance included alert systems for critical results described in radiology reports, such as appendicitis, acute lung injury, pneumonia, thromboembolic diseases, and malignant lesions. NLP can be used to not only identify disease but also its progression by analyzing a series of reports. Cohort building for epidemiologic studies included using NLP to improve the efficiency of epidemiologic research by identifying potential cases, such as renal cysts, pneumonia, pulmonary nodules, pulmonary embolism (PE), metastatic disease, adrenal nodules, abdominal aortic aneurysm, peripheral arterial disease, and various liver pathologies. Query-based case retrieval has benefited from leveraging NLP when searching for radiology reports for research, clinical decision support, quality improvement, education, and billing and coding. Quality assessment of radiologic practice can include the use of NLP systems to generate descriptive statistics on various quality metrics, such as recommendation behavior, report completeness, and communication of critical results. Finally, clinical support services include using NLP in assisting radiologists at the time of interpretation, including providing clinical decision support for the radiologists on various pathologies, detecting errors within radiology reports related to laterality and sex, and assisting with billing and coding.

More recently Chen et al. have compared the performance of a deep learning convolutional neural network (CNN) model with a traditional NLP model in extracting PE findings from thoracic CT reports [61, 62]. In their study, the CNN model performed with accuracy equivalent to or beyond that of an existing traditional NLP model. Tan et al. developed an NLP system from both rule-based and machine-learned models for identification of lumbar spine imaging findings related to low back pain on MRI and radiographs [87]. Their machine-learned models demonstrated higher sensitivity with slight loss of specificity and overall higher area under the curve (AUC). Hassanpour et al. developed an NLP method using a combination of machine learning and rule-based approaches to automatically extract clinical findings in radiology reports and characterize their level of change and significance [18]. Lastly, another group explored the effect of integrating NLP and machine learning algorithms to categorize oncologic response in radiology reports [88]. They evaluated cross-sectional abdomen/pelvis CT and MR exams with malignancy using various combinations of three NLP techniques (term frequency-inverse document frequency, term frequency weighting, and 16-bit feature hashing) and five machine learning algorithms (logistic regression, random decision forest, one-versus-all support vector machine, one-versus-all Bayes point machine, and fully connected neural network). They found that the best accuracy was achieved when both the NLP- and machine learning-based algorithms were optimized concurrently.

11.10 Text Summarization and Report Translation

With increasing adoption of EMRs, the radiologists have greater access to the patient's medical record. However, it can be time-consuming for the radiologist to review the medical history as well as prior imaging reports for relevant information. There has been work to summarize longer texts in compressed form using machine learning [89], and such text summarization technology can potentially be used to provide a concise summary for the radiologist to assist in imaging interpretation [90].

AI has been used to improve machine translation of one human language to another [91]. Google introduced Google Neural Machine Translation (GNMT) in November 2016 that uses an ANN to increase fluency and accuracy in Google Translate [92]. These show the possibility of using deep learning to automatically translate radiology reports from one language to another for patients and providers.

Finally, NLP and deep learning can be used to generate radiology reports that are tailored to different members of the overall patient's care. For example, a single radiology report can be generated by the radiologist, which can be modified by the AI for the patient, primary care physician, specialist, and surgeon. One way to help patients become more engaged in their own healthcare is through healthcare literacy. Multiple institutions have created programs to improve patient's knowledge of their disease by creating documents that can be understood by the average reader [93–95]. These reports, if created with machine learning, could adapt to the patients specifically at their own reading level. Inputs could include patient age, native language, and education level, but greater adaptability may be achieved if the patients interact with the machine and actively look up definitions while reading their report.

11.11 Speech Recognition

Speech recognition (SR) has been widely used by radiologists for many years. However, because SR can result in transcription errors, the radiologists must carefully proofread and edit reports. While there has been a significant improvement in the SR technology over the years, SR can cause errors like wrong-word substitution, nonsense phrases, and missing words. Advancement in AI in SR has the potential to reduce such errors in radiology reports [96, 97]. In recent years, researchers have been incorporating deep learning into SR [98–101]. In addition, with advancement in SR, it can be used to detect and alert the radiologists of reporting errors related to wrong laterality and sex, provide clinical decision support in real time, and remind radiologists of reporting guidelines related to compliance and billing. Finally, SR can be leveraged to provide a virtual assistant capability to the radiologists throughout the day. For example, one can use voice to retrieve information from the EMR or reference materials, to control viewers or reporting solutions, to set reminders, and to communicate critical findings.

11.12 Follow-up

Follow-up care for patients is becoming increasingly important as healthcare transitions from volume- to value-based care, where there is a greater emphasis on outcome-based reimbursement. Closing the loop on follow-up imaging care is essential as researchers have shown there are relatively high follow-up failure rates in radiology. For instance, Blagev et al. found 71% of incidentally detected pulmonary nodules are not appropriately followed up [102]. Cook et al. found out of the patients who had either indeterminate or suspicious findings that warranted a follow-up, 44% had not completed any abdominal imaging follow-up [103]. Failed follow-ups can lead to poor patient outcomes and medical malpractice. In addition, there is possible financial impact by losing potential technical and professional revenue from the follow-up exam. AI can be used to develop a system to extract follow-up information from radiology reports, which in turn can be used to track and monitor patient follow-ups [104]. Furthermore, AI has great potential to assist in identifying which patients should be followed more closely. For example, predictive intelligence has been used to improve patient follow-up for pulmonary nodules [105].

11.13 Worklist Optimization

Most modern worklists used by the radiologist in their daily workflow provide sorting of cases in the order of urgency, such as "stat" versus "routine." By using deep learning to analyze images for the presence of critical findings, one can more intelligently prioritize the worklist of the radiologists by the presence or lack of critical findings within the images, regardless whether or not they were marked "stat." For organizations with a large volume of exams, such triage through optimized workflow can reduce report turnaround times (RTAT) for critical exams with a potential to positively impact patient outcomes [2, 50]. For example, both researchers and startups have demonstrated a positive impact on RTAT for head

CT exams with acute intracranial hemorrhage, some by as much as 96% in RTAT reduction, through re-prioritization of worklist using deep learning [106, 107].

11.14 Staffing Optimization

Especially for large radiology practices and those with multiple sites that need coverage, appropriate scheduling of radiologists can be a challenging problem that can lead to lost revenue and decreased radiologist satisfaction if not done well [108]. One must consider multiple factors, including time of day, the day of the week, coverage location (emergency department, inpatient, outpatient), referral patterns, and exam modality, type, complexity, and volume [11]. There is additional challenge of maximizing the group's flexibility in choosing vacation and meeting times. For some, academic times also have to be factored in [109]. All of this needs to be done with a sense of fairness without compromising patient care. Boroumand et al. demonstrated better matching of radiology staffing with inpatient imaging workflow patterns led to improved turnaround times and critical results reporting [110]. AI has the potential to optimize staffing for radiology as seen in other industries [111].

11.15 Business Intelligence and Business Analytics

Radiology has high fixed costs due to the price of imaging equipment, personnel, and advanced knowledge required to deliver high-quality imaging, diagnosis, and treatment. Business intelligence (BI) and business analytics (BA) identify and quantify metrics such as diagnostic accuracy, turnaround time, imaging modality utilization, and wait times, among others [13, 112, 113]. With the ongoing shift of the medical payment models from volume to value, there will be growing pressure for the radiology groups to demonstrate value through improved patient outcomes with challenges in defining value and radiology's contribution. BI tools have been used in the financial industry for years and can be transitioned successfully to the healthcare environment for the process, quality, safety, and financial performance improvement [112]. The Society for Imaging Informatics in Medicine (SIIM) has described its SIIM Workflow Initiative in Medicine (SWIM) to optimize workflow, improve efficiency, decrease wasted time, and allow for cost savings through incorporating data analytics [16, 114]. Scorecards and outcome measures will likely incorporate machine learning-generated metrics in the near future [16]. BI and BA tools augmented by AI have the potential to add value to the entire imaging value chain, from improved workflow, efficiency, service, cost-effectiveness, revenue, quality and safety, and patient outcomes and satisfaction.

11.16 Content-Based Image Retrieval

In the Reporting section, we discussed the use of NLP and machine learning in query-based case retrieval to search the radiology report texts. Similar to reverse image search features available online, AI can be used to conduct searches for medical images using images instead of text. Content-based image retrieval (CBIR), also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR), involves using computer vision techniques in image retrieval, specifically searching for digital images in large databases. This is in contrast to the concept-based image retrieval, where searches are text-based and are on the metadata associated with the images rather than "contents" of the images. Here, the "content" refers to the features of the image pixel data, such as colors, shapes, textures, etc. [115]. In radiology, traditionally, the images were searched using the concept-based image retrieval method with the searches made using either the radiology report text associated with the images as well as report metadata, such as accession number, patient demographics, patient status, modality, exam description, CPT, dates and time stamps, exam location, ordering providers, and radiologists, or using the PACS image metadata, such as DICOM header information. However, with CBIR, instead of typing “clear cell renal cell carcinoma” in a search box and filtering it by MRI, one can draw a box around the lesion and ask the computer to find all exams that contain similar appearing lesion. While challenges remain in applying deep learning techniques in CBIR in radiology [3], advanced CBIR in radiology can assist in clinical decision support, research, and education within radiology and ultimately can be used to improve patient care [12].

11.17 Patient Safety

Patient care and safety is the ultimate end goal of quality healthcare delivery. Each of the preceding sections either directly or indirectly affects patient safety via a reduction in errors, improved clinical decision making and diagnosis, reduction in overutilization in a resource-limited environment, and reduction in cost which can allow spending to be used more strategically to enhance the delivery of healthcare. As such, the potential impact machine learning can have on patient safety is boundless. However, as AI is neither “astute nor intuitive,” physicians will be essential to design, enforce, validate, and revisit AI products to ensure patient safety while improving healthcare services [116].

11.18 Billing

Errors and omissions within radiology reports can lead to decreased reimbursement. In the United States, Duszak et al. evaluated abdominal ultrasound reports for frequency, characteristics, and financial impact of radiologist documentation deficiencies. They found incomplete documentation was common (9.3–20.2%), resulting in 2.5–5.5% in lost professional income [117]. As discussed previously, NLP and machine learning can be used to assist radiologists in ensuring complete documentation for maximum reimbursement. Healthcare organizations are beginning to look into machine learning for dealing with much larger complex issues of reimbursement denials, which can cost them 3–5% of their annual net revenue, if not higher [118]. While analyzing denials globally for hospitals and health systems can be complicated, radiology offers a much narrower focus to allow for potentially greater success rates by using machine learning.

In the United States, both positive and negative Medicare payment adjustments, under the Merit-based Incentive Payment System (MIPS), are dependent upon the quality and other performance measurements. Some quality measures apply to radiology where much of the data required for reporting rely on the content of the radiology reports. Hence, NLP and machine learning-based advanced text analytics can have a beneficial impact on allowing imaging practices to compete on these quality measures for improved reimbursement.

11.19 Patient Experience

In addition to translating radiology reports into a patient’s native language and reading level, the patient experience can be greatly enhanced with AI. Of late, online resources are the first and sometimes the only resources patients will use to guide their healthcare decisions. These online resources include organized websites such as WebMD and personalized patient portals but also include social media such as Twitter and Facebook [119, 120]. In conjunction with an increasing use of telemedicine, AI can help make information on social media searchable and targeted to the proper communities [119].

Further, communities for specific disease groups already exist, specifically for diabetes mellitus and coronary artery disease [121]. The combination of machine learning and digital imaging posted to these groups may allow for red flags to be picked up early, allowing for quicker diagnosis and treatment. AI could also form the foundation for patient- and family-centered care (PFCC) by automatically providing hyperlinks to online and local resources for findings noted in the radiology report. Overall, radiologists are in a great position to engage with patient communities regarding diagnosis, treatment, and prevention practices improving patient healthcare education and overall patient experience.

11.20 Challenges

While there has been significant advancement of AI in medical imaging in recent years, many challenges remain. Some of the challenges include brittleness and validation of the AI models. While these AI models work well within a research environment where they were created, they may operate poorly in real life when there are differences in variables like scanners, imaging protocols, patient population, and region. Another challenge is a relative lack of transparency and explainability as to how some of the AI models work or, more importantly, do not work, the so-called "black box" problem described earlier. Although the researchers are working to solve using various visualization techniques and other novel techniques like looking at pertinent negatives, more is needed for these algorithms to be accepted in medicine as the physicians will need to be able to trust the outputs of the AI models and to be able to explain them to their patients [122, 123]. This raises another particularly difficult challenge when working in healthcare, which is the regulatory hurdles. Some algorithms that do not involve image interpretation, as described in this chapter, face less regulatory burden. As a result, some of these AI algorithms may come to the market quicker. Other challenges include the difficulty of creating large, annotated training data sets, which are important for training deep learning models. In addition, there are no standards for clinical integration of these AI models. Finally, we must be aware of the potential for bias within AI algorithms and when present to be able to recognize and prevent their use. Despite many limitations and challenges, it is important to emphasize that AI holds tremendous potential to improve patient care and radiologist's work.

11.21 Conclusion

Although there has been a much greater focus in the AI medical imaging research and industry on using AI to make findings within images, this chapter explored the usefulness of AI in radiology beyond image interpretation. As shown, AI can potentially improve and have a positive impact on every aspect of the imaging value chain in radiology. In the era of value-based imaging, AI can augment and empower the radiologists to provide better patient care and increase their role in the overall patient care and the care team. The AI technology alone cannot accomplish these goals. As we learn to overcome the challenges and integrate AI into radiology, we must have a fuller field of view beyond interpretation to take full advantage of this emerging technology [124]. Those radiologists and practices that can embrace and take maximum advantage of AI where available will be able to position themselves well for the future.

11.22 Take-Home Points

- AI has many applications in radiology beyond image interpretation. AI will be an augmentation for all aspects of the radiology value chain.
- Uses of AI beyond image interpretation include improving imaging appropriateness and utilization, patient scheduling, exam protocoling, image quality, scanner efficiency, radiation exposure, radiologist workflow and reporting, patient follow-up and safety, billing, research and education, and more.
- We, in radiology, must use AI beyond image interpretation to take full advantage of this emerging technology, which can improve on every aspect of the imaging value chain and augment and empower the radiologists to provide better patient care.
- Radiologists should familiarize themselves with the benefits and limitations of AI.

References

1. Erickson BJ, Korfiatis P, Akkus Z, et al. Machine learning for medical imaging. Radiographics. 2017;37:505.
2. Prevedello LM, Erdal BS, Ryu JL, et al. Automated critical test findings identification and online notification system using artificial intelligence in imaging. Radiology. 2017;285:923.
3. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60.
4. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis. 2015;115:211.
5. Sahran S, Albashish D, Abdullah A, et al. Absolute cosine-based SVM-RFE feature selection method for prostate histopathological grading. Artif Intell Med. 2018;87:78–90.
6. Pedrosa M, Silva JM, Matos S, et al. SCREEN-DR – software architecture for the diabetic retinopathy screening. Stud Health Technol Inform. 2018;247:396.
7. Guo LH, Wang D, Qian YY, et al. A two-stage multi-view learning framework based computer-aided diagnosis of liver tumors with contrast enhanced ultrasound images. Clin Hemorheol Microcirc. 2018;69:343–54.
8. Saltz J, Gupta R, Hou L, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Rep. 2018;23:181.
9. Hramov AE, Frolov NS, Maksimenko VA, et al. Artificial neural network detects human uncertainty. Chaos. 2018;28:033607.
10. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115.
11. Lakhani P, Prater AB, Hutson RK, et al. Machine learning in radiology: applications beyond image interpretation. J Am Coll Radiol. 2018;15:350.
12. Boland GW, Duszak R Jr, McGinty G, et al. Delivery of appropriateness, quality, safety, efficiency and patient satisfaction. J Am Coll Radiol. 2014;11:7.
13. Brink JA, Arenson RL, Grist TM, et al. Bits and bytes: the future of radiology lies in informatics and information technology. Eur Radiol. 2017;27:3647.
14. Ip IK, Schneider L, Seltzer S, et al. Impact of provider-led, technology-enabled radiology management program on imaging. Am J Med. 2013;126:687.
15. Sistrom CL, Dang PA, Weilburg JB, et al. Effect of computerized order entry with integrated decision support on the growth of outpatient procedure volumes: seven-year time series analysis. Radiology. 2009;251:147.
16. Kruskal JB, Berkowitz S, Geis JR, et al. Big data and machine learning-strategies for driving this bus: a summary of the 2016 intersociety summer conference. J Am Coll Radiol. 2017;14:811.
17. Morey JM, Haney NM, Cooper PB. A predictive diagnostic imaging calculator as a clinical decision support tool. J Am Coll Radiol. 2014;11:736.
18. Hassanpour S, Langlotz CP. Predicting high imaging utilization based on initial radiology reports: a feasibility study of machine learning. Acad Radiol. 2016;23:84.
19. Muelly M, Vasanawala S. MRI schedule optimization through discrete event simulation and neural networks as a means of increasing scanner productivity. In: Radiology Society of North America (RSNA) 102nd scientific assembly and annual meeting. Chicago, IL, November 2016.
20. Muelly M, Stoddard P, Vasanwala S. Using machine learning with dynamic exam block lengths to decrease patient wait time and optimize MRI schedule fill rate. In: International society for magnetic resonance in medicine. Honolulu, HI, April 2017.
21. Li X, Wang J, Fung RYK. Approximate dynamic programming approaches for appointment scheduling with patient preferences. Artif Intell Med. 2018;85:16.
22. Hills LS. How to handle patients who miss appointments or show up late. J Med Pract Manage. 2009;25:166.
23. Blumenthal DM, Singal G, Mangla SS, et al. Predicting non-adherence with outpatient colonoscopy using a novel electronic tool that measures prior non-adherence. J Gen Intern Med. 2015; 30:724.
24. Torres O, Rothberg MB, Garb J, et al. Risk factor model to predict a missed clinic appointment in an urban, academic, and underserved setting. Popul Health Manag. 2015;18:131.
25. Huang Y, Hanauer DA. Patient no-show predictive model development using multiple data sources for an effective overbooking approach. Appl Clin Inform. 2014;5:836.
26. Percac-Lima S, Cronin PR, Ryan DP, et al. Patient navigation based on predictive modeling decreases no-show rates in cancer care. Cancer. 2015;121:1662.
27. Harvey HB, Liu C, Ai J, et al. Predicting no-shows in radiology using regression modeling of data available in the electronic medical record. J Am Coll Radiol. 2017;14:1303.
28. Kurasawa H, Hayashi K, Fujino A, et al. Machine-learning-based prediction of a missed scheduled clinical appointment by patients with diabetes. J Diabetes Sci Technol. 2016;10:730.
29. Chang JT, Sewell JL, Day LW. Prevalence and predictors of patient no-shows to outpatient endoscopic procedures scheduled with anesthesia. BMC Gastroenterol. 2015;15:123.
30. Kaplan-Lewis E, Percac-Lima S. No-show to primary care appointments: why patients do not come. J Prim Care Community Health. 2013;4:251.

31. Miller AJ, Chae E, Peterson E, et al. Predictors of repeated "no-showing" to clinic appointments. Am J Otolaryngol. 2015;36:411.
32. AlRowaili MO, Ahmed AE, Areabi HA. Factors associated with no-shows and rescheduling MRI appointments. BMC Health Serv Res. 2016;16:679.
33. Curran JS, Halpert RD, Straatman A. Patient "no-shows" – are we scheduling failure? Radiol Manage. 1989;11:44.
34. Guzek LM, Fadel WF, Golomb MR. A pilot study of reasons and risk factors for "no-shows" in a pediatric neurology clinic. J Child Neurol. 2015;30:1295.
35. Norbash A, Yucel K, Yuh W, et al. Effect of team training on improving MRI study completion rates and no-show rates. J Magn Reson Imaging. 2016;44:1040.
36. Samuels RC, Ward VL, Melvin P, et al. Missed appointments: factors contributing to high no-show rates in an urban pediatrics primary care clinic. Clin Pediatr (Phila). 2015;54:976.
37. McMullen MJ, Netland PA. Lead time for appointment and the no-show rate in an ophthalmology clinic. Clin Ophthalmol. 2015;9:513.
38. Trivedi H, Mesterhazy J, Laguna B, et al. Automatic determination of the need for intravenous contrast in musculoskeletal MRI examinations using IBM Watson's natural language processing algorithm. J Digit Imaging. 2018;31:245.
39. Rothenberg S, Patel J, Herschu M. Evaluation of a machine-learning approach to protocol MRI examinations: initial experience predicting use of contrast by neuroradiologists in MRI protocols. In: Radiology Society of North America (RSNA) 2012nd scientific assembly and annual meeting. Chicago, IL, November 2016.
40. Sohn J, Trivedi H, Mesterhazy J. Development and validation of machine learning based natural language classifiers to automatically assign MRI abdomen/pelvis protocols from free-text clinical indications. In: Society of Imaging Informations in Medicine (SIIM) annual meeting. Pittsburgh, PA, June 2017.
41. Brown AD, Marotta TR. Using machine learning for sequence-level automated MRI protocol selection in neuroradiology. J Am Med Inform Assoc. 2018;25:568.
42. Lee YH. Efficiency improvement in a busy radiology practice: determination of musculoskeletal magnetic resonance imaging protocol using deep-learning convolutional neural networks. J Digit Imaging. 2018;31:604–10.
43. Hyun CM, Kim HP, Lee SM, et al. Deep learning for undersampled MRI reconstruction. Phys Med Biol. 2018;63:135007.
44. Zhu B, Liu JZ, Cauley SF, et al. Image reconstruction by domain-transform manifold learning. Nature. 2018;555:487.
45. Eo T, Jun Y, Kim T, et al. KIKI-net: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images. Magn Reson Med. 2018;80:2188–201.
46. Golkov V, Dosovitskiy A, Sperl JI, et al. q-Space deep learning: twelve-fold shorter and model-free diffusion MRI scans. IEEE Trans Med Imaging. 2016;35:1344.
47. Hammernik K, Klatzer T, Kobler E, et al. Learning a variational network for reconstruction of accelerated MRI data. Magn Reson Med. 2018; 79:3055.
48. Quan TM, Nguyen-Duc T, Jeong WK. Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss. IEEE Trans Med Imaging. 2018;37:1.
49. Yang G, Yu S, Dong H, et al. DAGAN: deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. IEEE Trans Med Imaging. 2018;37:1.
50. Zaharchuk G, Gong E, Wintermark M, et al. Deep learning in neuroradiology. Am J Neuroradiol. 2018; https://doi.org/10.3174/ajnr.A5543.
51. Gong E, Pauly JM, Wintermark M, et al. Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI. J Magn Reson Imaging. 2018;48:330–40.
52. Geyer LL, Schoepf UJ, Meinel FG, et al. State of the art: iterative CT reconstruction techniques. Radiology. 2015;276:339.
53. Patino M, Fuentes JM, Singh S, et al. Iterative reconstruction techniques in abdominopelvic CT: technical concepts and clinical implementation. Am J Roentgenol. 2015;205:W19.
54. Ledig C, Theis L, Huszar F, et al. Photo-realistic single image super-resolution using a generative adversarial network. 2016. CoRR, abs/1609.04802.
55. Dong C, Loy CC, He K, et al. Image super-resolution using deep convolutional networks. IEEE Trans Pattern Anal Mach Intell. 2016;38:295.
56. Hayat K. Super-resolution via deep learning. 2017. CoRR, abs/1706.09077.
57. Johnson J, Alahi A, Li F-F. Perceptual losses for real-time style transfer and super-resolution. 2016. CoRR, abs/1603.08155.
58. Lim B, Son S, Kim H, et al. Enhanced deep residual networks for single image super-resolution. 2017. CoRR, abs/1707.02921.
59. Shi W, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. 2016. CoRR, abs/1609.05158.
60. Sajjadi MSM, Schölkopf B, Hirsch M. EnhanceNet: single image super-resolution through automated texture synthesis. 2016. CoRR, abs/1612.07919.
61. Chen H, Zhang Y, Zhang W, et al. Low-dose CT via convolutional neural network. Biomed Opt Express. 2017;8:679.

62. Chen H, Zhang Y, Kalra MK, et al. Low-dose CT with a residual encoder-decoder convolutional neural network. IEEE Trans Med Imaging. 2017;36:2524.
63. Yasaka K, Katsura M, Akahane M, et al. Model-based iterative reconstruction for reduction of radiation dose in abdominopelvic CT: comparison to adaptive statistical iterative reconstruction. Springerplus. 2013;2:209.
64. Moloney F, Twomey M, Fama D, et al. Determination of a suitable low-dose abdominopelvic CT protocol using model-based iterative reconstruction through cadaveric study. J Med Imaging Radiat Oncol. 2018; https://doi.org/10.1111/1754-9485.12733.
65. Murphy KP, Crush L, O'Neill SB, et al. Feasibility of low-dose CT with model-based iterative image reconstruction in follow-up of patients with testicular cancer. Eur J Radiol Open. 2016;3:38.
66. Cross NM, DeBerry J, Ortiz D, et al. Diagnostic quality of machine learning algorithm for optimization of low-dose computed tomography data. In: SIIM (Society for Imaging Informatics in Medicine) Annual Meeting, 2017
67. Xu J, Gong E, Pauly JM, et al. 200x low-dose PET reconstruction using deep learning. 2017. CoRR, abs/1712.04119.
68. Yang Q, Li N, Zhao Z, et al. MRI image-to-image translation for cross-modality image registration and segmentation. 2018. CoRR, abs/1801.06940.
69. Liu F, Jang H, Kijowski R, et al. Deep learning MR imaging–based attenuation correction for PET/MR imaging. Radiology. 2018;286:676.
70. Han X. MR-based synthetic CT generation using a deep convolutional neural network method. Med Phys. 2017;44:1408.
71. Wolterink JM, Dinkla AM, Savenije MHF, et al. Deep MR to CT synthesis using unpaired data. 2017. CoRR, abs/1708.01155
72. Ben-Cohen A, Klang E, Raskin SP, et al. Cross-modality synthesis from CT to PET using FCN and GAN networks for improved automated lesion detection. 2018. CoRR, abs/1802.07846
73. Li R, Zhang W, Suk H-I, et al. Deep learning based imaging data completion for improved brain disease diagnosis. Presented at the Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014, Cham. 2014.
74. Esses SJ, Lu X, Zhao T, et al. Automated image quality evaluation of T2 -weighted liver MRI utilizing deep learning architecture. J Magn Reson Imaging. 2018;47:723.
75. Kohli MD, Summers RM, Geis JR. Medical image data and datasets in the era of machine learning-whitepaper from the 2016 C-MIMI meeting dataset session. J Digit Imaging. 2017; 30:392.
76. Wang KC, Patel JB, Vyas B, et al. Use of radiology procedure codes in health care: the need for standardization and structure. Radiographics. 2017;37:1099.
77. Bulu H, Sippo DA, Lee JM, et al. Proposing new RadLex terms by analyzing free-text mammography reports. J Digit Imaging. 2018;31:596–603.
78. Percha B, Zhang Y, Bozkurt S, et al. Expanding a radiology lexicon using contextual patterns in radiology reports. J Am Med Inform Assoc. 2018;25:679–85.
79. Duda J, Botzolakis E, Chen P-H, et al. Bayesian network interface for assisting radiology interpretation and education. Presented at the SPIE medical imaging. 2018.
80. Chen R, Wang S, Poptani H, et al. A Bayesian diagnostic system to differentiate glioblastomas from solitary brain metastases. Neuroradiol J. 2013;26:175.
81. http://www.acrinformatics.org/acr-assist
82. Rubin DL, Kahn CE Jr. Common data elements in radiology. Radiology. 2017;283:837.
83. Zech J, Pain M, Titano J, et al. Natural language–based machine learning models for the annotation of clinical radiology reports. Radiology. 2018;287:570.
84. Meystre SM, Savova GK, Kipper-Schuler KC, et al. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008; 128:44.
85. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2:230.
86. Pons E, Braun LM, Hunink MG, et al. Natural language processing in radiology: a systematic review. Radiology. 2016;279:329.
87. Tan WK, Hassanpour S, Heagerty PJ, et al. Comparison of natural language processing rules-based and machine-learning systems to identify lumbar spine imaging findings related to low back pain. Acad Radiol. 2018; https://doi.org/10.1016/j.acra.2018.03.008.
88. Chen PH, Zafar H, Galperin-Aizenberg M, et al. integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. J Digit Imaging. 2018;31:178.
89. Liu P, Pan X. Text summarization with TensorFlow. Google Blogs. 2016. https://ai.googleblog.com/2016/08/text-summarization-with-tensorflow.html
90. Rush AM, Chopra S, Weston J. A neural attention model for abstractive sentence summarization. 2015. CoRR, abs/1509.00685.
91. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. 2014. CoRR, abs/1409.3215.
92. Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: bridging the gap be-

tween human and machine translation. 2016. CoRR, abs/1609.08144

93. Walker J, Darer JD, Elmore JG, et al. The road toward fully transparent medical records. N Engl J Med. 2014;370:6.
94. Oh SC, Cook TS, Kahn CE Jr. PORTER: a prototype system for patient-oriented radiology reporting. J Digit Imaging. 2016;29:450.
95. Bossen JK, Hageman MG, King JD, et al. Does rewording MRI reports improve patient understanding and emotional response to a clinical report? Clin Orthop Relat Res. 2013;471:3637.
96. Ringler MD, Goss BC, Bartholmai BJ. Syntactic and semantic errors in radiology reports associated with speech recognition software. Health Inform J. 2017;23:3.
97. Quint LE, Quint DJ, Myles JD. Frequency and spectrum of errors in final radiology reports generated with automatic speech recognition technology. J Am Coll Radiol. 2008;5:1196.
98. Zhang Y, Pezeshki M, Brakel P, et al. Towards end-to-end speech recognition with deep convolutional neural networks. 2017. CoRR, abs/1701.02720.
99. Hannun AY, Case C, Casper J, et al. Deep speech: scaling up end-to-end speech recognition. 2014. CoRR, abs/1412.5567.
100. Zhang Z, Geiger JT, Pohjalainen J, et al. Deep learning for environmentally robust speech recognition: an overview of recent developments. 2017. CoRR, abs/1705.10874.
101. Zhang Y, Chan W, Jaitly N. Very deep convolutional networks for end-to-end speech recognition. Presented at the 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), 5–9 March 2017. 2017.
102. Blagev DP, Lloyd JF, Conner K, et al. Follow-up of incidental pulmonary nodules and the radiology report. J Am Coll Radiol. 2016;13:R18.
103. Cook TS, Lalevic D, Sloan C, et al. Implementation of an automated radiology recommendation-tracking engine for abdominal imaging findings of possible cancer. J Am Coll Radiol. 2017; 14:629.
104. Xu Y, Tsujii J, Chang EIC. Named entity recognition of follow-up and time information in 20 000 radiology reports. J Am Med Inform Assoc. 2012;19:792.
105. Lacson R, Desai S, Landman A, et al. Impact of a health information technology intervention on the follow-up management of pulmonary nodules. J Digit Imaging. 2018;31:19.
106. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, et al. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. npj Digital Medicine. 2018;1:9.
107. Yaniv G, Kuperberg A, Walach E. Deep learning algorithm for optimizing critical findings report turnaround time. In: SIIM (Society for Imaging Informatics in Medicine) Annual Meeting. 2018.
108. Baum R, Bertsimas D, Kallus N. Scheduling, revenue management, and fairness in an academic-hospital radiology division. Acad Radiol. 2014;21:1322.
109. Avrin D. Faculty scheduling. Acad Radiol. 2014;21:1223.
110. Boroumand G, Dave JK, Roth CG. Shedding light on the off-hours coverage gap in radiology: improving turnaround times and critical results reporting. House Staff Quality Improvement and Patient Safety Posters. Poster 64. Jefferson Digital Commons. 2017. http://jdc.jefferson.edu/patientsafetyposters/64
111. Lazzeri F, Lu H, Reiter I. Optimizing project staffing to improve profitability with Cortana Intelligence. In: Microsoft learning blog, vol. 2018. 2017. https://blogs.technet.microsoft.com/machinelearning/2017/03/30/optimizing-workforce-staffing-to-improve-profitability-with-cortana-intelligence/
112. Prevedello LM, Andriole KP, Hanson R, et al. Business intelligence tools for radiology: creating a prototype model using open-source tools. J Digit Imaging. 2010;23:133.
113. Cook TS, Nagy P. Business intelligence for the radiologist: making your data work for you. J Am Coll Radiol. 2014;11:1238.
114. Meenan C, Erickson B, Knight N, et al. Workflow lexicons in healthcare: validation of the SWIM lexicon. J Digit Imaging. 2017;30:255.
115. Muramatsu C. Overview on subjective similarity of images for content-based medical image retrieval. Radiol Phys Technol. 2018;11:109–24.
116. Miller DD, Brown EW. Artificial intelligence in medical practice: the question to the answer? Am J Med. 2018;131:129.
117. Duszak R Jr, Nossal M, Schofield L, et al. Physician documentation deficiencies in abdominal ultrasound reports: frequency, characteristics, and financial impact. J Am Coll Radiol. 2012;9:403.
118. Report, B. s. H. C. Combatting denials using machine intelligence: how it works and why now is the time for it, vol. 2018. 2015. https://www.beckershospitalreview.com/finance/combatting-denials-using-machine-intelligence-how-it-works-and-why-now-is-the-time-for-it.html
119. Hawkins CM, DeLa OA, Hung C. Social media and the patient experience. J Am Coll Radiol. 2016;13:1615.
120. Gefen R, Bruno MA, Abujudeh HH. Online portals: gateway to patient-centered radiology. AJR Am J Roentgenol. 2017;209:987.
121. Partridge SR, Gallagher P, Freeman B, et al. Facebook groups for the management of chronic diseases. J Med Internet Res. 2018;20:e21.
122. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classi-

fication models and saliency maps. 2013. CoRR, abs/1312.6034.

123. Dhurandhar A, Chen P-Y, Luss R, et al. Explanations based on the missing: towards contrastive explanations with pertinent negatives. 2018. CoRR, abs/1802.07623.

124. Kim W. Beyond interpretation. ACR DSI Blog. 2017. https://www.acrdsi.org/Blog/Beyond-Interpretation

12 Artificial Intelligence and Computer-Assisted Evaluation of Chest Pathology

Edwin J. R. van Beek and John T. Murchison

12.1 Introduction

Computer-assisted diagnosis for chest pathology has been attempted for a number of decades. Over half a century ago, Lodwick et al. described the concept of computer coding of radiographic findings as a means of exploring the computer-aided diagnosis and management of lung cancer [1]. Some systems have made it into clinical applications, whereas others have failed to make a breakthrough until relatively recently. The current combination of the pressures on radiology in terms of workload, the increasing complexity of examinations, the need for accurate and quantitative diagnostic pathways and the key role of imaging in the overall management of patients create a situation of heightened enthusiasm to explore novel methods to be introduced into the central reporting set-up. It will be essential for any software to form a part of this pathway, to allow optimal utilisation and implementation of its findings.

In spite of many efforts, the current implementation of these tools is still in its infancy. This is due to strict regulations when implementing new tools towards the diagnostic management of patients, as well as the lack of integration of software tools in clinical workflow. However, more recently, there has been renewed and heightened interest in making applications that are both accurate and provide integration into normal workflow, thus adding value in terms of efficiency and quantifiable data.

In this chapter, we will explore a number of areas where there is a history of attempted computer-assisted diagnosis as well as where developments of increasing diagnostic detection and interpretation of imaging datasets are undertaken. By its very nature, and the current speed of developments, this will be far removed from an all-inclusive review!

12.2 General Chest Radiography

Chest radiographs still form the vast majority of investigations for the evaluation of chest pathology. This is due to the high frequency of both lung and heart pathology, which can be gleaned from interpretation of a simple chest radiograph. It is generally accurate in determining the overall status of the heart

E. J. R. van Beek (✉) · J. T. Murchison
Edinburgh Imaging, University of Edinburgh, Edinburgh, UK

Department of Radiology, Royal Infirmary of Edinburgh, Edinburgh, UK
e-mail: Edwin-vanbeek@ed.ac.uk; john.murchison@nhslothian.scot.nhs.uk

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_12

and lungs, while it is also necessary to check for medical support, such as pacemakers, lines and tubes. The flip side of the high number of chest radiographs is that they form a significant burden on reporting workflow, as the complexities of CT, MRI and other imaging techniques take up more and more time. Thus, improvements of interpretation of chest radiographs, either by replacing the need for radiologist interpretation or by aiding in answering repeated questions (e.g. line or tube placement) would lead to an immediate impact on care provision.

In a large study in rural Africa, involving 46,099 participants for screening of tuberculosis, formed the basis for a performance comparison between an automatic software to general practitioners in the field and central reporting radiologists [2]. Sputum cultures were used as a reference method. In 39,391 subjects, no further action was taken (asymptomatic and normal chest X-ray at field reading). Of 6708 subjects with symptoms or abnormal chest radiograph, culture results were unavailable in 585 subjects. Thus, culture examination results were available in 6123 subjects, of which 231 had confirmed tuberculosis. The study used 25,805 chest radiographs from the study, of which 24,296 had field reading results available. There were 20,025 cases where tuberculosis was excluded, while 2454 had symptoms (2239), abnormal chest radiograph (1098) or both (917). Of these, central readings led to normal chest radiograph in 2980, tuberculosis (148) or non-tuberculosis abnormality (1125), and ultimately 106 subjects had culture confirmed tuberculosis. The study demonstrated that the diagnostic accuracy of all three methods was very closely aligned, and the performance of the automated software was comparable to that of central expert readers. Therefore, the introduction of this system for application in a rural setting would be feasible and, in fact, has now been introduced in tuberculosis-prevalent, underserved areas in Africa. This allows immediate interpretation and patient care decisions to be made at the point of care, which is particularly useful in remote areas.

Paediatrics chest radiographs are performed with the minimal amount of radiation, compounding the fact that these tend to be small, premature infants requiring high-level supportive care. Patients who require ventilation are commonly encountered, and interpretation of lines and tubes may be difficult. Automated software to interpret the presence and position of endotracheal (ET) tubes are a potentially very useful adjunct, and one study in 1344 radiographs (528 with ET tubes and 816 without ET tubes) demonstrated a high detection rate of ET tube location of 94% with a small overall mean distance error of less than 2 mm and with 86% of cases a distance error of less than 5 mm [3]. It is therefore feasible that for the simple question of ET placement, an automated report could be issued to the bedside team, without reporting delays by radiology workflow.

Acute respiratory distress syndrome (ARDS) in children is a life-threatening condition due to a combination of sepsis, pneumonia and pulmonary oedema, with a high mortality rate [5]. Early intervention with lung protective ventilator support can significantly improve outcomes, and early recognition is therefore of utmost importance [5]. An automated computer-aided tool was developed, based on automatic segmentation of the rib cage and texture analysis of intercostal patches of the lungs, which was able to be tested on 90 chest radiographs [4]. The system was able to detect ARDS with high sensitivity (91%) and specificity (86%), but further validation studies will be required to prove its accuracy in clinical routine.

12.3 Lung Nodules

The detection and characterisation of lung nodules are a vital part of chest radiographic workflow. The need arises out of a number of pathologies, where nodules are a major discriminator of disease, ranging from infection to primary and metastatic lung cancer. Work has been ongoing for a number of decades to develop tools that can assist in the detection, both using chest radiographs and computed tomographic methods.

12.3.1 Chest Radiography

The miss rate of lung cancer was evaluated in a study of 259 patients with proven non-small cell lung cancer, who had presented with a nodular lesion [6]. Of these, 19% were initially missed, and missed lesions were smaller (mean 19 mm) and more frequently had superimposing structures. Due to diagnostic delays, 43% of patients were in a higher stage (from T1 to T2) at the final time of diagnosis. Another study from the same group demonstrated that, contrary to general belief, there was limited impact on diagnostic accuracy when including previous chest radiographs in the evaluation, nor were there significant improvements when radiographs were read by two observers [7].

The introduction of digital radiography has altered the way in which we can handle imaging data and brought the potential image manipulation and imaging interpretation into a computer-accessible domain. As a result, there have been various approaches to change the perceptual interactions of chest radiographs.

Two of these methods have been the introduction of performing dual-energy subtraction and temporal subtraction or a combination of these techniques. These are extremely well described in an excellent review on this topic [8]. The introduction of these methods has certainly improved the detection rates of both calcified and non-calcified lung nodules.

In one study, a commercial CAD software (Riverain Medical, Miamisburg, OH, USA) was applied to a series of studies of patients where lung cancers had been missed on chest radiographs [9]. A total of 89 lung cancers were missed in 114 radiographs, and these images were re-analysed using the CAD tool, which detected 46 or the 89 missed lesions (52%). In this group, 3.8 false-positive indicators were given for each true-positive lung cancer. The authors also used a control group of 89 similar studies without lung nodules, which resulted in false-positive results with a mean of 2.4 per radiograph. It is important to note that the missed lung cancer rate was highest in non-subspeciality trained radiologists, who were responsible for 99 (88%) of the overlooked cases.

In another study, a bone suppression method (Riverain Medical, Miamisburg, OH, USA) was evaluated in comparison with dual-energy radiography in 50 patients with 55 confirmed primary nodular cancers and 30 patients without cancer [10]. All studies were evaluated twice at 1-year intervals by ten observers with various clinical experiences. The reading accuracy improved with ROC curve area under the curve improving from 0.807 for standard chest radiograph to 0.867 for standard radiograph with bone suppression and 0.916 for standard radiograph with dual-energy subtraction. Importantly, the same statistically significant improvements were observed for the most experienced radiologists. Nevertheless, the authors concluded that bone suppression methodology, although not quite as good as dual-energy subtraction radiography, offers advantages in terms of costs (no special equipment required), reading time efficiency and radiation dose (which is higher for dual-energy subtraction).

In 45 patients with chest CT proven solitary lung nodules ranging from 8 to 25 mm and 45 normal controls, a commercial CAD system (EpiSight/XR, DEUS Technologies, Rockville, MD, USA) demonstrated a significant improvement of the ROC area under the curve from 0.924 (the average of 8 observers) to 0.986 [11]. The system was particularly useful for less experienced radiologists and those in training.

The diagnostic accuracy of three radiologists was compared with a commercial CAD system (xLNA Enterprise, Philips Medical Systems, Hamburg, Germany) for detection of lung nodules on 117 chest radiographs, using computed tomography as reference standard [12]. In 75 patients, CT was without lung nodules, while 66 nodules were present in 42 patients. The CAD system had a sensitivity of 39% for nodule detection in the range of 5–15 mm, compared to 18–30% for the radiologists. There were 2.5 false-positive indicators on average.

Follow-up chest radiographs of 324 patients returning for any type of cancer formed the basis for the study of another commercial product

(IQQA Chest, EDDA Technologies, Princeton Junction, NJ, USA) [13]. This system is based on an algorithm that includes a nodule-specific enhancement, nodule segmentation and nodule analysis and subsequently suggests highlighted areas on the chest radiograph display (Fig. 12.1). The system is immediate and interactive, allowing the reporting radiologist to immediately accept or decline suggested markup, and the combined report was used to determine the reading accuracy. There were 214 patients with appropriate follow-up included in this study, and lung nodules were confirmed in 35 without CAD and in 51 with CAD, resulting in a significant improvement of sensitivity from 64% to 93% (Figs. 12.2 and 12.3). The false-positive rate increased from three to six cases, resulting in a non-significant decrease in specificity from 98% to 96%. There were 153 true-negative cases (71%).

Overall, chest radiographic methods for automated lung nodule detection have been introduced with several products in the market that are integrated within reporting workstations. The systems offer enhanced sensitivity at a price of 0–4 false-positive indicators and require a radiologist opinion for final reporting. Thus, they are improving efficiency of reporting for lung nodules by allowing faster reading times while offering a second reader approach for detection.

12.3.2 Computed Tomography

The use of computed tomography (CT) for lung cancer screening has paved the way for increased patient detection of early lung cancer [14] and appears cost-effective [15]. A subsequent follow-up article directly compared the performance of low-dose CT with chest radiography at the T1 and T2 rounds of the study and demonstrated much greater sensitivity of CT (94%) versus radiography (60%), although the positive predictive value was better with radiography (5%) versus CT (2%) [16]. This experience, together with that of many others, has led to the widespread support to introduce lung cancer screening in high-risk patients [17]. Apart from lung cancer screening, the utility of CT as a primary tool for staging of various cancers has also improved detection of lung metastases. This has resulted in a significant increase in demand for chest CT investigations.

The problem with the enhanced capabilities of CT over chest radiographs is that it also requires a much greater number of images to be evaluated, and this immense stream of imaging data renders the method prone to "missed" nodules by radiologists of all backgrounds. Indeed, in a study from the NELSON study, the initial baseline screen report was adjusted by expert chest radiologist in 195/2796 participants (5.9%) [18].

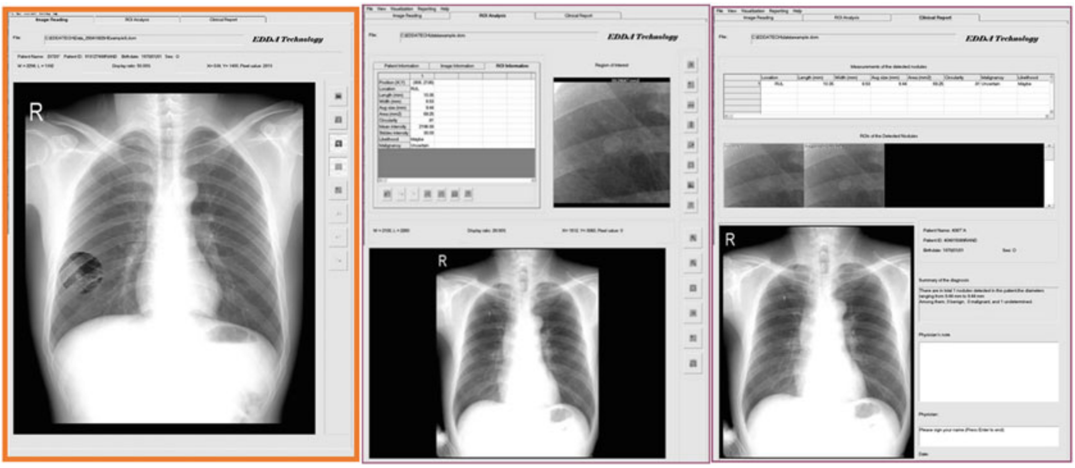


Fig. 12.1 Chest radiography computer-assisted diagnosis set-up for lung nodule detection, where potential nodules are highlighted, segmented and characterised. Reports can saved in PACS for follow-up purposes

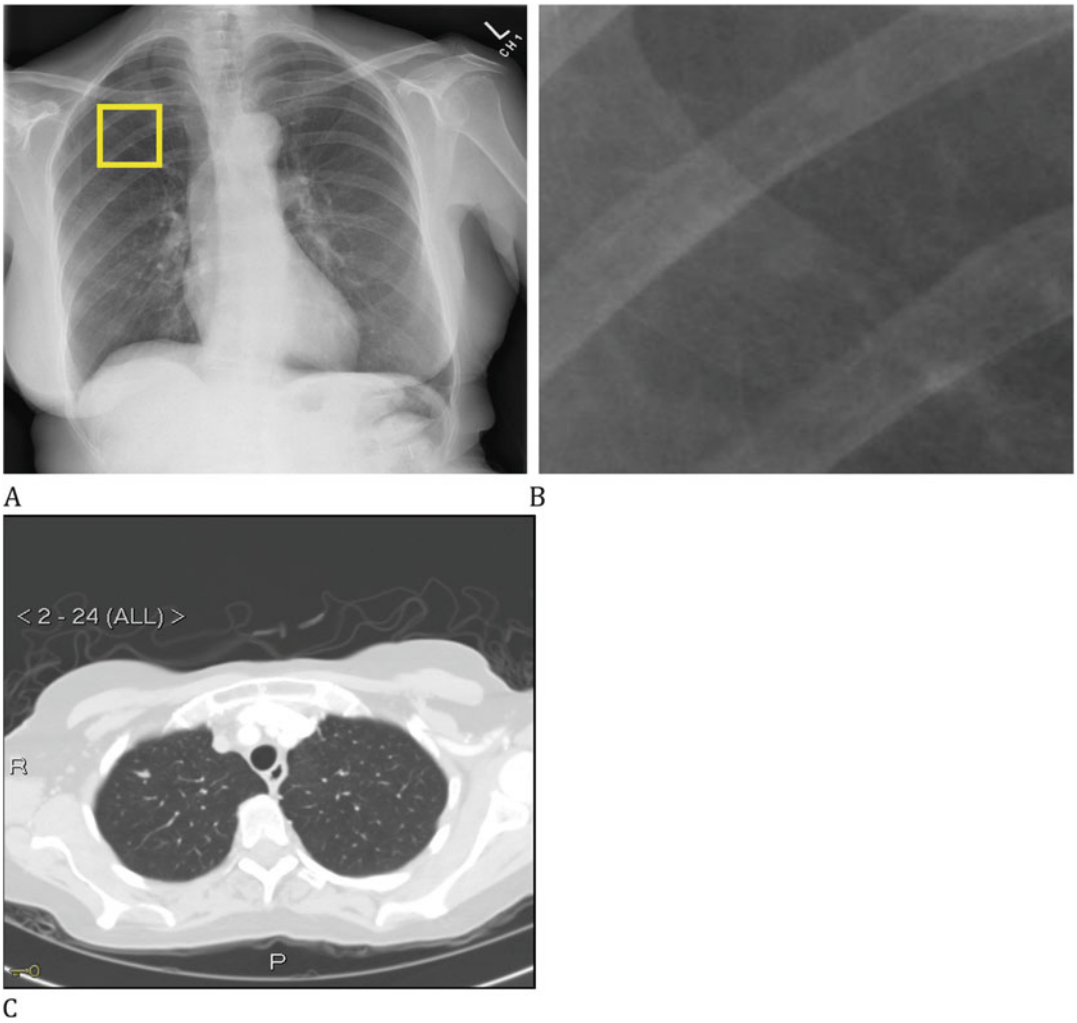


Fig. 12.2 Example of chest radiograph in a patient with renal cell cancer (**a**) with nodule highlighted by CAD system (**b**) and subsequently confirmed at CT (**c**)

In 95% of these incidents, lung nodules were downgraded and none subsequently developed lung cancer, thus leading to a decrease in false-positive results.

Based on the increased need, work pressures and the large number of negative results in a screening population, computer-assisted software systems would be an immense step forward to allow greater efficiency of reporting while allowing greater sensitivity. We need to realise that the number of images of a simple chest CT has increased from less than 30 slices, when single-detector systems were used, to more than 1500 images with the introduction of 1.25 mm slice thickness images using 64-multidetector systems or better. In combination with other chest CT indications, such as screening for lung metastases, the overall interest in developing tools to develop software tools to assist the detection and reporting of lung nodules has really taken off.

In a study evaluating the impact of a lung nodule detection software (GE Digital Contrast Agent, GE Healthcare, Waukesha, WI, USA), it was shown that radiologists had greater confidence in recording the presence or absence

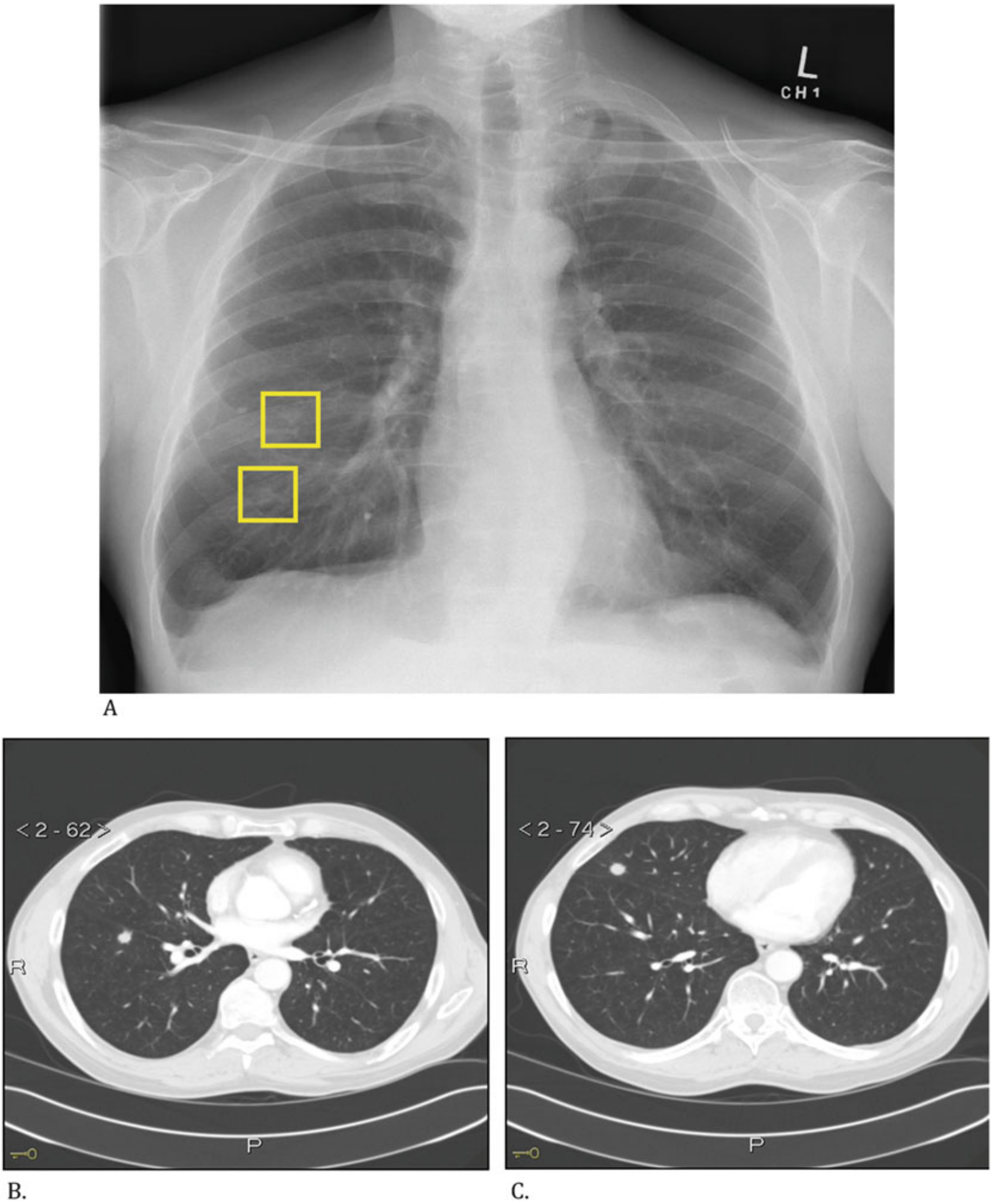


Fig. 12.3 Patient with head and neck squamous cancer (T3N2M0) at follow-up. CAD system identified two lung nodules (**a**), subsequently confirmed by CT (**b, c**)

of lung nodules, although this software did not improve diagnostic accuracy of lung nodule detection [19].

A small study of 20 outpatient CT scans with 195 non-calcified nodules greater than 3 mm was evaluated by three radiologists and subsequently by a CAD algorithm [20]. The study demonstrated that at a "cost" of three false-positive detections per CT scan, sensitivity significantly increased from a mean of 50% for the radiologists to 76% with application of CAD software.

Another small study in 25 patient CT scans with 116 nodules directly compared the performance of 2 commercially available CAD systems, ImageChecker CT (R2 Technologies, Sunnyvale, CA, USA) and Nodule Enhanced Viewing (Siemens Medical Solutions, Forchheim, Germany) [21]. This study showed that both systems performed similarly and improved sensitivity when used in addition to the radiologists.

Finally, four CAD systems were evaluated for potential detection of missed lung cancers in a screening trial, which manifested as solid nodules [22]. Cases were retrospectively identified at follow-up, as having been present on baseline CT scans. Of the total 50 lung cancers identified, the four CAD systems detected 56–70% at baseline and 74–82% at first follow-up, but there were 0.6–7.4 false positives at baseline. Thus, CAD could function as a second reader but would have missed more than 20% of lung cancers if used independently from radiologist reads.

An excellent review of the state of play in 2015 was written by Rubin, demonstrating the various tools available for nodule detection, ranging from image interpretation improvements using multiplanar reconstruction and maximum intensity projection with thicker slice thickness to utility of various CAD detection tools [23]. At that time, lung CAD was available but rarely used in routine clinical practice due to a combination of factors, including workflow, limited validation and limited overall gain in efficiency and detection of nodules that would truly affect patient's management.

It is important to realise that not all nodules are equal. For instance, calcified nodules can be disregarded as they represent old granulomata and don't require follow-up as they are benign. Nodules below a certain threshold similarly do not require further attention. This is relevant as most patients over the age of 30 will have some minor nodular changes, but these are inconsequential. Thus, the term "actionable nodule" has been coined, as these are those that are relevant with several recent guidelines pointing out the management of detected nodules [24, 25]. As part of these guidelines, one needs to consider that measurement of nodule size and volume are important parameters that directly affect the subsequent management of these lung nodules [26, 27].

Since the advent of machine learning tools into the development of dedicated detection software, more sophisticated systems are now making their way into the clinical domain. A recent version of computer-assisted diagnosis uses a combination of vessel subtraction and nodule detection to allow better identification of nodules [28]. A study in 324 cases (95 with proven cancer and 83 proven benign), derived from the National Lung Screening Trial dataset, demonstrated that the VIS/CADe (ClearRead CT, Riverain Technologies, Miamisburg, OH, USA) detected 89% of malignant and 82% of benign nodules of 5 mm or greater with a false-positive rate of 0.58 per study. This enhanced diagnosis tool improved the ROC area under the curve from 0.633 for unaided radiologist reads to 0.773 for aided detection of all nodules. For actionable nodules, the area under the curve improved from 0.584 to 0.692. This system improved lung cancer detection from 64% to 80%, while radiologist reporting time decreased by 26%. This system has now been approved by the FDA and is being implemented more widely for use as a second reader.

Another computer model was compared with 11 radiologists using a study group of 300 CT scans from the Danish Lung Cancer Screening Trial [29, 30]. The model was based on the Pan-Canadian Lung Cancer Screening Trial, which developed a risk stratification of pulmonary nodules [31]. The study cohort included 60 patients with proven lung cancer, 120 randomly selected participants with at least one benign nodule and 120 participants with at least one benign nodule in the range 3–16 mm with preference for nodules greater than 10 mm. There was no difference in overall risk assessment of malignant and benign nodules, but human observers were better at differentiating malignant from size-matched benign nodules. Importantly, the authors state that morphological criteria need to be addressed to develop more sophisticated software tools.

This latter point was further highlighted in a review article, demonstrating the importance of not just size but also of morphological characteristics such as those predicting benign features (calcification, internal fat, perifissural location and triangular shape) versus malignant features (speculation, lobulation, pleural indentation, vascular convergence, associated cystic airspace, irregular air bronchogram and part-solid appearance) [32].

In 2017, the Kaggle Data Science Bowl was held, inviting people to develop machine learning and artificial intelligence tools for the prediction of lung cancer diagnosis given a single chest CT investigation [33]. The competition made available a standard dataset from the National Cancer Institute to all participants to aid in the machine learning process of their algorithms and then a second dataset on which to validate their algorithm. The final test set was performed independently from the developers, and results were posted. Nearly 400 valid entries were received, and there were around five software tools that outperformed all others, demonstrating the wide differences between software tools.

One of these software tools recently received CE marking (Veye Chest, Aidence, Amsterdam, the Netherlands). Recently, several pilot studies were undertaken, demonstrating a very high FROC (free receiver operating characteristic), which was comparable to that of experienced readers (unpublished data). The Veye Chest product is an example of deep learning algorithms being applied to diagnostic tasks, in this case the detection and segmentation of pulmonary nodules. The detection task consists of two separate deep learning models to generate candidate locations and to filter out false-positive candidates. Two additional deep learning models determine the nodule composition and segmentation. This makes diameter and volume measurements available to the radiologist without the need for manual input. Furthermore, the design of deep learning-based detection models enables the output of a probability score for each nodule (not to be confused with a malignancy score). This allows a threshold to be applied on this probability score, so the user can select the optimal trade-off between sensitivity and false-positive rate to reflect the clinical situation. Some examples of this nodule detection software are given in Fig. 12.4.

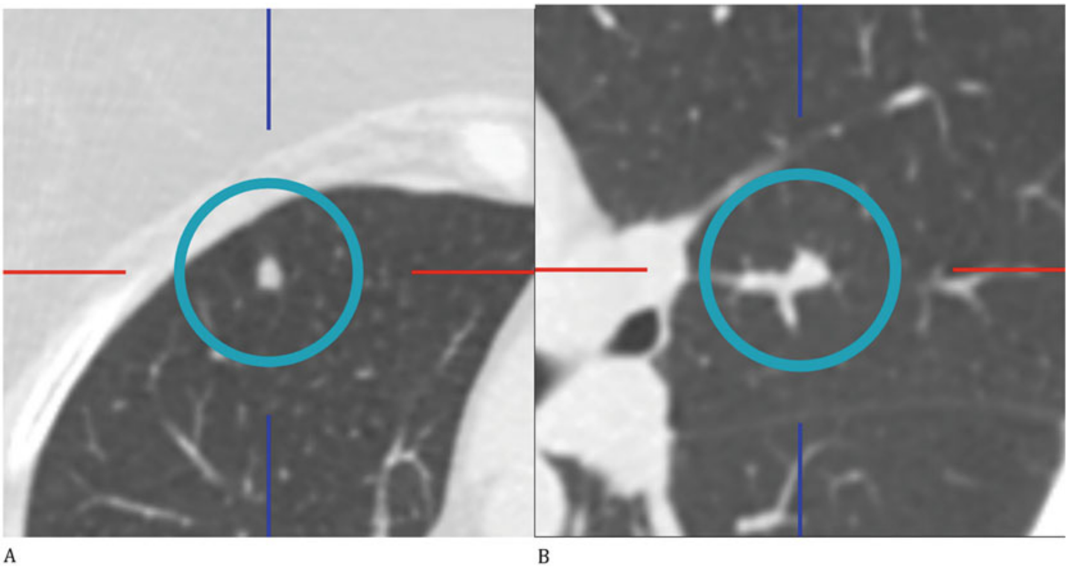


Fig. 12.4 Two examples of application of nodule detection software (VeyeChest, Aidence, Amsterdam, The Netherlands) where both the radiologist and the software detected a nodule (**a**) and where the software detected a nodule which had been missed by the radiologist (**b**)

12.4 Lung Cancer Radiomics

Another area of potential interest in relation to lung nodule and lung cancer evaluation is based on the previously mentioned point of morphological assessment. There is increasing evidence that the use of computer algorithms can derive additional data to enhance the predictive value of both the type of lung cancer, its level of aggressiveness and the subsequent stratification in treatment arms and long-term prognosis. We recently wrote a review of the importance of radiomics and its evolving role in lung cancer management [34]. Radiomics is defined as the field of study in which high-throughput data is extracted and large amounts of advanced quantitative imaging feature are analysed from medical imaging in the context of other pathological, genetic and clinical data (Fig. 12.5). A previous review demonstrated that a large number of features can be derived from CT data, which can be used to help differentiate the phenotype of cancers [35].

In a study of 41 pathologically classified resected lung adenocarcinomas, CT data were annotated for volumes, after which automatically generated computer measurements were performed: mass, volume and percentage solid tumour were used to model the probability of invasive non-lepidic adenocarcinoma, lepidic predominant adenocarcinoma and adenocarcinoma in situ/minimally invasive adenocarcinoma [36]. The authors were able to accurately differentiate tumour types based on multivariate models of percentage solid volume in 73% and based on percentage solid mass and total nodule mass in 76%.

Two studies based on imaging data from the National Lung Screening Trial also evaluated the potential role of morphological features derived from computer-added segmentation and characterisation. One study, which was a matched case-control sample of 186 CT studies with 4–20 mm non-calcified lung nodules, where biopsy was used for final diagnosis [37]. The datasets were split into a training set (70 cancers and 70 benign controls) and a validation set (20 cancers and 26 benign controls). The extracted 1342 image features including 1108 radiomics features as described by Aerts [35]. These included shape of the lesion, extra nodular features (e.g. emphysema, fibrosis, pleural thickening, vessel involvement, cavity, calcification), perinodular features (speculation, scar, calcification) and other features [percentage solid portion, margin coarseness, size of adjacent lymph nodes and summary cancer-like feature (size, attenuation, speculation, etc.)]. Based on this model, it was possible to have high diagnostic prediction for invasive versus noninvasive tumour types with a sensitivity of 95% and a specificity of 88% [37].

A nested control study based on the National Lung Screening Trial dataset, evaluated the radiologic features of small pulmonary nodules and the risk of lung cancer [38]. They included 73 patients with proven lung cancer and 157 control subjects who had three consecutive negative screening results. Nine features were significantly different, of which five were included in a predictive model: total emphysema score, attachment to vessel, nodule location, border definition and concavity. This model yielded a ROC area under the curve of 0.932, a specificity of 92% and a sensitivity of 76% for the identification of malignant nodules. Although in itself, this model will not predict all nodules, it can assist risk stratification in conjunction with radiologist interpretation.

A similar approach used only the volume, volume doubling time and volumetry-based diameter from the NELSON study cohort, consisting of 7155 participants with 9681 non-calcified nodules [39]. This study derived direct clinically meaningful guidelines, including the fact that nodules with a volume less than 100 mm^3 or less than 5 mm in diameter are not predictive for lung cancer. Furthermore, volume doubling times should be used for 100–300 mm^3 or 5–10 mm lung nodules only to have an impact on patient outcomes. Volume doubling times were helpful in predicting lung cancer, with doubling times of 600 days or more, 400–600 days and less than

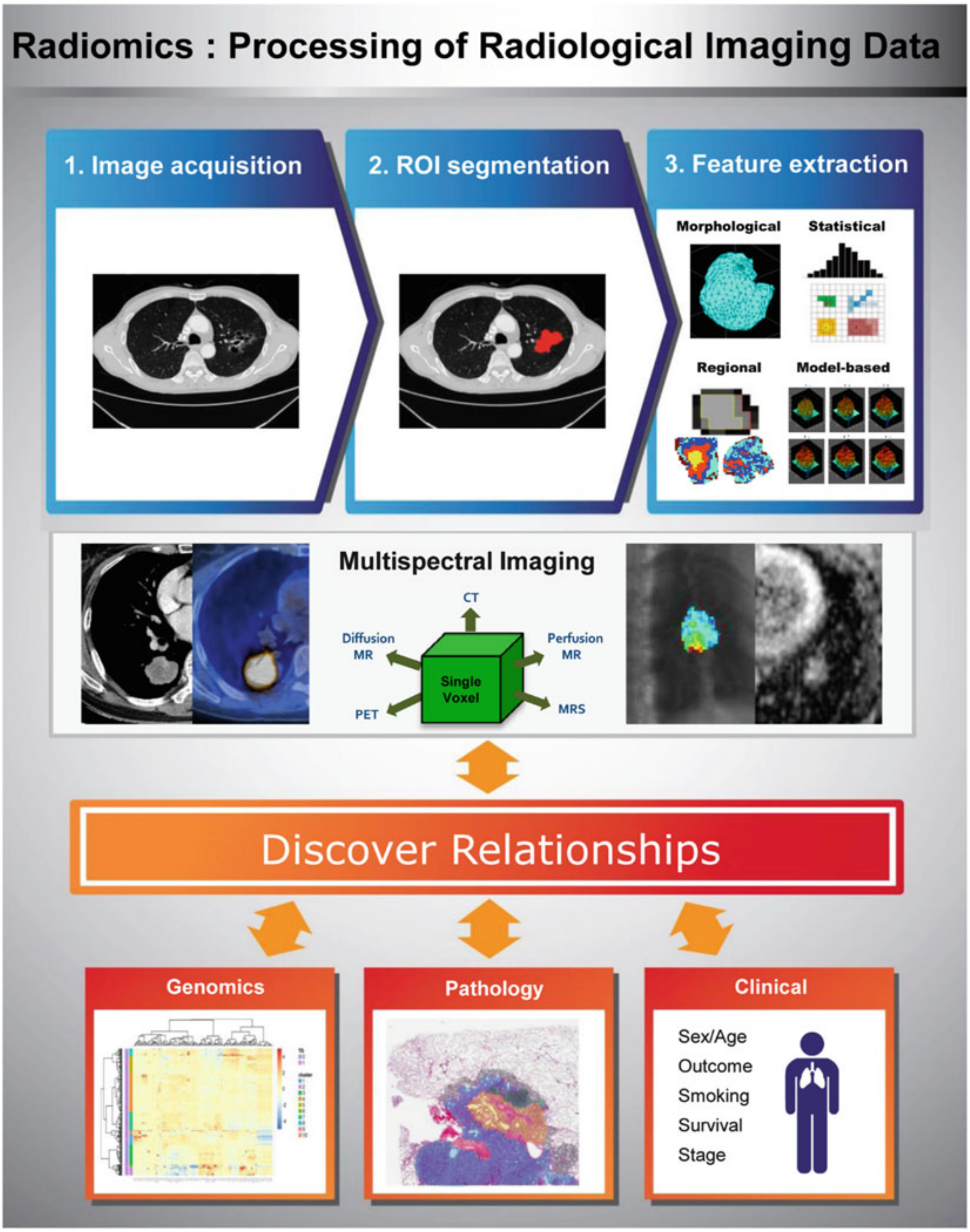


Fig. 12.5 Radiomics overview: quantitative features of lung cancer are extracted, and the information is combined with other imaging modalities to improve tumour characterisation. This enables the discovery of relationships with other tumour-related features. From: Lee G, Lee HY, Park H, et al. Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: state of the art. Eur J Radiol 2017;86:297–307

400 days yielding probabilities for lung cancer of 0.8%, 4.0% and 9.9%, respectively. Lesions that are 300 mm^3 or larger or with diameter 10 mm or larger lead to lung cancer in 16.9% and 15.2%, respectively.

In conclusion, the use of machine learning and artificial intelligence tools will increasingly yield improvements on nodule detection, will assist in prediction of malignancy versus benign and will guide clinicians as to the subsequent management of these findings. This will also allow better prognostication of individual patients.

12.5 Pulmonary Embolism

Pulmonary embolism remains a very frequent clinical problem and heavily relies on a combination of clinical likelihood assessment, plasma D-dimer testing and CT pulmonary angiography for diagnosis and exclusion [40]. However, the assessment of CT pulmonary angiograms is prone to misinterpretation for various reasons, including observer experience, quality of examinations and patient factors.

A study comparing different levels of expertise and background demonstrated that there was good interobserver agreement between radiologist and radiology residents and significantly less agreement among emergency physicians [41]. In addition, there was significant overreporting of pulmonary embolism by emergency physicians.

Two studies applied a commercial system (ImageCheckerCT, R2 Technology, Sunnyvale, CA, USA). One study used a set of 36 consecutive patients undergoing 16 multidetector row CT pulmonary angiography for suspected pulmonary embolism [42]. All studies were retrospectively reviewed by two experience radiologists, who identified 130 segmental emboli and 107 subsegmental emboli in 23 patients with pulmonary embolism. There were five patients with isolated subsegmental emboli. All 23 patients were positively identified as having PE, while vessel-by-vessel analysis demonstrated a sensitivity of 92% and 90% for segmental and subsegmental emboli, respectively.

The second, larger study prospectively enrolled 125 CT pulmonary angiogram studies using a less advanced 4 multidetector row CT scanner [43]. A total of 45 emboli were diagnosed in 15 patients, and 26 of these were confirmed by the software in 8 patients, with 19 emboli in 7 patients missed (sensitivity 42%). At the same time, there were 97 false-positive results (specificity 77%). The authors conclude that the system is insufficiently sensitive and causes a lot of false-positive result and requires modification for it to be clinically useful.

Another computer-aided diagnosis prototype (Version 3.0, Siemens Medical Solution, Malvern, PA, USA) for detection of pulmonary embolism was attempted for 16 and 64 multidetector row CT scanners, using 40 datasets containing 18 patients with pulmonary emboli and 22 normal examinations [44]. There were 212 expert panel confirmed emboli, of which 65 were centrally located, while 147 were in peripheral vessels. The studies were primarily read by 6 general radiologists, who detected 157 of 212 emboli (sensitivity 74%), with 97% of the central emboli and 70% of peripheral emboli detected, while 9 indicated emboli were considered false positive. The CAD software detected 74% of central and 82% of peripheral emboli, while 154 false positives (average 3.8) were indicated. The study suggested that CAD may improve the detection of peripheral emboli.

A comparison of computer-assisted diagnosis of pulmonary embolism (PE-CAD, Siemens Medical Solutions, Malvern, PA, USA) as a second reader in the context of level of experience of chest radiologist was undertaken in a prospective study of 56 patients with the use of 64 multidetector row CT [45]. The study demonstrated that the experienced readers outperformed the inexperienced reader. The software had the greatest impact on sensitivity in the least experienced readers. Yet, the addition of software-aided detection did improve the sensitivity of experienced readers in segmental and subsegmental arteries. The authors conclude that this software should be used as a second reader.

Several further studies have evaluated the impact of reader experience on CAD system performance. In one study, 79 CT pulmonary angiogram studies using a variety of CT systems ranging from 16 to 64 multidetector row were evaluated and then re-evaluated after 3 months with application of a prototype CAD system () by two inexperienced readers [46]. There were 32 positive studies with a total of 119 emboli of varying size and location. The sensitivity for the readers went up from 50% for all PE to 88% and 71%, respectively. There were an average of 3.27 false positives per patient which resulted in a small but not significant increase on false-positive rates by these inexperienced readers.

In another study with six readers of different experience, a software prototype (Philips Healthcare, Best, The Netherlands) was applied to 158 studies without and 51 studies with pulmonary embolism [47]. The cases were obtained using a combination of 16 and 64 multidetector row CT scanners. Both sensitivity and reading time were measured, which demonstrated a significant improvement of sensitivity without a decrease in specificity. The increased sensitivity was greatest for the least experienced reader. In 15 patients, pulmonary embolism would have been missed by readers, which were diagnosed once CAD was applied. There was a minimal increase in reading time (less than half a minute) associated with the CAD readout needing assessed by the readers.

A study in 43 patients with suspected pulmonary embolism retrospectively evaluated another iteration of the prototype software (Light viewer version 1.6, Siemens Medical Solutions, Malvern, PA, USA) using 16 multidetector row CT pulmonary angiography [48]. There was a high prevalence of pulmonary embolism (33 patients) with a total of 215 thrombi. The software sensitivity alone was 83% versus three radiologists 77–87%, but the combined radiologist with software sensitivity improved to 92–98%. As expected, subsegmental emboli had lowest sensitivity.

The largest study to date assessed 6769 consecutive CT pulmonary angiogram studies for the presence of pulmonary emboli that were missed at initial reporting [49]. There were 53 studies in which 146 pulmonary emboli were deemed missed by a panel of three experts, and a prototype software (PE-CAD, Siemens Medical Systems, Malvern, PA, USA) was applied. The software correctly identified 103 emboli, while offering two additional marks which also proved to be pulmonary emboli, for an overall sensitivity of 72% and a per-study sensitivity of 77%.

The impact of various technical conditions of CT pulmonary angiograms, including ECG gating (30) or non-gating (30), the use of dual-energy imaging (14) and image quality on the performance of a pulmonary embolism CAD system (PE-CAD version 7, Siemens Medical Systems, Forchheim, Germany) [50]. The study demonstrated excellent sensitivity for peripheral emboli detection (97%), which was not influenced by scanning conditions or overall image quality.

With the advent of more sophisticated reconstruction methods and techniques, new software tools have been developed. Dedicated software tools for application in dual-energy pulmonary CT angiography (lung PBV and lung vessels) were compared with standard available CAD software (PE-CAD, Siemens Medical Solutions, Forchheim, Germany) in 37 patients of whom 21 had proven pulmonary emboli at segmental level or more peripheral, while 16 had no pulmonary embolism [51]. The lung PBV software was used to depict peripheral defects of the colour-coded iodine maps, while the lung vessels were used to prove vessel enhancement based on pixel-by-pixel density analysis. The study demonstrated improved detection of peripheral pulmonary emboli when using the CAD software as well as the new software tools, with the best result obtained with lung vessels.

Another study evaluated the potential advancement of pulmonary embolism detection using iterative reconstruction methodology, when compared with filtered back projection [52]. Pulmonary CT angiography from 40 patients were collected and reconstructed using standard filtered back projection and 6 levels of a hybrid iterative reconstruction algorithm and subsequently evaluated using a standard CAD system (PE-CAD, Siemens Medical Solutions, Malvern, PA, USA). The study demonstrated that increasing iterative reconstruction resulted in a decrease of sensitivity, whereas specificity improved with fewer false negatives. The best result was obtained with the lowest level of iterative reconstruction.

The introduction of spectral detector CT allows the use of a single X-ray source with a dual layer, spectral detector, leading to the ability to perform dual-energy imaging retrospectively. Using a 128 multidetector row spectral CT system (IQon, Philips Healthcare, Best, The Netherlands) in 15 patients with CT pulmonary angiography and 18 patients with CT venography and a diagnosis of pulmonary embolism, a study evaluated the application of a commercial CAD tool (Pulmonary Artery Assessment, Intellispace Portal, version 9, Philips Healthcare, Best, The Netherlands) [53]. The authors found a positive impact on combining high-attenuation monoenergetic reconstruction (greater 500 HU) with CAD software, resulting in excellent sensitivity and decreased false positives when compared to standard reconstruction.

A different approach was used in a study that aimed to correlate automatically segmented cardiac chamber volume to the outcome of patients with pulmonary embolism [54]. Thus, although this software tool did not directly aim to visualize pulmonary embolism, it was applied as a predictor for mortality. Of 756 consecutive patients, segmentation was successful to determine cardiac chamber volumes in 636 patients (84% success rate) and was correlated with mortality (84 patients died within 30 days with a diagnosis of pulmonary embolism). The most predictive marker was a decreased left atrial volume of 62 ml or less, which increased the mortality risk 2.4-fold.

A similar study aimed to link a fully automatic CAD detected severity/extent of pulmonary emboli and the development of right heart failure in 557 patients with proven pulmonary embolism and without underlying cardiopulmonary disease [55]. The study compared to overall burden of emboli with the ratio of the right ventricle over the left ventricle (as a parameter of right heart strain) and demonstrated a significant correlation. The reporting time for radiologists decreased from 15 to less than 5 min.

More recently, it has been suggested that machine learning may be applied to accurately predict the presence of pulmonary embolism without any need for imaging [56]. In this study of 1427 patients at risk of pulmonary embolism, they extracted 28 diagnostic features, which were incorporated in a neural hypernetwork, which was able to predict those going on to develop pulmonary embolism in 94% of cases. Although this may be helpful to better identify patients at risk of developing pulmonary embolism, it will not aid in the actual diagnostic process of pulmonary embolism.

Although several clinical CAD systems exist for the identification of pulmonary embolism, they have been fairly limited in gaining access to routine clinical workloads. This may in part be due to the excellent diagnostic quality of CT pulmonary angiograms with later CT systems or the utility of easier visualisation of emboli, for instance, through dual-energy iodine mapping. It is important to recognise the various influences on the performance of these CAD systems, such as type of CT scanner, slice thickness and implementation by radiologists of different experience levels. The role for machine learning approaches to pulmonary embolism is not clear at present, but this should be a promising area of research.

12.6 Parenchymal Lung and Airways Diseases

Various software tools have been introduced to help classify and quantify lung parenchymal and airway diseases. Initially, a simple approach based on lung density was applied to derive the "emphysema index", and this was further subdivided into upper, middle and lower thirds and core and rind assessment [57]. This method was subsequently used in a large number of studies but most notably predicted the outcomes of the National Lung Treatment Trial [58]. Subsequent developments and improvements of software tools have allowed better visualisation with isovolumetric CT images offering better evaluation of both lung parenchyma and airways (Fig. 12.6).

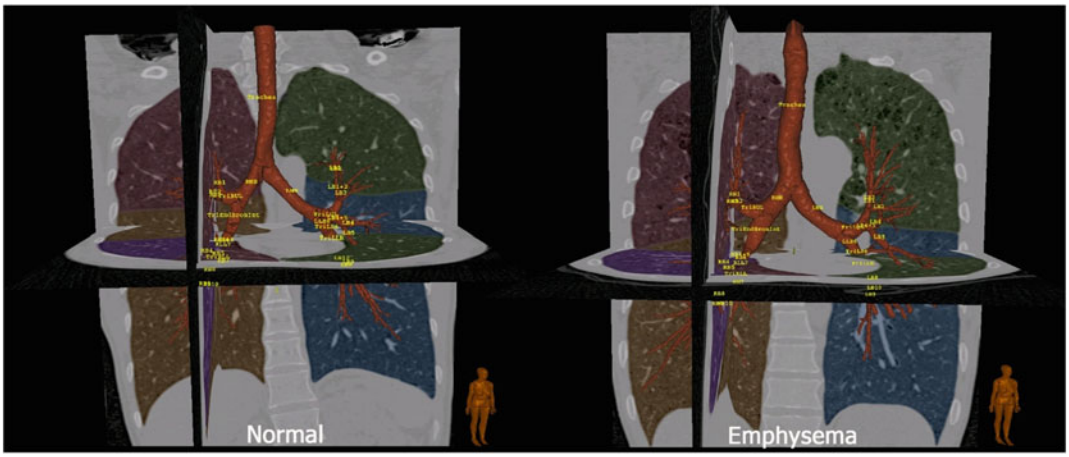


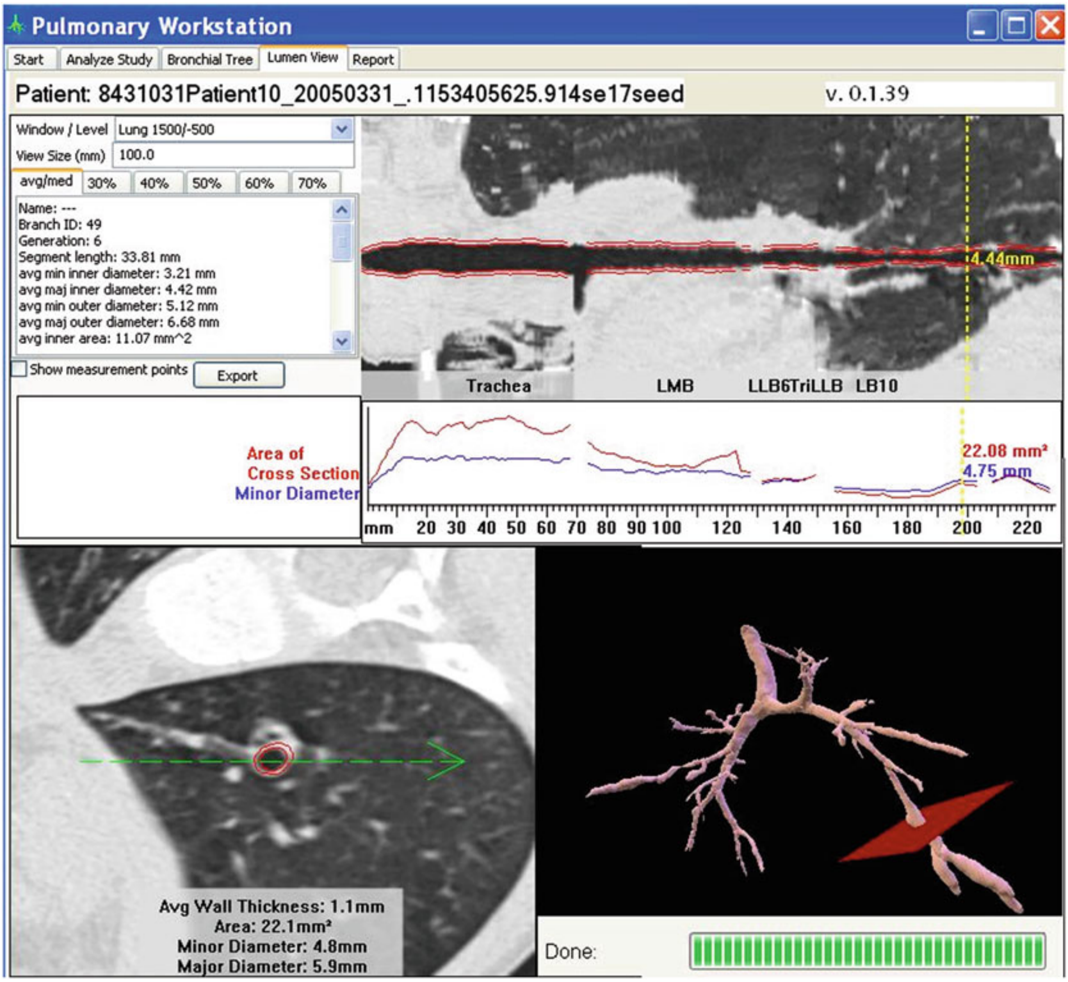
Fig. 12.6 Examples of isovolumetric reconstruction of 0.64 mm slice thickness CT images with automated labelling of airway tree and a normal subject and a patient with mild upper lobes emphysema (VIDA Diagnostics, Coralville, IA, USA)

The use of sophisticated software has increasingly been applied to study the pathophysiology of lung and airway diseases, often in conjunction with large-scale genetic profiling. Several large cohort studies, including COPDGene [59, 60], SPIROMICS [61] and MESA [62] have applied these software tools. In addition, the enhanced visualisation of airway remodelling and quantification of air trapping have formed a significant part of the SARP study in asthma [63, 64]. Similarly, this advanced software has been applied to bronchiectasis, which may be particularly helpful in patients with cystic fibrosis (Fig. 12.7). Although this software is increasingly widely available, it is not automated to an extent that it easily fits into workflow. Furthermore, it relies heavily on manual and expert input to derive meaningful results.

It is important to realise that there are many factors that can affect the lung density, which include level of inspiration (this can be overcome by active coaching of patients). Iterative reconstruction and kernel selection have a significant impact on quantitative density measurements [65]. Means to mitigate significant measurement shifts include combined use of high-frequency (bone) kernels with iterative reconstruction [66]. The benefit of a combination of bone kernel and iterative reconstruction is that it allows better visualisation and quantification of the airway tree.

A new method aimed to derive a biomarker based on voxel-wise subtraction of inspiratory and expiratory CT images, leading to a parametric response map (PRM) [67]. This technique uses two CT acquisitions, one at full inspiration and a low-dose study at expiration, which are segmented and then co-registration through a deformable algorithm. This then leads to a map based on density change on a voxel-by-voxel basis, which is displayed as a colour-coded map. The study used 194 CT scans from the 10,000 COPDGene study cohort, with varying GOLD status based on post-bronchodilator pulmonary function tests, and successfully classified these subjects based on the PRM method. This product is now fully automated and can be integrated into normal workflow (Lung Density Analysis, Imbio, Minneapolis, MN, USA).

The role of quantified CT imaging in emphysema, asthma and airway disease is clearly still developing. The advent of more automated assessment techniques will enable workflow integration, giving quantifiable information and distribution of disease to help better plan for individualised treatments.


Pulmonary Workstation

Start | Analyze Study | Bronchial Tree | Lumen View | Report

Patient: 8431031Patient10_20050331_.1153405625.914se17seed v. 0.1.39

Window / Level: Lung 1500/-500

View Size (mm): 100.0

avg/med 30% 40% 50% 60% 70%

Name: ---
Branch ID: 49
Generation: 6
Segment length: 33.81 mm
avg min inner diameter: 3.21 mm
avg maj inner diameter: 4.42 mm
avg min outer diameter: 5.12 mm
avg maj outer diameter: 6.68 mm
avg inner area: 11.07 mm^2

☐ Show measurement points Export

Avg Wall Thickness: 1.1mm
Area: 22.1mm²
Minor Diameter: 4.8mm
Major Diameter: 5.9mm

Done:

Fig. 12.7 Example of airway analysis in a low-dose CT study of a paediatric patient with cystic fibrosis, demonstrating airway measurements and distal bronchiectasis

12.7 Interstitial Lung Disease

Lung texture analysis has been developed for the quantification of interstitial lung disease to enable improvements over visual scoring methods. One system used a pixel-wise density analysis of lung CT images, which are classified using texture features and classified as fibrosis or no fibrosis [68]. This system was initially developed and tested in 129 patients with scleroderma-related pulmonary fibrosis and was able to detect at a threshold of 1% and quantify at a threshold of 25%.

Another method used a 3D voxel-based method and 3D adaptive multiple feature method. [69]. This method used expert observers to classify the voxels of interest, which were subsequently placed into a Bayesian support vector machine learning method. The method was subsequently employed in a large-scale pulmonary fibrosis trial (PANTER-IPF study) [70, 71]. This study used 355 CT scans from 211 subjects, which were all visually scored. Thirty-four investigations were technically inadequate, leaving 199 subjects for further analysis. The feasibility of performing quantified CT texture

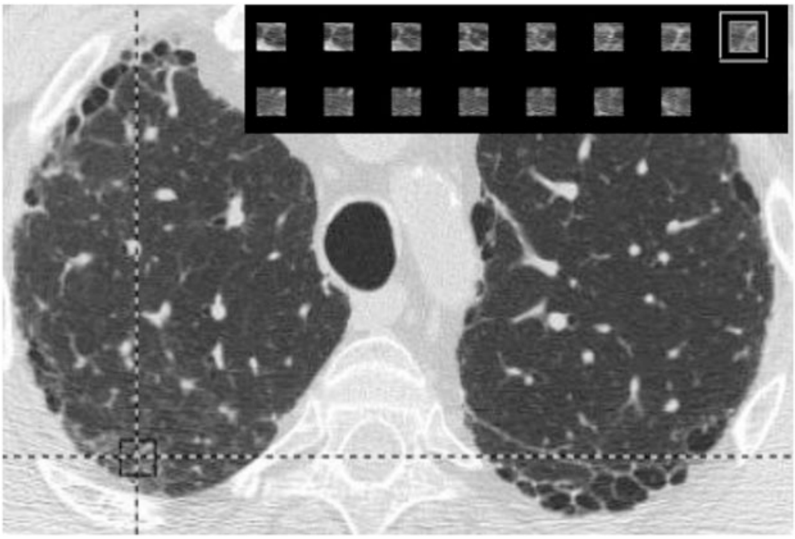


Fig. 12.8 Expert radiologist labelling of volumes of interest for 3D adaptive multiple features method (AMFM) training. The cross hear demonstrates the volume, with the individual slices within the region of interest in the top corner. Reprinted with permission of the American Thoracic Society. Copyright © 2018 American Thoracic Society. Salisbury ML, et al. Idiopathic pulmonary fibrosis: the association between the adaptive multiple features method and fibrosis outcome. Am J Respir Crit Care Med 2017;195:921–929. From: The American Journal of Respiratory and Critical Care Medicine is an official journal of the American Thoracic Society

analysis (Fig. 12.8) and its direct correlation with outcome was demonstrated, while it also showed that the AMFM method was slightly better at predicting event-free survival than visual assessment with a cut-off of 10% ground glass lung involvement [71].

A machine learning method retrospectively enrolled 280 subjects with baseline inspiratory CT and 72 with CT scans at least 15 months post baseline from the IPF Network relied on a data-driven texture analysis (DTA) approach [72]. Visual semi-quantitative scoring was performed, as well as a CT histogram analysis and a data-driven textural analysis, which is based on an unsupervised feature learning paradigm using a large number of 3 × 3 mm patches in lung CT scans of patients with pulmonary fibrosis and those with normal lungs (Fig. 12.9). The method was able to accurately specify changes in lung texture when compared to pulmonary function tests, including diffusing capacity for carbon monoxide. Incorporating this method into the CT assessment significantly improved prediction of lung function over this time period.

Another method, which has been incorporated into normal workflow, is a commercialised software (Lung Texture Analysis, Imbio, Minneapolis, MN, USA) [73]. This method uses a CT post processing technique (CALIPER), which allows for quantification of CT parenchymal patterns and was derived in 284 consecutive patients with pulmonary fibrosis. This software proved superior to visual scoring when compared to pulmonary function tests. A subsequent longitudinal study in 66 patients with follow-up between 6 and 24 months demonstrated that all computer variables (ground glass opacity, reticulation and honeycombing) all exhibited stronger links to forced vital capacity than visual scoring, but computer-derived pulmonary vessel volume was even more strongly related to predicting pulmonary function decline [74]. The software has excellent potential to demonstrate small and larger changes and appears more

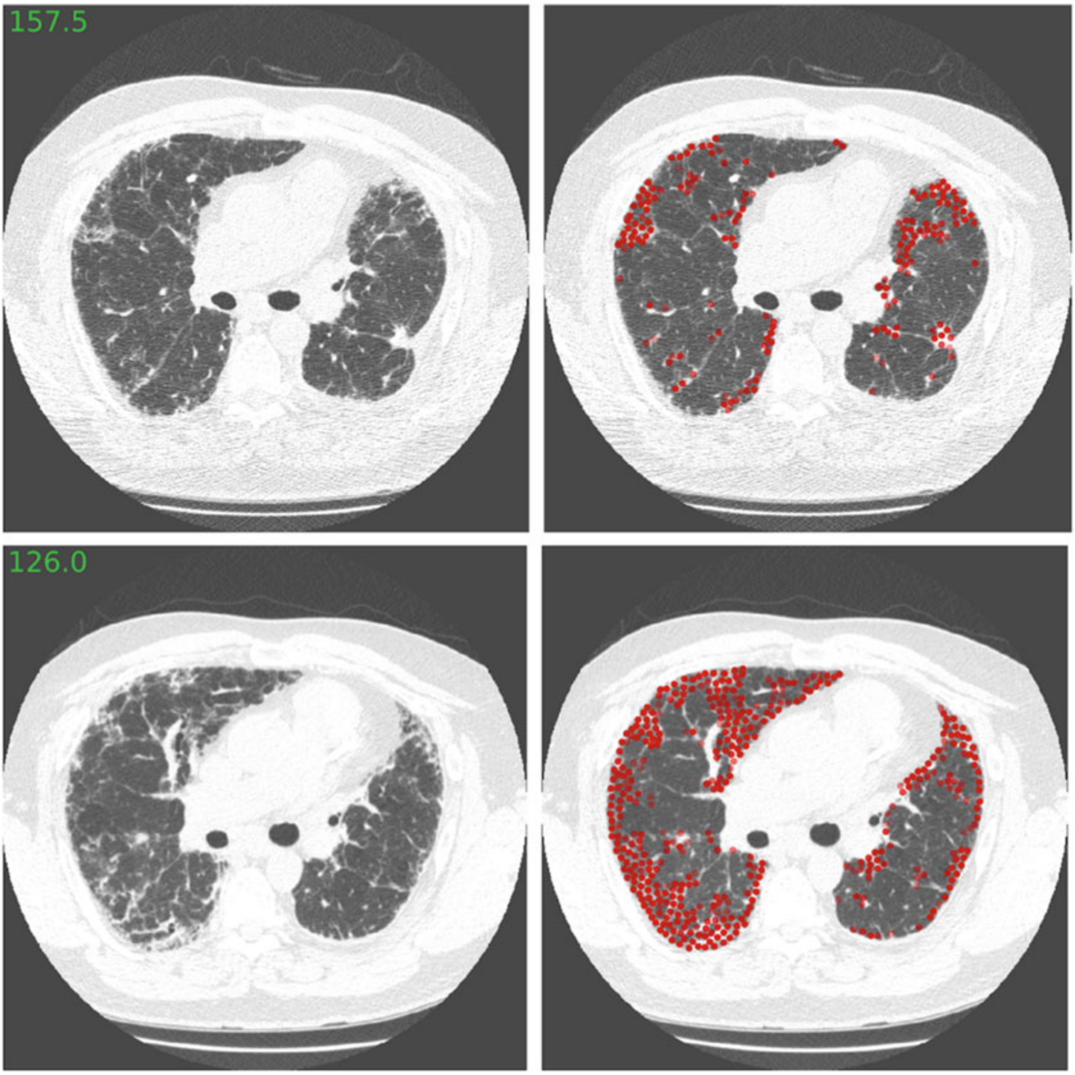


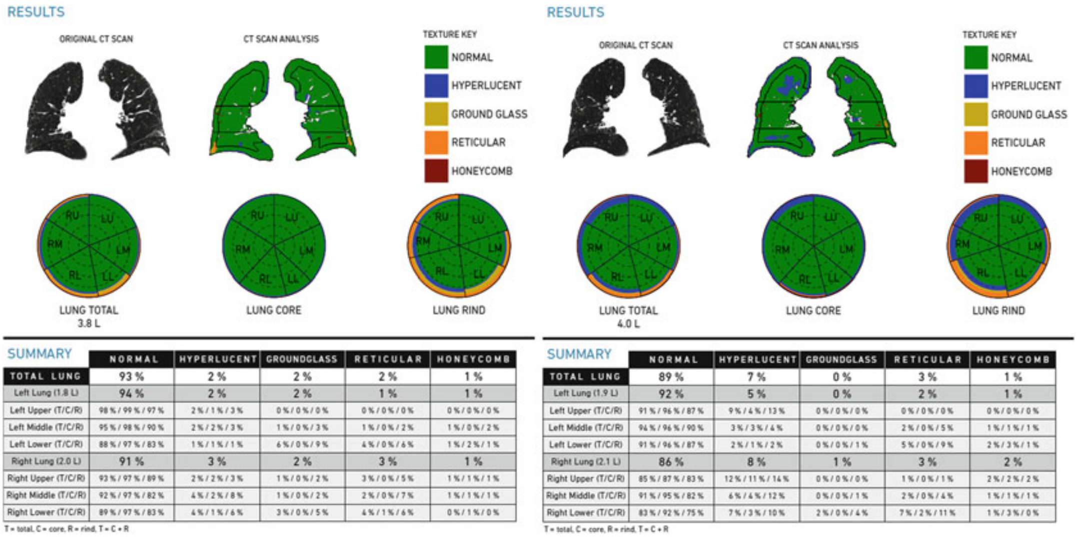
Fig. 12.9 Selected baseline image with fibrosis marked (top row) and follow-up study at 18 months (bottom row). The DTA method demonstrates 22% fibrosis at baseline and 41% at follow-up. From: Humphries SM, Yagihashi K, Huckleberry J, et al. Idiopathic pulmonary fibrosis: data-driven textural analysis of extent of fibrosis at baseline and 15-month follow-up. Radiology 2017;285:270–278. Copyright: Radiological Society of North America

sensitive than standard pulmonary function tests in demonstrating change (Figs. 12.10 and 12.11).

A recent review suggests that quantitative CT biomarkers for the evaluation of idiopathic pulmonary fibrosis is both necessary to often insight into the extent of disease and to allow detection of changes in response to novel treatments as a potential surrogate outcome and ultimately can lead to prognostication of individual patients [75].

12.8 Conclusions

The role of computer-assisted and machine learning/artificial intelligence software tools in chest imaging has only just begun. Some software tools



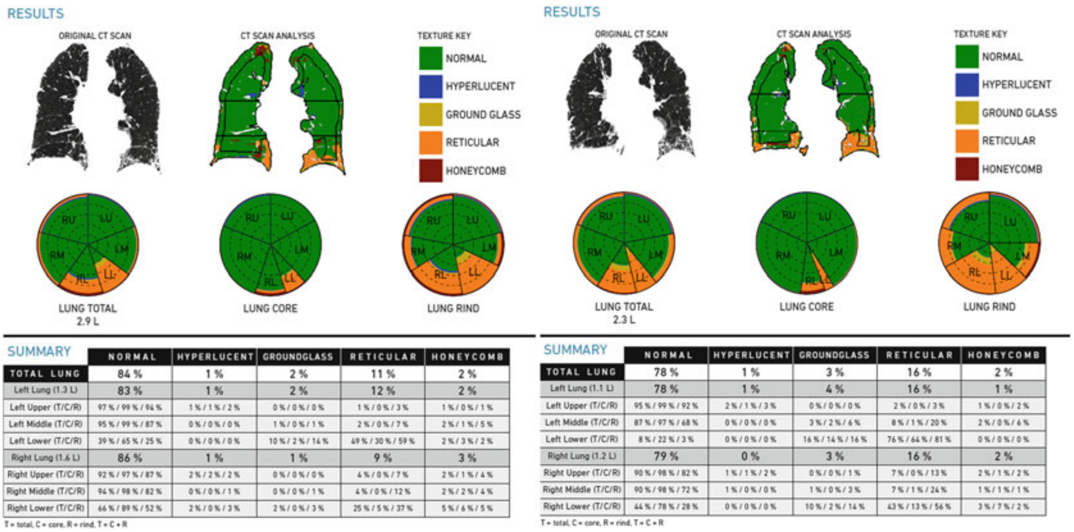
SUMMARY	NORMAL	HYPERLUCENT	GROUNDGLASS	RETICULAR	HONEYCOMB
TOTAL LUNG	93 %	2 %	2 %	2 %	1 %
Left Lung (1.8 L)	94 %	2 %	2 %	1 %	1 %
Left Upper (T/C/R)	98 % / 99 % / 97 %	2 % / 1 % / 3 %	0 % / 0 % / 0 %	0 % / 0 % / 0 %	0 % / 0 % / 0 %
Left Middle (T/C/R)	95 % / 98 % / 90 %	2 % / 2 % / 3 %	1 % / 0 % / 3 %	1 % / 0 % / 2 %	1 % / 0 % / 2 %
Left Lower (T/C/R)	88 % / 97 % / 83 %	1 % / 1 % / 1 %	6 % / 0 % / 9 %	4 % / 0 % / 6 %	1 % / 2 % / 1 %
Right Lung (2.0 L)	91 %	3 %	2 %	3 %	1 %
Right Upper (T/C/R)	93 % / 97 % / 89 %	2 % / 2 % / 3 %	1 % / 0 % / 2 %	3 % / 0 % / 5 %	1 % / 1 % / 1 %
Right Middle (T/C/R)	92 % / 97 % / 82 %	4 % / 2 % / 8 %	1 % / 0 % / 2 %	2 % / 0 % / 7 %	1 % / 1 % / 1 %
Right Lower (T/C/R)	89 % / 97 % / 83 %	4 % / 1 % / 6 %	3 % / 0 % / 5 %	4 % / 1 % / 6 %	0 % / 1 % / 0 %

T = total, C = core, R = rind, T = C + R

SUMMARY	NORMAL	HYPERLUCENT	GROUNDGLASS	RETICULAR	HONEYCOMB
TOTAL LUNG	89 %	7 %	0 %	3 %	1 %
Left Lung (1.9 L)	92 %	5 %	0 %	2 %	1 %
Left Upper (T/C/R)	91 % / 96 % / 87 %	9 % / 4 % / 13 %	0 % / 0 % / 0 %	0 % / 0 % / 0 %	0 % / 0 % / 0 %
Left Middle (T/C/R)	94 % / 96 % / 90 %	3 % / 3 % / 4 %	0 % / 0 % / 0 %	2 % / 0 % / 5 %	1 % / 1 % / 1 %
Left Lower (T/C/R)	91 % / 96 % / 87 %	2 % / 1 % / 2 %	0 % / 0 % / 1 %	5 % / 0 % / 9 %	2 % / 3 % / 1 %
Right Lung (2.1 L)	86 %	8 %	1 %	3 %	2 %
Right Upper (T/C/R)	85 % / 87 % / 83 %	12 % / 11 % / 14 %	0 % / 0 % / 0 %	1 % / 0 % / 1 %	2 % / 2 % / 2 %
Right Middle (T/C/R)	91 % / 95 % / 82 %	6 % / 4 % / 12 %	0 % / 0 % / 1 %	2 % / 0 % / 4 %	1 % / 1 % / 1 %
Right Lower (T/C/R)	83 % / 92 % / 75 %	7 % / 3 % / 10 %	2 % / 0 % / 4 %	7 % / 2 % / 11 %	1 % / 3 % / 0 %

T = total, C = core, R = rind, T = C + R

Fig. 12.10 Patient with mild interstitial pulmonary fibrosis. Lung function tests were FEV1 1.98 l (80% predicted), VC 3.04 l (93% predicted) and TCO 0.87 mmol/min/kPa (73% predicted). During the same time course of over 4 years, the pulmonary function tests remained stable, whereas quantitative CT assessment demonstrated minor progression



SUMMARY	NORMAL	HYPERLUCENT	GROUNDGLASS	RETICULAR	HONEYCOMB
TOTAL LUNG	84 %	1 %	2 %	11 %	2 %
Left Lung (1.3 L)	83 %	1 %	2 %	12 %	2 %
Left Upper (T/C/R)	97 % / 99 % / 94 %	1 % / 1 % / 2 %	0 % / 0 % / 0 %	1 % / 0 % / 3 %	1 % / 0 % / 1 %
Left Middle (T/C/R)	95 % / 99 % / 87 %	0 % / 0 % / 0 %	1 % / 0 % / 1 %	2 % / 0 % / 7 %	2 % / 1 % / 5 %
Left Lower (T/C/R)	39 % / 65 % / 25 %	0 % / 0 % / 0 %	10 % / 2 % / 14 %	49 % / 30 % / 59 %	2 % / 3 % / 2 %
Right Lung (1.6 L)	86 %	1 %	1 %	9 %	3 %
Right Upper (T/C/R)	92 % / 97 % / 87 %	2 % / 2 % / 2 %	0 % / 0 % / 0 %	4 % / 0 % / 7 %	2 % / 1 % / 4 %
Right Middle (T/C/R)	94 % / 98 % / 82 %	0 % / 0 % / 1 %	0 % / 0 % / 1 %	4 % / 0 % / 12 %	2 % / 2 % / 4 %
Right Lower (T/C/R)	66 % / 89 % / 52 %	2 % / 0 % / 3 %	2 % / 0 % / 3 %	25 % / 5 % / 37 %	5 % / 6 % / 5 %

T = total, C = core, R = rind, T = C + R

SUMMARY	NORMAL	HYPERLUCENT	GROUNDGLASS	RETICULAR	HONEYCOMB
TOTAL LUNG	78 %	1 %	3 %	16 %	2 %
Left Lung (1.1 L)	78 %	1 %	4 %	16 %	1 %
Left Upper (T/C/R)	95 % / 99 % / 92 %	2 % / 1 % / 3 %	0 % / 0 % / 0 %	2 % / 0 % / 3 %	1 % / 0 % / 2 %
Left Middle (T/C/R)	87 % / 97 % / 68 %	0 % / 0 % / 0 %	3 % / 2 % / 6 %	8 % / 1 % / 20 %	2 % / 0 % / 6 %
Left Lower (T/C/R)	8 % / 22 % / 3 %	0 % / 0 % / 0 %	16 % / 14 % / 16 %	76 % / 64 % / 81 %	0 % / 0 % / 0 %
Right Lung (1.2 L)	79 %	0 %	3 %	16 %	2 %
Right Upper (T/C/R)	90 % / 98 % / 82 %	1 % / 1 % / 2 %	0 % / 0 % / 1 %	7 % / 0 % / 13 %	2 % / 1 % / 2 %
Right Middle (T/C/R)	90 % / 98 % / 72 %	1 % / 0 % / 0 %	1 % / 0 % / 3 %	7 % / 1 % / 24 %	1 % / 1 % / 1 %
Right Lower (T/C/R)	44 % / 78 % / 28 %	0 % / 0 % / 0 %	10 % / 2 % / 14 %	43 % / 13 % / 56 %	3 % / 7 % / 2 %

T = total, C = core, R = rind, T = C + R

Fig. 12.11 Patient with moderate interstitial pulmonary fibrosis. Lung function tests were FEV1 1.74 (94% predicted), VC 2.3 l (94% predicted), TCO 3.29 mmol/min/kPa (73% predicted). After 2 years of follow-up, progressive abnormal function tests correlating with progressive quantitative CT: FEV1 0.99 (66% predicted), VC 1.08 l (58% predicted), TCO 2.29 (39% predicted)

are making their inroads into routine management, provided they can be incorporated into routine workflow and are time efficient (either by saving on reading time or by providing important data that assist patient management in a reasonable time frame). Many software tools are still stand alone, and these will need to transition into more automated mode, running in the background of routine workflow, and being available at the time of reporting.

It is clear that software tools can greatly enhance the radiologist's reporting, either by providing greater diagnostic accuracy and certainty or by yielding important quantifiable data that impact on patient management. Provided that these tools undergo rigorous clinical assessment prior to clinical introduction, they should be easily adopted by radiologists.

12.9 Take-Home Points

The utility of machine learning approaches for lung nodule detection and lung cancer diagnosis, including the utility of treatment advice and outcome prediction, will likely make a significant impact on how CT lung screening will be implemented.

The use of pulmonary embolism computer-assisted diagnosis, although feasible, has not been widely accepted due to the limited additional value and lack of workflow integration. However, the incorporation of CAD with deep learning for not only detection but also prognostication may be a very powerful adjunct in this field in the future. More research is needed.

The use of software tools to enhance the capabilities of chest radiographs is not only very promising but already shows direct impact on point of care diagnosis in a wide area of clinical application. It is likely that this field will expand and helps provide better health care (particularly in underserved areas).

Quantification of lung diseases is a primary target to assist in the development of novel treatments and for identification of patients who will benefit from them. Software tools are available already but will require adapting to be incorporated into routine clinical workflow. It is likely that these imaging biomarkers will play a huge role in how clinicians will use imaging into their clinical decision making.

References

1. Lodwick GW, Keats TE, Dorst JP. The coding of roentgen images for computer analysis as applied to lung cancer. Radiology. 1963;81:185–200.
2. Melendez J, Philipsen RHHM, Chanda-Kapata P, Sunkutu V, Kapata N, van Ginneken B. Automatic versus human reading of chest X-rays in the Zambia national tuberculosis prevalence study. Int J Tuberc Lung Dis. 2017;21:880–6.
3. Kao EF, Jaw TW, Li CW, Chou MC, Liu GC. Automated detection of endotracheal tubes in paediatric chest radiographs. Comput Methods Prog Biomed. 2015;118:1–10.
4. Zaglam N, Jouvet P, Flechelles O, Emeriaud G, Cheriet F. Computer-aided diagnosis system for the acute respiratory distress syndrome from chest radiographs. Comp Biol Med. 2014;52:41–8.
5. Zimmerman JJ, Akhtar SR, Caldwell E, Rubenfield GD. Incidence and outcomes of pediatric acute lung injury. Pediatrics. 2009;124:87–95.
6. Quekel LGBA, Kessels AGH, Goei R, van Engelshoven JMA. Miss rate of lung cancer on the chest radiograph in clinical practice. Chest. 1999;115:720–4.
7. Quekel LGBA, Goie R, Kessels AGH, van Engelshoven JMA. Detection of lung cancer on the chest radiograph: impact of previous films, clinical information, double reading, and dual reading. J Clin Epidemiol. 2001;54:1146–50.
8. MacMahon H, Li F, Engelmann R, Roberts R, Armato S. Dual energy subtraction and temporal subtraction chest radiography. J Thorac Imaging. 2008;23:77–85.
9. White CJ, Flukinger T, Jeudy J, Chen JJ. Detection system to detect missed lung cancer at chest radiography. Radiology. 2009;252:273–81.
10. Li F, Engelmann R, Pesce LL, Doi K, Metz CE, MacMahon H. Small lung cancers: improved detection by use of bone suppression imaging – comparison with dual-energy subtraction chest radiography. Radiology. 2011;261:937–49.
11. Kakeda S, Moriya J, Sato H, et al. Improved detection of lung nodules on chest radiographs using a commercial computer-aided diagnosis system. Am J Roentgenol. 2004;182:505–10.
12. Bley TA, Baumann T, Saueressig U, et al. Comparison of radiologist and CAD performance in the detection of CT-confirmed subtle pulmonary nodules on digital chest radiographs. Invest Radiol. 2008;43:343–8.
13. Van Beek EJR, Mullan B, Thompson B. Evaluation of a real-time interactive pulmonary nodule analysis system on chest digital radiographic images. A prospective study. Acad Radiol. 2008;15:571–5.
14. The National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365:395–409.
15. Black WC, Gareen IF, Soneji SS, et al. Cost-effectiveness of CT screening in the National Lung Screening Trial. N Engl J Med. 2014;371:1793–802.

16. Aberle DR, DeMello S, Berg CD, et al. Results of the two incidence screenings in the National Lung Screening Trial. N Engl J Med. 2013;369:920–31.
17. Oudkerk M, Deveraj A, Vliegenthart R, et al. European position statement on lung cancer screening. Lancet Oncol. 2017;18:e754–66.
18. Heuvelmans MA, Oudkerk M, de Jong PA, Mali WP, Groen HJM, Vliegenthart R. The impact of radiologists' expertise on screen result decisions in a CT lung cancer screening trial. Eur Radiol. 2015;25:792–9.
19. Nietert PJ, Ravenel JG, Taylor KK, Silvestri GA. Influence of nodule detection software on radiologists' confidence in identifying pulmonary nodules with computed tomography. J Thorac Imaging. 2011;26:48–53.
20. Rubin GD, Lyo JK, Palk DS, et al. Pulmonary nodules on multi-detector row CT scans: performance comparison of radiologists and computer-aided detection. Radiology. 2005;234:274–83.
21. Das M, Muhlenbruch G, Mahnken AH, et al. Small pulmonary nodules: effect of two computer-aided detection systems on radiologist performance. Radiology. 2006;241:564–71.
22. Liang M, Tang W, Xu DM, et al. Low-dose CT screening for lung cancer: computer-aided detection of missed lung cancers. Radiology. 2016;281:279–88.
23. Rubin GD. Lung nodule and cancer detection in computed tomography screening. J Thorac Imaging. 2015;30:130–8.
24. Callister MEJ, Baldwin DR, Akram AR, et al. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. Thorax. 2015;70(Suppl 2):ii1–ii54.
25. MacMahon H, Naidich DP, Goo JM, et al. Guidelines for management of incidental pulmonary nodules detected on CT imaging: from the Fleischner Society 2017. Radiology. 2017;284:228–43.
26. Bankier AA, MacMahon H, Goo JM, Rubin GD, Schaefer-Prokop CM, Naidich DP. Recommendations for measuring pulmonary nodules at CT: a statement from the Fleischner Society. Radiology. 2017;285:584–600.
27. Deveraj A, van Ginneken B, Nair A, Baldwin D. Use of volumetry for lung nodule management: theory and practice. Radiology. 2017;284:630–44.
28. Lo SCB, Freeman MT, Gillis LB, White CS, Mun SK. Computer-aided detection of lung nodules on CT with a computerized pulmonary vessel suppressed function. Am J Roentgenol. 2018;210:1–9.
29. Van Riel SJ, Ciompi F, Winkler Wille MM, Dirksen A, et al. Malignancy risk estimation of pulmonary nodules in screening CTs: comparison between a computer model and human observers. PLoS One. 2017;12:e0185032.
30. Wille MM, Dirksen A, Ashraf H, et al. Results of the randomized Danish lung cancer screening trial with focus on high-risk profiling. Am J Respir Crit Care Med. 2013;193:542–51.
31. McWilliams A, Tammemagi MC, Mayo JR, et al. Probability of cancer in pulmonary nodules detected on first screening CT. N Engl J Med. 2013;369:910–9.
32. Snoeckx A, Reyntiens P, Desbuquoit D, et al. Evaluation of the solitary pulmonary nodule: size matters, but do not ignore the power of morphology. Insights Imaging. 2018;9:73–86.
33. https://www.kaggle.com/c/data-science-bowl-2017. Accessed 16 July 2018.
34. Lee G, Lee HY, Park H, et al. Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: state of the art. Eur J Radiol. 2017;86:297–307.
35. Aerts HJWL, Velasquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
36. Ko JP, Suh J, Ibidapo O, et al. Lung adenocarcinoma: correlation of quantitative CT findings with pathologic findings. Radiology. 2016;280:931–9.
37. Huang P, Park S, Yan R, et al. Lung cancer diagnosis with small pulmonary nodules: a matched case-control study. Radiology. 2018;286:286–95.
38. Liu Y, Wang H, Li Q, et al. Radiologic features of small pulmonary nodules and lung cancer risk in the National Lung Screening Trial: a nested case-control study. Radiology. 2018;286:298–306.
39. Horeweg N, van Rosmalen J, Heuvelmans MA, et al. Lung cancer probability in patients with CT-detected pulmonary nodules: a prespecified analysis of data from the NELSON trial of low-dose CT screening. Lancet Oncol. 2014;15:1332–41.
40. Konstantinides SV, Torbicki A, Agnelli G, et al. 2014 ESC guidelines on the diagnosis and management of acute pulmonary embolism. Eur Heart J. 2014;35:3033–80.
41. Hochhegger B, Alves GRT, Chaves M, et al. Interobserver agreement between radiologists and radiology residents and emergency physicians in the detection of PE using CTPA. Clin Imaging. 2014;38:445–7.
42. Schoepf UJ, Schneider AC, Das M, Wood SA, Cheema JI, Costello P. Pulmonary embolism: computer-aided detection at multidetector row spiral computed tomography. J Thorac Imaging. 2007;22:319–23.
43. Maizlin ZV, Vos PM, Godoy MB, Cooperberg PL. Computer-aided detection of pulmonary embolism on CT angiography. J Thorac Imaging. 2007;22:324–9.
44. Buhmann S, Herzog P, Liang J, et al. Clinical evaluation of a computer-aided diagnosis (CAD) prototype for the detection of pulmonary embolism. Acad Radiol. 2007;14:851–8.

45. Engelke C, Schmidt S, Bakai A, Auer F, Marten K. Computer-assisted detection of pulmonary embolism: performance evaluation in consensus with experienced and inexperienced chest radiologists. Eur Radiol. 2008;18:298–307.
46. Blackmon KN, Florin C, Bogoni L, et al. Computer-aided detection of pulmonary embolism at CT pulmonary angiography: can it improve performance of inexperienced readers? Eur Radiol. 2011;21:1214–23.
47. Wittenberg R, Berger FH, Peters JH, et al. Acute pulmonary embolism: effect of a computer-assisted detection prototype on diagnosis – an observer study. Radiology. 2012;262:305–13.
48. Das M, Mühlenbruch G, Helm A, et al. Computer-aided detection of pulmonary embolism: influence on radiologists' detection performance with respect to vessel segments. Eur Radiol. 2008;18:1350–5.
49. Kligerman S, Lahiji K, Galvin JR, Stokum C, White CS. Missed pulmonary embolism on CT angiography: assessment with pulmonary embolism – computer aided detection. Am J Roentegenol. 2013;202:65–73.
50. Dewailly M, Remy-Jardin M, Duhamel A, et al. Computer-aided detection of acute pulmonary embolism with 64-slice multi-detector row computed tomography: impact of the scanning conditions and overall image quality in the detection of peripheral clots. J Comput Assist Tomogr. 2010;34:23–30.
51. Lee CW, Seo JB, Song JW, et al. Evaluation of computer-aided detection and dual energy software in detection of peripheral pulmonary embolism on dual-energy pulmonary CT angiography. Eur Radiol. 2011;21:54–62.
52. Lahiji K, Kligerman S, Jeudy J, White C. Improved accuracy of pulmonary embolism computer-aided detection using iterative reconstruction compared with filtered back projection. Am J Roentgenol. 2014;203:763–71.
53. Kröger JR, Hickethier T, Pahn G, Gerhardt F, Maintz D, Bunck AC. Influence of spectral detector CT based monoenergetic images on the computer-aided detection of pulmonary artery embolism. Eur J Radiol. 2017;95:242–8.
54. Aviram G, Soikher E, Bendet A, et al. Prediction of mortality in pulmonary embolism based on left atrial volume measured on CT pulmonary angiography. Chest. 2016;149:667–75.
55. Li Y, Dai Y, Deng L, Guo Y. Computer-aided detection for the automated evaluation of pulmonary embolism. Technol Health Care. 2017;2015:S135–42.
56. Rucco M, Sousa-Rodrigues D, Merelli E, et al. Neural hypernetwork approach for pulmonary embolism diagnosis. BMC Res Notes. 2015;8:617.
57. Uppaluri R, Mitsa T, Sonka M, Hoffman EA, McLennan G. Quantification of pulmonary emphysema from lung computed tomography images. Am J Respir Crit Care Med. 1997;156:248–54.
58. National Emphysema Treatment Trial Research Group. A randomized trial comparing lung volume reduction surgery with medical therapy for severe emphysema. N Engl J Med. 2003;348:2059–73.
59. Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDGene) study design. COPD. 2010;7:32–43.
60. Wan ES, Hokanseon JE, Murphy JR, et al. Clinical and radiographic predictors of GOLD-unclassified smokers in the COPDGene study. Am J Respir Crit Care Med. 2011;184:57–63.
61. Couper D, LaVange LM, Han M, et al. Design of the subpopulations and intermediate outcomes in COPD study (SPIROMICS). Thorax. 2014;69:491–4.
62. Hoffman EA, Ahmed FS, Baumhauer H, et al. Variation in the percent of emphysema-like lung in a healthy, non-smoking multi-ethnic sample. The MESA lung study. Ann Am Thorac Soc. 2014;11:898–907.
63. Busacker A, Newell JD Jr, Keefe T, et al. A multivariate analysis of risk factors for the air-trapping asthmatic phenotype as measured by quantitative CT analysis. Chest. 2009;135:48–56.
64. Witt CA, Sheshadri A, Carlstrom L, et al. Longitudinal changes in airway remodelling and air trapping in severe asthma. Acad Radiol. 2014;21:986–93.
65. Martin SP, Gariani J, Hachulla AL, et al. Impact of iterative reconstructions on objective and subjective emphysema assessment with computed tomography: a prospective study. Eur Radiol. 2017;27:2950–6.
66. Rodriguez A, Ranallo FN, Judy PF, Fain SB. The effects of iterative reconstruction and kernel selection on quantitative computed tomography measures of lung density. Med Phys. 2017;44:2267–80.
67. Galban CJ, Han MK, Boes JL, et al. Computed tomography-based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression. Nat Med. 2012;18:1711–5.
68. Kim HG, Tashkin DP, Clements PJ, et al. A computer-aided diagnosis system for quantitative scoring of extent of lung fibrosis in scleroderma patients. Clin Exp Rheumatol. 2010;28(Suppl 62):S26–35.
69. Xu Y, van Beek EJ, Hwanjo Y, Guo J, McLennan G, Hoffman EA. Computer-aided classification of interstitial lung diseases via MDCT: 3D adaptive multiple feature method (3D AMFM). Acad Radiol. 2006;13:969–78.
70. Martinez FJ, de Andrade JA, Anstrom KJ, King TE Jr, Raghu G, Idiopathic Pulmonary Fibrosis Clinical Research Network. Randomized trial of acetylcysteine in idiopathic pulmonary fibrosis. N Engl J Med. 2014;370:2093–101.
71. Salisbury ML, Lynch DA, van Beek EJR, et al. Idiopathic pulmonary fibrosis: the association between the adaptive multiple features method and fibrosis outcome. Am J Respir Crit Care Med. 2017;195:921–9.

72. Humphries SM, Yagihashi K, Huckleberry J, et al. Idiopathic pulmonary fibrosis: data-driven textural analysis of extent of fibrosis at baseline and 15-month follow-up. Radiology. 2017;285:270–8.
73. Jacob J, Bartholmai BJ, Rajagopalan S, et al. Automated quantitative computed tomography versus visual computed tomography scoring in idiopathic pulmonary fibrosis: validation against pulmonary function. J Thorac Imaging. 2016;31:304–11.
74. Jacob J, Bartholmai BJ, Rajagopalan S, et al. Serial automated quantitative CT analysis in idiopathic pulmonary fibrosis: functional correlations and comparison with changes in visual CT scores. Eur Radiol. 2018;28:1318–27.
75. Wu X, Kim GH, Barber D, et al. Computed tomographic biomarkers in idiopathic pulmonary fibrosis: the future of quantitative analysis. Am J Respir Crit Care Med. 2018; https://doi.org/10.1164/rccm.201803-0444PP. [Epub ahead of print]

Cardiovascular Diseases

13

Johan Verjans, Wouter B. Veldhuis, Gustavo Carneiro,
Jelmer M. Wolterink, Ivana Išgum, and Tim Leiner

Key Points

- Machine learning and deep learning will have a big impact on the diagnosis and workup of cardiovascular diseases.
- The entire imaging chain including patient scheduling, image acquisition, image reconstruction, image interpretation, classification of findings and derivation of prognostic information will be impacted by advances in machine learning and deep learning.
- Machine learning and deep learning are already being applied to echocardiography, CT and MRI of the heart as well as nuclear myocardial perfusion scintigraphy.
- Machine learning and deep learning are ideally suited to reveal prognostic information in diagnostic imaging studies.
- Machine learning and deep learning will enable truly personalized medicine by combining information from multiple sources in addition to the results from imaging studies.
- Bringing machine learning algorithms to the clinic is not straightforward and is ideally done using a vendor-neutral AI platform.

13.1 Introduction

Over the past few decades, different diagnostic techniques have been used for the early detection, diagnosis, monitoring and treatment of cardiovascular diseases. Medical imaging has become an especially indispensable tool for this purpose. The application of artificial intelligence in this field has shown great promise, but its penetration into daily cardiovascular imaging practice has thus far been limited. As a result, the typical daily work of the cardiovascular imaging specialist has continued to be largely dominated by human skill for image acquisition, quantification and interpretation with limited use of computer-aided diagnosis.

Computer vision and artificial intelligence (AI) researchers aim to create intelligent methods to see and comprehend an image as well as

J. Verjans
Department of Cardiology, South Australian Health and Medical Research Institute, University of Adelaide, Adelaide, SA, Australia

Department of Cardiology, Utrecht University Medical Center, Utrecht, The Netherlands

W. B. Veldhuis · T. Leiner (✉)
Department of Radiology, Utrecht University Medical Center, Utrecht, The Netherlands
e-mail: T.Leiner@umcutrecht.nl

G. Carneiro
Department of Cardiology, South Australian Health and Medical Research Institute, University of Adelaide, Adelaide, SA, Australia

J. M. Wolterink · I. Išgum
Image Sciences Institute, Utrecht University Medical Center, Utrecht, The Netherlands

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_13

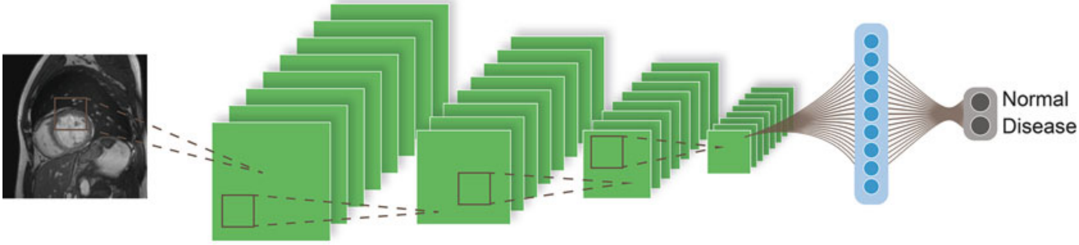


Fig. 13.1 A particularly powerful set of ML tools in medical imaging are artificial neural networks. These networks are modelled on the visual system of the brain and typically contain a large number of interconnected processing elements or 'neurons'. Typically, these elements are structured in multiple layers which are capable of extracting features from data at increasing levels of abstraction. Deep learning refers to artificial convolutional neural networks (CNNs) with multiple (also known as 'hidden') layers between the input and output layers. Over the past years, it has become clear that DL is very well suited for many relevant tasks in radiology such as detecting abnormalities in images, delineating anatomical structures or classifying findings

humans can—or even better. Although computer vision has been applied to analysing medical data for many years [1], until recently this was always based on carefully handcrafted algorithms that performed limited, well-described tasks in imitation of the human process for analysing images. This was primarily due to three reasons: (1) the lack of adequate and efficient methods to train algorithms to perform tasks accurately and under a variety of clinical conditions, (2) the lack of sufficiently large high-quality digital datasets to train an automated system to develop its own approach and (3) the lack of affordable hardware and standardized, open-source software to develop algorithms. Over the past few years, all of these limitations have been overcome, and large-scale medical image acquisition and analysis have enabled high-throughput extraction of imaging features to quantify the changes in end-organ tissue structure and function associated with manifestations of cardiovascular diseases.

This chapter covers practical use cases of the young but rapidly developing field of medical AI and its most promising avenues in the field of cardiovascular imaging. We describe recent advances in machine learning, especially with regard to deep learning, which are helping to identify, classify and quantify cardiovascular disease from echocardiography, CT, MRI and nuclear medicine (Fig. 13.1). In addition, we discuss other potential applications of AI beyond image interpretation.

13.2 Impact of AI on Cardiovascular Imaging

Although most ongoing AI research in the cardiovascular field has focused on image interpretation and prognosis, it is important to realize that AI can and will impact the *entire* imaging chain from choosing a particular imaging test, patient scheduling, image acquisition, image reconstruction and image interpretation to derivation of prognostic information (Fig. 13.2).

13.2.1 Decision Support

Cardiovascular medicine is increasingly guideline driven. Many societies have issued guidelines with the aim to provide standardized and evidence-based care to patients with suspected or known cardiovascular disease. Using this information in clinical practice can be a daunting task. For example, the European Society of Cardiology currently lists 49 categories of guidelines [2]. It is self-evident that no single person can master the intricacies of all these guidelines on a day-to-day basis. It is expected that ML-based decision support systems can help the imaging specialist select the best imaging tests in individual patients.

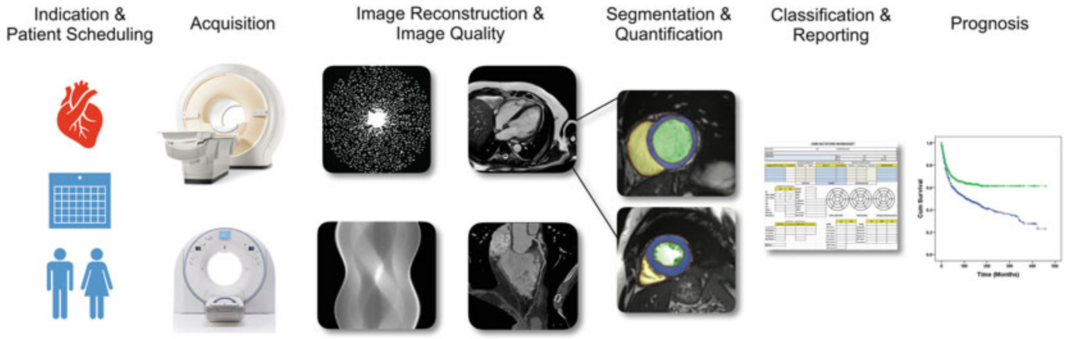

Indication & Patient Scheduling
Acquisition
Image Reconstruction & Image Quality
Segmentation & Quantification
Classification & Reporting
Prognosis
Cum Survival
Time (Months)

Fig. 13.2 Artificial intelligence will impact the entire imaging chain from scheduling to prognosis

13.2.2 Image Acquisition

Hardware vendors are now selling the first AI-based commercial products that help the radiographer during the examination to select the optimal imaging protocol in individual patients, including selection of the precise location and extent of image acquisition. Automated identification of the heart and prescription of scan planes are now possible. This may be especially advantageous for less experienced operators, both for follow-up imaging and complex cases.

13.2.3 Image Reconstruction and Improvement of Image Quality

Machine learning has shown great promise in CT and MR image reconstruction and to a lesser extent in echocardiography and nuclear imaging. For example, deep learning with convolutional neural networks (CNNs) has been successfully applied for very fast reconstruction of highly accelerated cardiac MR acquisitions as an alternative to much slower current state-of-the-art methods such as compressed sensing [3]. Conversely, DL has been applied in CT image reconstruction as well. Several research groups have shown that high-quality CT images can be reconstructed from undersampled projection data [4] or low-radiation-dose, noisy images [5–9]. In addition, deep learning techniques have been used to create CT images with improved spatial resolution on the basis of lower spatial resolution information [10].

13.2.4 Post-processing and Image Analysis

One of the most obvious applications of ML and DL in cardiovascular imaging is image post-processing and analysis. One of the most important and also labour-intensive tasks in cardiac imaging is contouring of the left and right ventricles at end-systole and end-diastole in order to obtain cardiac ejection fractions and myocardial mass. Many research groups have now shown that this process can be fully automated with highly reliable results using ML algorithms [11, 12], and several commercial software packages have already incorporated this technology. Another application in cardiovascular imaging that comes to mind is automated determination of aortic volumes to determine the degree of expansion over the cardiac cycle or to assess volume and rate of growth of aortic aneurysms.

13.2.5 Interpretation and Diagnosis

In addition to using images labelled with a ground truth in terms of segmentation or diagnosis, researchers are now attempting to generate complete radiology reports from images only [13]. DL algorithms are being trained to do this by showing them large datasets of hundreds of thousands to millions of combinations of imaging and the corresponding radiology reports. Although this has not been attempted specifically in cardiovascular radiology, it is expected that this will happen in the future. However, in many patients, information obtained from cardiac imaging tests is just one part of the full clinical picture. This underscores the need for solutions that take into account other information than just images.

13.2.6 Opportunistic Screening and Prognosis

One of the most promising applications of AI and ML in cardiovascular imaging is fully automated detection, quantification and reporting of relevant prognostic information. This may be more detailed information relevant to the clinical problem for which imaging was requested, but also information visible in the images outside of the organ of interest. For instance, more detailed analysis of cardiac motion patterns in patients with pulmonary hypertension has been shown to have a better predictive value for adverse outcomes compared to right ventricular ejection fraction, which is currently used for this purpose [14]. Another example is fully automated identification of vascular calcifications in lung cancer screening CT scans [15–17].

13.2.7 Combining Imaging with Other Data Sources

A final area where AI can be of high value in the future is by combining the results from imaging examinations with other data such as information in electronic health records, laboratory data, genetic analyses and medication use. Combining these data will, for instance, yield new insights into which combinations of clinical and genetic variables are associated with certain imaging findings, outcomes or effectiveness and side effects of new cardiovascular drugs [18, 19].

13.3 Practical Use of AI in Different Cardiovascular Imaging Modalities

13.3.1 Echocardiography

Echocardiography is the most widely used imaging modality in cardiology and is an indispensable tool in the clinical cardiovascular imaging toolbox. Portability and affordability are key advantages of echocardiography, but there are also limitations, such as operator dependency and the complete workflow from acquisition to reporting being a lengthy process [20, 21]. Machine learning strategies could aid echocardiography by not only speeding up reporting time but also could improve accuracy and reduce variability with an ultimate goal to develop a real-time digital 'assistant' that automatically interprets echo images.

Automation is not a novel concept in echocardiography, and disease classification was already attempted back in the 1970s, using Fourier analysis of M-mode images to classify normal subjects, and patients with 'idiopathic hypertrophic subaortic stenosis', mitral valve prolapse and mitral stenosis [22–24, 76]. Using a classification scheme, the investigators demonstrated the potential of an automated detection system of the diagnoses. Recent advances in computer-aided echocardiogram segmentation and diagnosis have shown to improve feasibility, accuracy and reproducibility of real-time full-volume 3D transthoracic echocardiography to measure volumes and function fully automatically, both in sinus rhythm and atrial fibrillation. This was followed by more reports demonstrating rapid and easy automatic segmentation of various left ventricular, right ventricular and atrial parameters including strain, supported by different vendors [25–29]. In comparison with expert contours, automation generation of contours generally accounted for larger ventricular and atrial volumes but resulted in better agreement with the reference standard MRI, which is encouraging. In addition, several deep learning approaches have been described and proven successful, for example, Carneiro et al. achieved similar segmentation performance using only 20 training images. When increased to a training set of 400 images, the approach produced generally more accurate LV segmentation [29]. Another important and perhaps the first step towards full automation of the entire echocardiography imaging chain including automated classification of disease is automatic detection of the standard echocardiography views. Madani et al. anonymized more than 800,000 transthoracic echocardiogram (TTE) images from 267 patients and built a multilayer convolutional neural network to do supervised learning using 15 standard echocardiographic views [30]. The vendor-agnostic model recognized and classified 15 standard views from echocardiography movie clips with 98% overall test accuracy, without significant overfitting. Even on single low-resolution images, test accuracy among 15 standard views was 91.7% versus 70.2–83.5% for board-certified echocardiographers. Elegantly, they used occlusion testing and saliency mapping to trace that the algorithm's classification depended on similar features as used by human experts (Fig. 13.3).

One of the first attempts to validate a clinically useful algorithm for classification of disease was described by Narula et al. [31], who obtained speckle-tracking data from 77 athletes and 62 hypertrophic cardiomyopathy patients and used 3 different algorithms (support vector machines, random forests and artificial neural networks) to investigate the ability to distinguish physiological from pathological cardiac remodelling. The machine learning model showed increased sensitivity and specificity compared with average early diastolic tissue velocity, early-to-late diastolic mitral velocity ratio and strain ($p < 0.01$; $p < 0.01$; $p = 0.04$). The same group of investigators studied 94 patients with either constrictive pericarditis (CMP) or restrictive cardiomyopathy (RCM) to improve accuracy in this challenging echocardiographic problem [32]. In both studies, speckle-tracking echocardiography datasets were used to distinguish restrictive from constrictive pericarditis, which was achieved with an AUC of 89% without and 96% with selected echocardiographic variables, while traditional imaging biomarkers achieved 63–82%. This study demonstrates feasibility of a cognitive machine learning approach for learning and recalling patterns observed during echocardiographic evaluations.

Current guidelines recommend quantitative and also semiquantitative echocardiographic techniques to assess the severity of valvular disease which is essential for therapeutic and perioperative management [33]. Since geometrical assessment of valves can be based on pattern recognition, this application is likely to be enhanced and automated using machine learning. Many investigations that aim to automate valve assessment have focused on the mitral valve [34]. Jin et al. used so-called anatomically intelligent ultrasound (AIUS; Philips Medical Systems, Andover, MA, USA) to semiautomatically track the annulus and leaflet anatomy to demonstrate that AIUS required significantly less time for image analysis (1.9 ± 0.7 min vs. 9.9 ± 3.5 min, $p < 0.0001$), improved sensitivity (60% vs. 90%, $p < 0.001$), specificity (91% vs. 97%, $p = 0.001$) and accuracy (83% vs. 95%, $p < 0.001$) [35]. This example illustrates that semiautomated algorithms can improve performance in localizing mitral valve disease, which can support less experienced operators. In aortic disease, Calleja et al. tested in patients with either aortic stenosis, dilated aorta or aortic regurgitation an automated 3D algorithm (Aortic Valve Navigator; Philips Medical Systems, Andover, MA, USA) to model and quantify the aortic root, using both 3D transoesophageal echocardiography and CT data to assist in planning transcatheter aortic valve replacement [36]. They demonstrated excellent reproducibility in quantification of aortic regurgitation and stenosis. Valve annulus diameters were significantly underestimated ($p < 0.05$) by 2D echo images compared to 3D echo and CT. Besides being shorter in duration, the analysis was accurate compared to CT, which was used as the standard of reference. Although most commercial software packages still require significant user input, packages for valvular assessment will soon require minimal to no user input, incorporating artificial intelligence to achieve their goal [37]. However, there is still

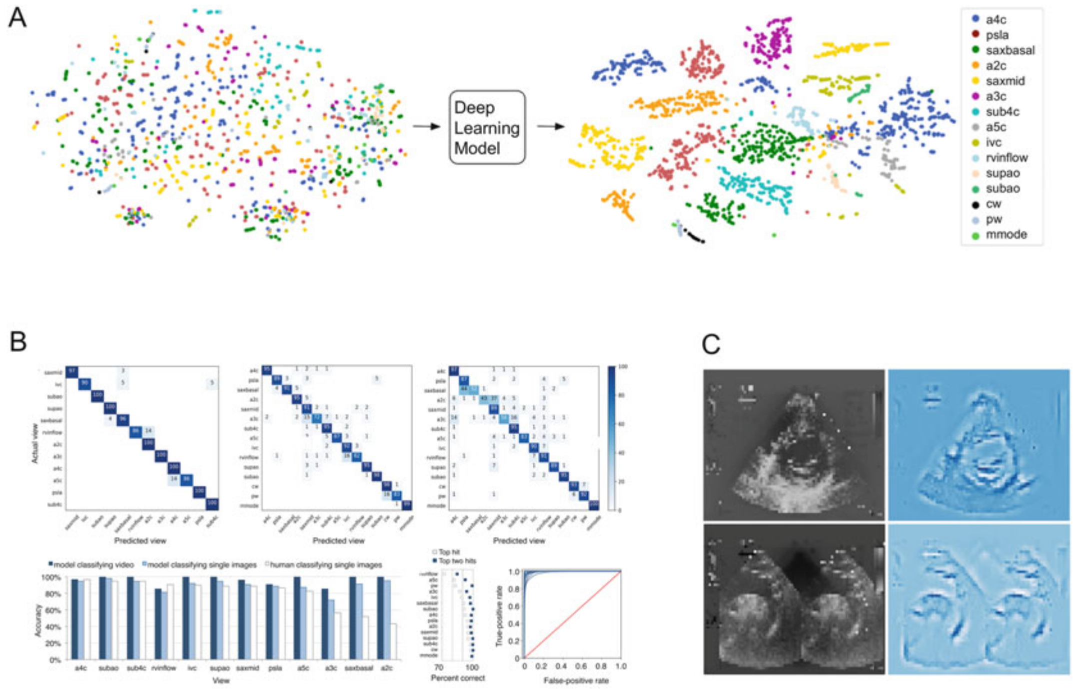


Fig. 13.3 Deep learning model demonstrating human performance in distinguishing 15 standard echocardiographic views using t-SNE clustering analysis of image classification. (**a**) On the left the input echocardiogram images, plotted in 4800-dimensional space projected to 2-dimensional space for visualization. Different colours represent different classes. After neural network analysis, there appears organization into clusters. (**b**) Confusion matrix with real view labels on *y*-axis and neural network-predicted view labels on the *x*-axis by view category for video classification (left) and still-image classification (middle) compared with a representative board-certified echocardiographer (right). The numbers in the boxes indicate percentage of labels predicted for true category. Bottom left graph compares comparison of deep. Learning accuracy for video classification (dark blue), still-image classification (light blue) and still-image classification by a representative echocardiographer (white). Lower right graph indicates receiver operating characteristic curves for view categories ranging from 0.985 to 1.00 (mean 0.996). (**c**) Saliency maps (occlusion map not shown). The input pixels weighted most heavily in the neural network's classification of the original images (left). The most important pixels (right) make an outline of relevant structures demonstrating similar patterns that humans use to classify the image. Abbreviations: *a4c* apical four-chamber, *psla* parasternal long axis, *saxbasal* short axis basal, *a2c* apical two-chamber, *saxmid* short axis mid/mitral, *a3c* apical three-chamber, *sub4c* subcostal four-chamber, *a5c* apical five-chamber, *ivc* subcostal ivc, *rvinflow* right ventricular inflow, *supao* suprasternal aorta/aortic arch, *subao* subcostal/abdominal aorta, *cw* continuous-wave Doppler, *pw* pulsed-wave Doppler, *mmode* m-mode. Adapted with permission of the authors from reference [30]

extensive work to be done to validate feasibility and reproducibility of these software packages in clinical routine.

13.3.2 Computed Tomography

Cardiac computed tomography is a widely used method to assess the presence and extent of coronary atherosclerosis and to evaluate cardiac anatomy. Coronary artery calcium scoring (CACS) and coronary CT angiography (CCTA) are highly sensitive techniques to rule out coronary artery disease. Wolterink et al. have demonstrated in a phantom as well as patient study that radiation dose for CACS can be reduced by up to 80% by training a convolutional neural network (CNN) jointly with an adversarial CNN to estimate routine-dose CT images from low-dose CT images and hence reduce noise (Fig. 13.4). Noise reduction improved quantification

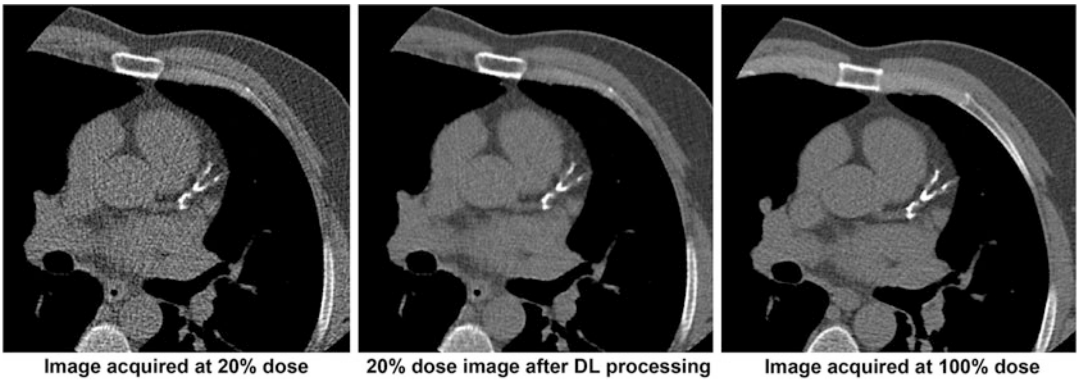


Image acquired at 20% dose
20% dose image after DL processing
Image acquired at 100% dose

Fig. 13.4 Deep learning methods can be used to transform a CT image acquired at extremely low-radiation dose (20% of routine dose; left panel) into a high-quality image resembling an image of the same patient acquired at routine radiation dose (middle panel). On the right the corresponding image slice from the full-dose acquisition is shown

of low-density calcified inserts in phantom CT images and allowed coronary calcium scoring in low-dose patient CT images with high noise levels [5].

Machine learning has also made inroads in quantification of coronary calcium and atherosclerotic plaque. Coronary artery plaque burden, expressed as either the total amount of coronary calcium in non-contrast enhanced CT scans [38] or the volume of calcified, non-calcified and mixed atherosclerotic plaque, is associated with chest pain symptoms [39] and highly indicative of future cardiovascular events [40] the ultimate goal is fully automated quantification of both. Detailed manual analysis of the coronary vasculature to determine atherosclerotic plaque burden can be tedious and time-consuming and is therefore impractical in current clinical practice. For this reason, coronary evaluation is primarily done by assigning ordinal scores to coronary segments [41]. To solve this problem, Wolterink et al. [42, 43] as well as several other groups of investigators [44] have described a method capable of fully automated quantification of coronary artery calcium (Fig. 13.5). The method uses supervised learning to directly identify and quantify CAC without a need for manual annotations as commonly used in existing methods. The study included cardiac CT exams of 250 patients, and agreement with the reference mass score was excellent with an intraclass correlation coefficient of 0.94. In further work, the same investigators have described a method to determine the presence and classify the type of coronary artery plaque, as well as to determine the presence and the degree of a coronary stenosis [45]. Based on manually annotated CCTA from 131 patients with suspected or known coronary artery disease, a recurrent CNN was trained to perform fully automated analysis of coronary artery plaque and stenosis with high accuracy and reliability compared to the manual segmentations.

Once coronary plaques have been identified, it can be desirable to use machine learning for a more detailed evaluation using radiomic techniques. Radiomics is a process typically referring to supervised ML that consists of extracting a large number of quantitative features from radiology images and subsequent classification using a ML classifier to determine diagnosis or perform prediction. Thus, radiomics, also sometimes called texture analysis (TA), objectively quantifies texture of radiological images by exploiting interpixel relationships. Kolossvary et al. used this approach to improve identification of high-risk ‘napkin-ring sign’ coronary plaques and found that radiomic analysis improved identification of these plaques over conventional quantitative parameters [46]. Mannil et al. have used radiomics to enable

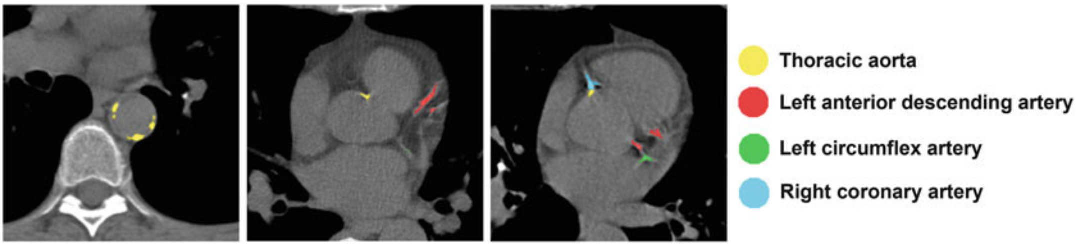


Fig. 13.5 Fully automatic identification of calcifications in the thoracic aorta and coronary arteries in chest CT using deep learning without human interaction. Algorithms such as the one shown may not only facilitate faster and more robust identification of arterial calcium deposits in patients referred for evaluation of coronary artery disease but also in patients referred for other clinical questions in which the heart is depicted in the field of view

the differentiation between patients with prior myocardial infarction and control subjects using non-contrast-enhanced low-radiation CACS images. They studied 30 control subjects and 57 patients with acute or chronic MI and found that TA in combination with deep learning was able to identify patients with prior MI with an area under the receiver operating curve of 0.78. This study was the first to demonstrate the ability of TA to unmask acute or chronic MI on non-contrast-enhanced low-radiation dose CACS images with high accuracy [47].

Another area where machine learning has been of value is for identification of hemodynamically significant coronary stenoses, since the specificity of the CCTA for this purpose is low when using visual analysis [48]. Zreik et al. published a method for automatic identification of patients with functionally significant coronary artery stenoses, employing deep learning analysis of the left ventricle (LV) myocardium in resting CCTA scans. They studied CCTA scans of 166 patients who underwent invasive fractional flow reserve (FFR) measurements [49]. To identify patients with a functionally significant coronary artery stenosis, analysis was performed in several stages. First, the LV myocardium was segmented using a multiscale convolutional neural network (CNN). To characterize the segmented LV myocardium, it was subsequently encoded using an unsupervised convolutional autoencoder (CAE). Subsequently, LV myocardium was divided into a number of spatially connected clusters, and statistics of the encodings were computed as features. Thereafter, patients were classified according to the presence of functionally significant stenosis using a support vector machine (SVM) classifier based on the extracted features. Images of 20 patients with invasive FFR measurements were used to train the LV myocardium encoder. Classification of patients was evaluated in the remaining 126 patients using 50 tenfold cross-validation experiments. The use of the algorithm resulted in an area under the receiver operating characteristic curve (AUC) of 0.74 ± 0.02. At sensitivity levels 0.60, 0.70 and 0.80, the corresponding specificity was 0.77, 0.71 and 0.59, respectively. The results demonstrate that automatic analysis of the LV myocardium in a single CCTA scan acquired at rest, without assessment of the anatomy of the coronary arteries, can be used for fully automated identification of patients with functionally significant coronary artery stenosis. Coenen et al. took a different approach and focused on improving an existing computational fluid dynamics (CFD) method for identification of significant CAD [50]. The CFD method is a workstation-based algorithm which relies on manual extraction of the coronary tree, a process that takes approximately 30–60 min. Once the coronary tree is extracted, coronary flow and pressure are simulated both at rest and in a hyperaemic state by virtual reduction of the microvascular resistance, thereby simulating the effect of adenosine infusion. The intracoronary blood pressure in the hyperaemic state is then divided by assumed blood pressure in the aorta to calculate the pressure drop across coronary stenoses. To improve the accuracy of this approach, the authors used a method trained using 12,000 synthetic 3D coronary models of various anatomies and degrees of CAD, for which the CFD-based CT-FFR values were computed. The CFD-based results from the 12,000 synthetic coronary models were then used as the ground truth training data for the ML-based CT-FFR application. Subsequently, the ML-based CT-FFR model was trained using a deep learning model incorporating 28 features extracted from the coronary tree geometry. Application of their method improved diagnostic accuracy on a per-vessel basis from 58% to 78% in a cohort of 351 patients in whom 525 coronary arteries were analysed. At the patient level, diagnostic accuracy improved from 71% to 85% compared to visual analysis of the CCTA images.

Machine learning has also been studied with regard to its ability to improve the prognostic value of CCTA. Motwani et al. investigated the value of ML to predict 5-year all-cause mortality (ACM) over traditional clinical and CCTA metrics in a 10,030 patient substudy of the COronary CT Angiography EvaluatioN For Clinical Outcomes: An InteRnational Multicenter (CONFIRM) registry [51]. Prediction of ACM in these patients was compared between Framingham risk score (FRS), CCTA-derived metrics such as the segment stenosis score (SSS) and segment involvement scores (SIS) and a ML algorithm, based on 44 CCTA parameters and 25 clinical parameters. Compared to the conventional metrics, ML exhibited a substantially and significantly higher AUC compared with FRS or CCTA data alone for prediction of 5-year ACM (ML, 0.79 vs. FRS, 0.61; SSS, 0.64; SIS, 0.64; $p < 0.001$ for all). In this first large-scale evaluation of ML for prognostic risk assessment using CCTA data, the observed efficacy suggests ML has an important clinical role in evaluating prognostic risk in individual patients with suspected CAD.

13.3.3 Magnetic Resonance Imaging

Machine learning is bound to take MRI to the next level on all levels of the imaging continuum [52]. As discussed in the introductory section of this chapter, Schlemper et al. have recently described a novel method based on a cascade of CNNs for dynamic MR image reconstruction that consistently outperformed state-of-the-art methods based on compressed sensing. Cardiac MRI relies on undersampling to achieve clinically feasible scan times, but complex algorithms are needed to de-alias the resulting images. Existing methods rely on the requirements of sparsity and incoherence between the sampling and sparsity domains and are computationally intensive. The presented deep learning method suggests that the CNN was capable of learning a generic strategy to de-alias the images without explicitly formulating rules how to do so [3].

Similar to echocardiography, CMR requires highly skilled radiographers with knowledge of physics, anatomy and pathology to obtain diagnostic images. There are considerable ongoing efforts to automate and accelerate CMR acquisition for non-complex scan protocols without human intervention.

In the journey towards practical applications for imaging specialists, machine learning methods have been mainly applied for segmentation and increasingly also for classification. It must be noted that CMR is inherently variable per patient and per scanner and is subject to many parameters that impose more difficulties in interpretation and require preprocessing. Alignment to a common orientation, bias correction algorithms and normalization are important aspects to consider before feeding data into a machine learning algorithm.

Current diagnostic criteria are largely derived from quantitative measures that indicate the difference between normal and pathologic processes of different degrees. These measures are derived from manually delineated contours from imaging specialists. The main challenges for automating this common practice are variability of heart shape, basal and apical slices, variability among different scanners and overlap between cardiac and background structures with noisy and fuzzy margins [53].

Nevertheless, a multitude of reports suggest that the problem of automatic LV segmentation

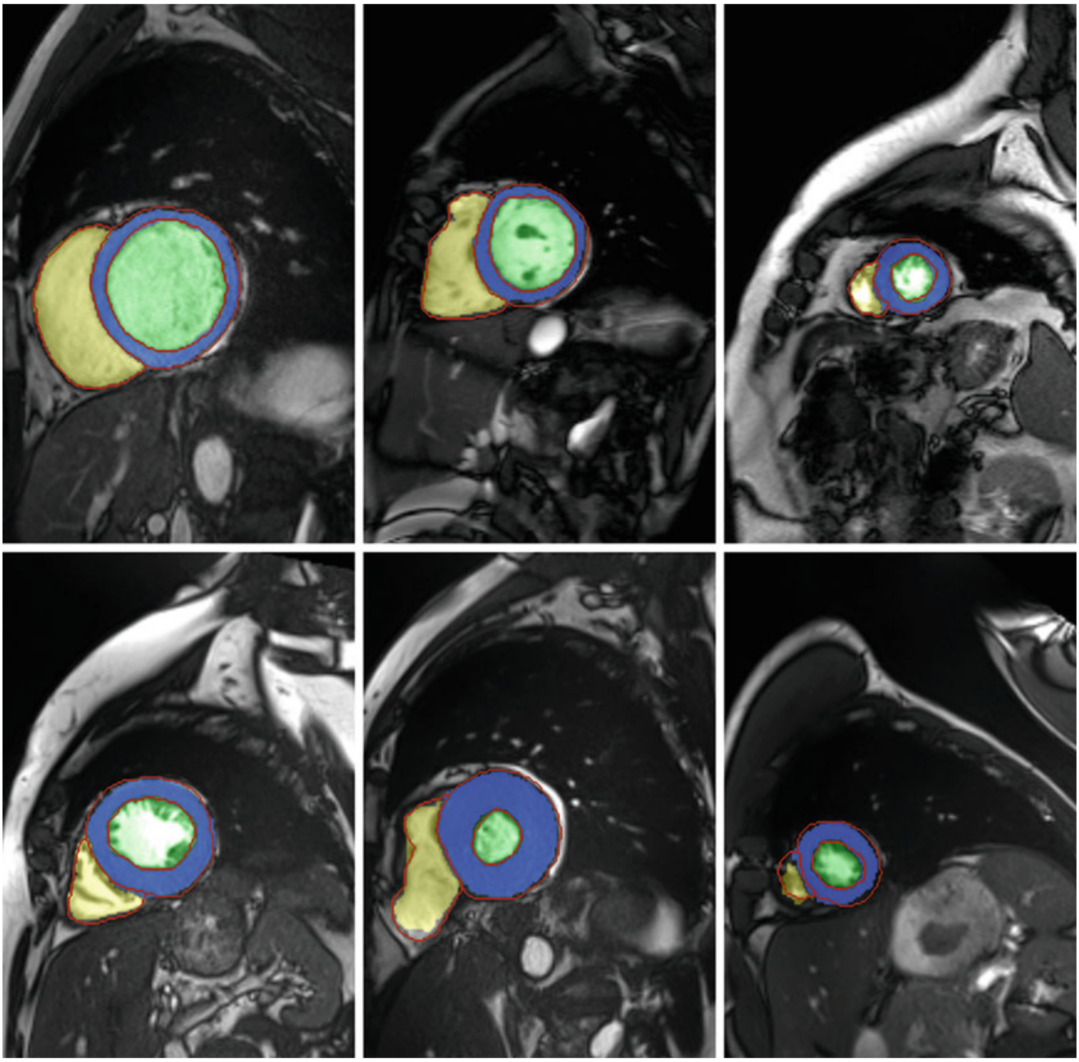


Fig. 13.6 Results of fully automatic segmentation of the left ventricular cavity (green), left ventricular myocardium (blue) and right ventricular cavity (yellow) in cardiac cine MR. Manual reference annotations are shown in red

is bound to be solved by machine learning in the near future (Fig. 13.6). Although there have been attempts to define ground truth in segmentation [54], this remains an issue when training and validation datasets have been analysed by operators with different levels of expertise. In one of the largest studies to date, manually analysed CMR data from 4500 patients from the UK Biobank was compared in terms of segmentation performance to a commercially available automated algorithm (Siemens syngo InlineVF; Siemens Healthcare, Erlangen, Germany) [55]. After excluding grossly misplaced contours in patients, the remaining patients showed good agreement of ESV −6.4 ± 9.0 ml, 0.853 (mean ± SD of the differences, ICC); EDV −3.0 ± 11.6 ml, 0.937; SV 3.4 ± 9.8 ml, 0.855; and EF 3.5 ± 5.1%, 0.586. LV mass was overestimated (29.9 ± 17.0 g, 0.534) due to larger epicardial contours. This study is one of the first to show feasibility of large-scale automated analysis in a relatively healthy population. The more variable RV geometry poses a greater challenge, but commercial software packages have started incorporating machine learning algorithms for automatic LV and also RV segmentation for clinical use, resulting in better workflow. However, the accuracy, and in particular RV segmentation, still has room for improvement.

The ability of ML to aid in outcome prediction based on CMR images was demonstrated in a report by Dawes et al., where semiautomatic segmentation was applied to study the right ventricle in 256 patients with pulmonary hypertension [14]. The investigators applied a supervised strategy based on principal component analysis (PCA) to identify patterns of systolic motion that were most strongly predictive of survival in this cohort with 36% mortality at follow-up. When added to conventional imaging and biomarkers, 3D cardiac motion improved survival prediction with an AUC of 0.73 versus 0.60, respectively ($p < 0.001$), and provided greater differentiation in median survival between high- and low-risk groups (13.8 vs. 10.7 years; $p < 0.001$). This exemplifies how a machine learning survival model can help discover novel independent predictors of outcome from non-standard imaging biomarkers.

Stress perfusion imaging on CMR has entered the stage after studies have convincingly shown to be a great alternative for detecting ischemia compared to other imaging modalities [56, 57]. Furthermore, there have been some exciting studies demonstrating the feasibility of quantitation of flow [58]. The ability to quantitate, in combination with automation, would form an immediate useful application. An example of near-automation was recently shown in a small study using stress perfusion images, demonstrating same diagnostic accuracy of automated compared to manual analysis (AUROC 0.73 vs. 0.72) [59].

An exciting area of CMR that could significantly influence cardiovascular disease classification is the ability to noninvasively quantify blood flow anywhere in the body, also called 4D flow [60, 61]. The large amount of data that is generated per patient with 4D flow imaging is currently still posing challenges for the interpretation by human imaging specialists. Therefore, the field of flow quantification might even benefit most from machine learning algorithms, especially when large datasets become available. Flow quantification in heart failure, for example, has significant potential as it would measure the resultant of flow, diastolic and systolic function [61].

An initial report has been published on automated detection of late gadolinium enhancement (LGE), a marker of fibrosis [62]. Standardized T1 and T2 mapping also assess fibrosis and myocardial edema are also starting to enter clinical routine. Ultimately, automated quantification of LGE, T1, T2, cardiac function and flow could help stratify these hearts according to their risk or phenotype, as shown by Rayatzadeh et al., potentially paving the way for a personalized decision to implant an ICD in a patient [63].

Congenital heart disease is often hampered by limited numbers of patients and heterogeneous populations. As a result, this field could greatly benefit from a computer-assisted approach using limited datasets and also quantitative flow. Samad et al. investigated 153 tetralogy of Fallot patients with the aim to predict which patients would deteriorate over a median time of 2.7 years (scans >6 months apart) [64]. Support vector machine classifiers including cross-validation to identify useful variables. The mean AUC for major or minor versus no deterioration was 0.82 ± 0.06 and for major versus no or minor was 0.77 ± 0.07, demonstrating the utility of baseline variables that were not uncovered using regression analyses. Although it is a relatively small study, it is elegant in that it underscores the potential using less sophisticated algorithms.

13.3.4 Nuclear Imaging

As is the case with CT imaging, nuclear imaging techniques such as single-photon emission computed tomography (SPECT) and positron emission tomography (PET) are dependent on ionizing radiation to obtain insight into cardiac structure and function. It is paramount that the lowest possible radiation dose be used to enable diagnostic evaluation. Several studies have now been published that utilized CNNs to reconstruct diagnostic images out of low-radiation dose, i.e. noisy and blurry, acquired images [65, 66], but the proposed methods have not been widely used on clinical patient data.

Machine learning has been studied, however, with regard to its ability to improve the diagnostic accuracy of SPECT myocardial perfusion imaging (MPI). Arsanjani and colleagues found significant improvements in the diagnostic accuracy for detection of significant CAD when utilizing a ML algorithm which combined multiple quantitative perfusion and clinical variables [67]. The ML algorithm outperformed both automatically calculated total stress perfusion deficit (TPD) and two expert observers in terms of AUC (0.94 vs. 085–0.89; $p < 0.001$) for the detection of obstructive CAD in 1181 patients referred for SPECT MPI. The largest study to date to apply deep learning to improve identification of flow-limiting CAD is by Betancur et al. who developed a CNN trained from obstructive stenosis correlations by invasive coronary angiography to estimate the probability of obstructive coronary artery disease in the main coronary arteries [68]. The CNN computed a probability of obstructive CAD in large epicardial coronary vessels without predefined subdivision of the polar MPI map. This network was trained on both raw data and quantitative polar MPI maps of 1638 patients without known CAD undergoing 99mTc-sestamibi or tetrofosmin MPI. During training, the feature extraction units learn to recognize key polar map features, and the fully connected layers learn how to combine these features to predict per-vessel disease. Multivessel disease prediction was based on the patterns of predicted probabilities for each vessel. Using the CNN the AUC for identification of significant CAD improved from 80% to 82% at patient level and from 64% to 70% at vessel level. Of note, in this study the standard of reference was the degree of coronary stenosis as seen on the invasive coronary angiography and not invasive FFR.

The same group of investigators also studied the ability of ML to predict early revascularization and death in patients with suspected CAD. With regard to the outcome of revascularization, the ML algorithm was able to identify patients likely to undergo percutaneous coronary intervention as opposed to medical therapy alone with an AUC of 0.81 and performed equal to or better on this task compared to two experienced observers (AUC, 0.72–0.81) [69]. The ability of ML to predict the broader outcome of major adverse cardiovascular events (MACE) up to 3 years after undergoing SPECT MPI was studied in 2619 consecutive patients referred for clinically indicated exercise or pharmacological stress MPI [70]. Again, a ML algorithm based on 28 clinical, 17 stress test and 25 imaging variables outperformed physician expert prediction of MACE as well as predictions based on automated determination of stress TPD with an AUC of 0.81 for ML versus 0.65 for the physician expert and 0.73 for TPD. Since MACE can be considered subjective to a certain extent, it is important to assess the ability of ML to improve prediction of cardiac death. Precisely this problem was studied by Haro Alonso et al. in 8321 patients who had undergone dual-isotope SPECT MPI with adenosine stress and adenosine stress with walking at a large US medical centre [71]. In a comprehensive study that included 122 clinical and imaging variables, the ability to predict cardiac death was studied using logistic regression (LR) and several ML models. Follow-up duration was 3.2 ± 2.0 years. All of the ML algorithms outperformed LR in terms of AUC for prediction of cardiac death. Best performance was achieved using a SVM that yielded an AUC of 0.83 versus 0.76 for LR.

13.3.5 Outcome Prediction Based on Composite Data

The added value of ML for cardiovascular event prediction becomes evident when multiple data sources are combined. One of the most significant studies to appear on this topic is the recent work by Ambale-Venkatesh et al. [72] who used random survival forests ML to identify the top 20 predictors of death, all cardiovascular disease, stroke, coronary heart disease, heart failure and atrial fibrillation from over 700 variables measured in 6814 multi-ethnic study of atherosclerosis (MESA) participants, free of cardiovascular disease at baseline. In addition to traditional cardiovascular risk factors, information about medication use and several questionnaires were collected, as well as the ankle-brachial index (ABI), information about coronary calcium, atherosclerotic plaque in the carotid arteries, a long list of magnetic resonance imaging and laboratory and electrocardiographic biomarkers. Variables from imaging markers, ankle-brachial index (ABI) and serum biomarkers were of intermediate to high prediction importance, whereas, apart from age, traditional risk factors, questionnaires and medication exposure were of lower importance. ECG indices related to the ST segment were of intermediate importance, whereas other ECG indices had low-intermediate importance [72]. This work is important because it clearly demonstrates that currently used standard risk scores can be improved substantially by using so-called deep phenotyping (multiple evaluations of different aspects of a specific disease process), which facilitates efficient prediction of specific outcomes. In current clinical practice, many of these measures from imaging, biomarker panels and ECG signals are frequently ignored by many clinicians. Although MESA studied subjects free of cardiovascular disease at baseline, these findings hold promise for clinical practice because many of the collected variables will also be available in patients with suspected and known cardiovascular disease. Thus, ML can take into account the full breadth and depth of phenotypic data and improve our understanding of the determinants of cardiovascular disease.

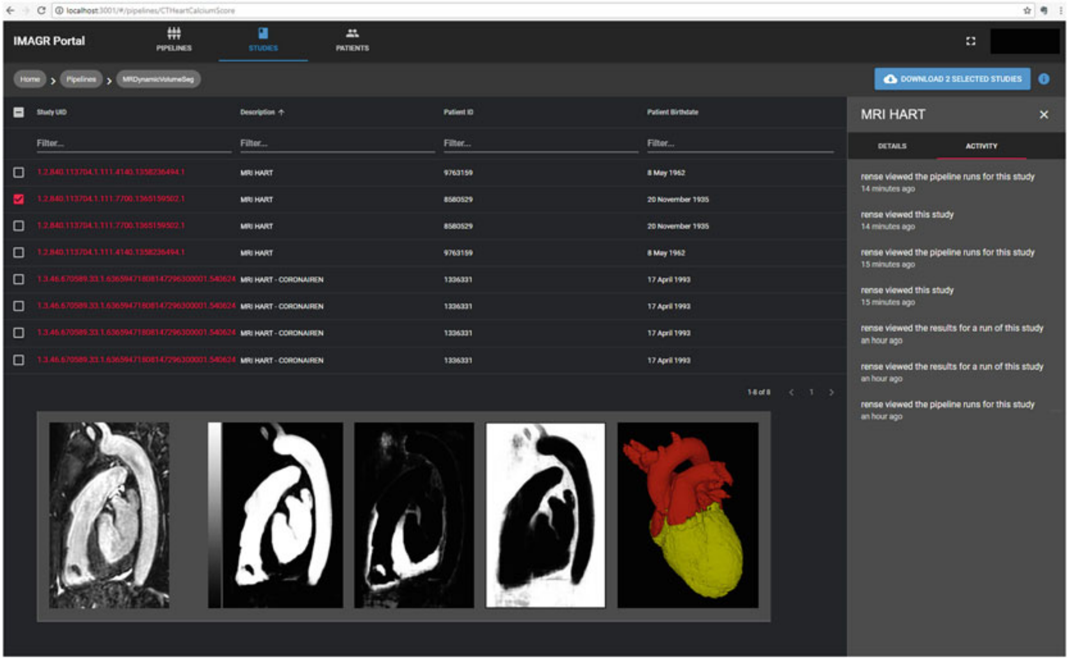
13.3.6 Deployment of Algorithms in Clinical Practice

Despite the promising developments described above, actual, structural implementation of AI in the clinical workflow has, to the best of our knowledge, not been achieved anywhere. One of the main reasons for this is that it is a lot of work, i.e. up to several months per application, to transform a scientifically proven technique into a product that can be used in the clinical setting. This is work that requires a dedicated infrastructure as well as dedicated personnel.

Here we propose a general solution to this problem and show how this solution was implemented at Utrecht University Medical Center. In addition to bridging the gap from the lab to the clinic, the proposed solution brings more general benefits that will be described below. Similar to using a Vendor Neutral Archive (VNA) for storage, we propose to build a Vendor Neutral AI Platform (VNAP) for vendor-neutral clinical implementation of AI.

The VNAP infrastructure takes care of the boilerplate activity needed to place an algorithm in the hands of physicians. Using a standardized build process, an algorithm can be containerized and employed on the VNAP. Containerization has many advantages: it is very straightforward; it decouples deployment from the programming language and tools used to implement the algorithm or trained neural network (NN); and it makes the AI implementation independent of (changes in) the underlying VNAP infrastructure and also independent of other algorithms, and their requirements, simultaneously running on the VNAP. Importantly, it also allows for versioning which is the ability to run a previous version of the algorithm alongside a newer version on the same patient—a crucial ability that will likely be a new standard requirement when using constantly evolving, learning neural networks in a clinical setting. An important more general benefit is that instead of implementing tens or hundreds of separate cloud solutions from different vendors, the IT department only has to take care of one infrastructure and only has to take care of interoperability between one infrastructure and the picture archiving and communication system (PACS) and the hospital information system (HIS) instead of managing hundreds often fragile connections. For end-users, the advantage is that only one way of interacting with AI needs to be learned. Lastly, investments in hardware upgrades are more easily justified because the whole range of AI applications benefits at once.

The above are the specifications of IMAGR, a UMCU-designed VNAP implementation, built on top of a virtual hardware cluster—



	Study UID	Description ↑	Patient ID	Patient Birthdate
	Filter...	Filter...	Filter...	Filter...
□	1.2.840.113704.1.111.4140.1358236494.1	MRI HART	9763159	8 May 1962
☑	1.2.840.113704.1.111.7700.1365159502.1	MRI HART	8580529	20 November 1935
□	1.2.840.113704.1.111.7700.1365159502.1	MRI HART	8580529	20 November 1935
□	1.2.840.113704.1.111.4140.1358236494.1	MRI HART	9763159	8 May 1962
□	1.3.46.670589.33.1.63659471808147296300001.540624	MRI HART - CORONAIREN	1336331	17 April 1993
□	1.3.46.670589.33.1.63659471808147296300001.540624	MRI HART - CORONAIREN	1336331	17 April 1993
□	1.3.46.670589.33.1.63659471808147296300001.540624	MRI HART - CORONAIREN	1336331	17 April 1993
□	1.3.46.670589.33.1.63659471808147296300001.540624	MRI HART - CORONAIREN	1336331	17 April 1993

rense viewed the pipeline runs for this study
14 minutes ago

rense viewed this study
14 minutes ago

rense viewed the pipeline runs for this study
15 minutes ago

rense viewed this study
15 minutes ago

rense viewed the results for a run of this study
an hour ago

rense viewed the results for a run of this study
an hour ago

rense viewed the pipeline runs for this study
an hour ago

Fig. 13.7 Screenshot of vendor-neutral AI platform user interface. Cardiac MR examinations are shown with resulting segmentation below the list of studies. Source image of the cardiac MRI is shown on left lower panel, and voxel-wise tissue classes (blood pool, myocardium and background) after segmentation are shown in the middle three panels. Right lower panel shows resulting three-dimensional rendering

distributing CPU and GPU power—from proven open-source components, like Apache Airflow, Celery RabbitMQ, Redis and Docker, which manage the planning, monitoring and execution of pipelines of tasks and the associated data flow. Via IMAGR, any AI or other image processing application, from any vendor or any PhD student, can now be made available on every PACS station, whether for research purposes or for patient care (Fig. 13.7).

13.3.7 Outlook and Conclusions

Imaging modalities, such as echocardiography, CT, MRI and nuclear imaging techniques, have moved beyond the stage of simply imaging the heart but can now be used to transform the depiction of living biological tissues into quantitative data. Excitingly, recent studies have confirmed that it is possible to achieve significant improvements in prediction accuracy and to predict the previously unpredictable. The proliferation of cardiovascular imaging data, and particularly a set of international projects aiming to accumulate large datasets, has overcome this problem, making machine learning using big imaging data a very promising field. The potential impact of these predictions on patient health and cost of treatment could be immense. Clinicians will need to prepare for a paradigm shift in the next decades, from trusting their eye to potentially trusting a prospectively validated 'black box' that answers their diagnostic questions. Although ML impacts the entire imaging chain, it is our belief that the true impact of AI in the short term will be in risk prediction algorithms that guide clinicians. It is clear that machine learning has vast potential to change the way clinicians work. Encouragingly, the US Food and Drug Administration is now making efforts to stimulate practical use of artificial intelligence in medicine [73]. Nevertheless, because of its potential to change the way we generate knowledge, interpret data and make decisions, artificial intelligence may trigger uncertainties and reservations among healthcare providers and clinicians. It is evident that cardiovascular research is on the front-line in helping make practical clinical applications a reality. Randomized trials of computer versus human or versus computer-assisted humans will be a major step that could make clinical application reality.

In conclusion, cardiac imaging has advanced greatly in the last decade. The availability of large imaging biobanks such as MESA (>2500 scans) [74] and the UK Biobank (up to 100,000 CMR scans) [75] will significantly contribute to advancement of ML in cardiac imaging. However, despite these advances, large-scale clinical adoption of machine learning algorithms will take time [52].

References

1. Kukar M, Kononenko I, Groselj C, Kralj K, Fettich J. Analysing and improving the diagnosis of ischaemic heart disease with machine learning. Artif Intell Med. 1999;16:25–50.
2. https://www.escardio.org/Guidelines/Clinical-Practice-Guidelines. European Society of Cardiology. Accessed July 22 2018.
3. Schlemper J, Caballero J, Hajnal JV, Price AN, Rueckert D. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. IEEE Trans Med Imaging. 2018;37:491–503.
4. Zhang Z, Liang X, Dong X, Xie Y, Cao G. A sparse-view CT reconstruction method based on combination of DenseNet and deconvolution. IEEE Trans Med Imaging. 2018;37:1407–17.
5. Wolterink JM, Leiner T, Viergever MA, Išgum I. Generative adversarial networks for noise reduction in low-dose CT. IEEE Trans Med Imaging. 2017;36:2536–45.
6. Yang Q, Yan P, Zhang Y, Yu H, Shi Y, Mou X, Kalra MK, Zhang Y, Sun L, Wang G. Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. IEEE Trans Med Imaging. 2018;37:1348–57.
7. Kang E, Min J, Ye JC. A deep convolutional neural network using directional wavelets for low-dose X-ray CT reconstruction. Med Phys. 2017;44:e360–75.
8. Chen H, Zhang Y, Zhang W, Liao P, Li K, Zhou J, Wang G. Low-dose CT via convolutional neural network. Biomed Opt Express. 2017;8:679–94.
9. Chen H, Zhang Y, Kalra MK, Lin F, Chen Y, Liao P, Zhou J, Wang G. Low-dose CT with a residual encoder-decoder convolutional neural network. IEEE Trans Med Imaging. 2017;36:2524–35.
10. Park J, Hwang D, Kim KY, Kang SK, Kim YK, Lee JS. Computed tomography super-resolution using deep convolutional neural network. Phys Med Biol. 2018;63:145011.
11. https://www.kaggle.com/c/second-annual-data-science-bowl. Accessed July 22 2018.
12. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng PA, Cetin I, Lekadir K, Camara O, Ballester MAG, Sanroma G, Napel S, Petersen S, Tziritas G, Grinias E, Khened M, Kollerathu VA, Krishnamurthi G, Rohe MM, Pennec X, Sermesant M, Isensee F, Jager P, Maier-Hein KH, Baumgartner CF, Koch LM, Wolterink JM, Išgum I, Jang Y, Hong Y, Patravali J, Jain S, Humbert O, Jodoin PM. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans Med Imaging. 2018; https://doi.org/10.1109/TMI.2018.2837502. [Epub ahead of print]
13. Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports. https://arxiv.org/abs/1711.08195
14. Dawes TJW, de Marvao A, Shi W, Fletcher T, Watson GMJ, Wharton J, Rhodes CJ, Howard LSGE, Gibbs JSR, Rueckert D, Cook SA, Wilkins MR, O'Regan DP. Machine learning of three-dimensional right ventricular motion enables outcome prediction in pulmonary hypertension: a cardiac MR imaging study. Radiology. 2017;283:381–90.
15. Išgum I, Rutten A, Prokop M, Staring M, Klein S, Pluim JP, Viergever MA, van Ginneken B. Automated aortic calcium scoring on low-dose chest computed tomography. Med Phys. 2010;37:714–23.
16. Išgum I, Prokop M, Niemeijer M, Viergever MA, van Ginneken B. Automatic coronary calcium scoring in low-dose chest computed tomography. IEEE Trans Med Imaging. 2012;31:2322–34.
17. Lessmann N, van Ginneken B, Zreik M, de Jong PA, de Vos BD, Viergever MA, Išgum I. Automatic calcium scoring in low-dose chest CT using deep neural networks with dilated convolutions. IEEE Trans Med Imaging. 2018;37:615–25.
18. Schafer S, de Marvao A, Adami E, Fiedler LR, Ng B, Khin E, Rackham OJ, van Heesch S, Pua CJ, Kui M, Walsh R, Tayal U, Prasad SK, Dawes TJ, Ko NS, Sim D, Chan LL, Chin CW, Mazzarotto F, Barton PJ, Kreuchwig F, de Kleijn DP, Totman T, Biffi C, Tee N, Rueckert D, Schneider V, Faber A, Regitz-Zagrosek V, Seidman JG, Seidman CE, Linke WA, Kovalik JP, O'Regan D, Ware JS, Hubner N, Cook SA. Titin-truncating variants affect heart function in disease cohorts and the general population. Nat Genet. 2017;49:46–53.
19. Biffi C, de Marvao A, Attard MI, Dawes TJW, Whiffin N, Bai W, Shi W, Francis C, Meyer H, Buchan R, Cook SA, Rueckert D, O'Regan DP. Three-dimensional cardiovascular imaging-genetics: a mass

univariate framework. Bioinformatics. 2018;34:97–103.

20. Nagueh SF, Appleton CP, Gillebert TC, Marino PN, Oh JK, Smiseth OA, Waggoner AD, Flachskampf FA, Pellikka PA, Evangelista A. Recommendations for the evaluation of left ventricular diastolic function by echocardiography. J Am Soc Echocardiogr. 2009;22:107–33.
21. Picano E, Lattanzi F, Orlandini A, Marini C, L'Abbate A. Stress echocardiography and the human factor: the importance of being expert. J Am Coll Cardiol. 1991;17:666–9.
22. Chu WK, Raeside DE. Fourier analysis of the echocardiogram. Phys Med Biol. 1978;23:100–5.
23. Chu WK, Raeside DE, Chandraratna PA, Brown RE, Poehlmann H. Echocardiogram analysis in a pattern recognition framework. Med Phys. 1979;6:267–71.
24. Thavendiranathan P, Liu S, Verhaert D, Calleja A, Nitinunu A, Van Houten T, De Michelis N, Simonetti O, Rajagopalan S, Ryan T, Vannan MA. Feasibility, accuracy, and reproducibility of real-time full-volume 3D transthoracic echocardiography to measure LV volumes and systolic function: a fully automated endocardial contouring algorithm in sinus rhythm and atrial fibrillation. JACC Cardiovasc Imaging. 2012;5:239–51.
25. Knackstedt C, Bekkers SC, Schummers G, Schreckenberg M, Muraru D, Badano LP, Franke A, Bavishi C, Omar AM, Sengupta PP. Fully automated versus standard tracking of left ventricular ejection fraction and longitudinal strain: the FAST-EFs multicenter study. J Am Coll Cardiol. 2015;66:1456–66.
26. Tsang W, Salgo IS, Medvedofsky D, Takeuchi M, Prater D, Weinert L, Yamat M, Mor-Avi V, Patel AR, Lang RM. Transthoracic 3D echocardiographic left heart chamber quantification using an automated adaptive analytics algorithm. JACC Cardiovasc Imaging. 2016;9:769–82.
27. Otani K, Nakazono A, Salgo IS, Lang RM, Takeuchi M. Three-Dimensional echocardiographic assessment of left heart chamber size and function with fully automated quantification software in patients with atrial fibrillation. J Am Soc Echocardiogr. 2016;29:955–65.
28. Tamborini G, Piazzese C, Lang RM, Muratori M, Chiorino E, Mapelli M, Fusini L, Ali SG, Gripari P, Pontone G, Andreini D, Pepi M. Feasibility and accuracy of automated software for transthoracic three-dimensional left ventricular volume and function analysis: comparisons with two-dimensional echocardiography, three-dimensional transthoracic manual method, and cardiac magnetic resonance imaging. J Am Soc Echocardiogr. 2017;30:1049–58.
29. Carneiro G, Nascimento JC. Combining multiple dynamic models and deep learning architectures for tracking the left ventricle endocardium in ultrasound data. IEEE Trans Pattern Anal Mach Intell. 2013;35:2592–607.
30. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. Digit Med. 2018;1:6.
31. Narula S, Shameer K, Salem Omar AM, Dudley JT, Sengupta PP. Machine-learning algorithms to automate morphological and functional assessments in 2D echocardiography. J Am Coll Cardiol. 2016;68:2287–95.
32. Sengupta PP, Huang YM, Bansal M, Ashrafi A, Fisher M, Shameer K, Gall W, Dudley JT. Cognitive machine-learning algorithm for cardiac imaging: a pilot study for differentiating constrictive pericarditis from restrictive cardiomyopathy. Circ Cardiovasc Imaging. 2016;9:e004330.
33. Baumgartner H, Falk V, Bax JJ, De Bonis M, Hamm C, Holm PJ, Iung B, Lancellotti P, Lansac E, Rodriguez Muñoz D, Rosenhek R, Sjögren J, Tornos Mas P, Vahanian A, Walther T, Wendler O, Windecker S, Zamorano JL, ESC Scientific Document Group. 2017 ESC/EACTS guidelines for the management of valvular heart disease. Eur Heart J. 2017;38:2739–91.
34. Gandhi S, Mosleh W, Shen J, Chow CM. Automation, machine learning, and artificial intelligence in echocardiography: a brave new world. Echocardiography. 2018;35(9):1402–18. https://doi.org/10.1111/echo.14086.
35. Jin CN, Salgo IS, Schneider RJ, Kam KK, Chi WK, So CY, Tang Z, Wan S, Wong R, Underwood M, Lee AP. Using anatomic intelligence to localize mitral valve prolapse on three-dimensional echocardiography. J Am Soc Echocardiogr. 2016;29:938–45.
36. Calleja A, Thavendiranathan P, Ionasec RI, Houle H, Liu S, Voigt I, Sai Sudhakar C, Crestanello J, Ryan T, Vannan MA. Automated quantitative 3-dimensional modeling of the aortic valve and root by 3-dimensional transesophageal echocardiography in normals, aortic regurgitation, and aortic stenosis: comparison to computed tomography in normals and clinical implications. Circ Cardiovasc Imaging. 2013;6:99–108.
37. Warraich HJ, Shahul S, Matyal R, Mahmood F. Bench to bedside: dynamic mitral valve assessment. J Cardiothorac Vasc Anesth. 2011;25:863–6.
38. Budoff MJ, Young R, Burke G, Jeffrey Carr J, Detrano RC, Folsom AR, Kronmal R, Lima JAC, Liu KJ, McClelland RL, Michos E, Post WS, Shea S, Watson KE, Wong ND. Ten-year association of coronary artery calcium with atherosclerotic cardiovascular disease (ASCVD) events: the multi-ethnic study of atherosclerosis (MESA). Eur Heart J. 2018;39:2401–8.
39. Lee SE, Sung JM, Rizvi A, Lin FY, Kumar A, Hadamitzky M, Kim YJ, Conte E, Andreini D, Pontone G, Budoff MJ, Gottlieb I, Lee BK, Chun EJ, Cademartiri F, Maffei E, Marques H, Leipsic JA, Shin S, Hyun Choi J, Chinnaiyan K, Raff G, Virmani R, Samady H, Stone PH,

Berman DS, Narula J, Shaw LJ, Bax JJ, Min JK, Chang HJ. Quantification of coronary atherosclerosis in the assessment of coronary artery disease. Circ Cardiovasc Imaging. 2018;11:e007562.

40. Naoum C, Berman DS, Ahmadi A, Blanke P, Gransar H, Narula J, Shaw LJ, Kritharides L, Achenbach S, Al-Mallah MH, Andreini D, Budoff MJ, Cademartiri F, Callister TQ, Chang HJ, Chinnaiyan K, Chow B, Cury RC, DeLago A, Dunning A, Feuchtner G, Hadamitzky M, Hausleiter J, Kaufmann PA, Kim YJ, Maffei E, Marquez H, Pontone G, Raff G, Rubinshtein R, Villines TC, Min J, Leipsic J. Predictive value of age- and sex-specific nomograms of global plaque burden on coronary computed tomography angiography for major cardiac events. Circ Cardiovasc Imaging. 2017;10:e004896.
41. Cury RC, Abbara S, Achenbach S, Agatston A, Berman DS, Budoff MJ, Dill KE, Jacobs JE, Maroules CD, Rubin GD, Rybicki FJ, Schoepf UJ, Shaw LJ, Stillman AE, White CS, Woodard PK, Leipsic JA. CAD-RADS(TM) coronary artery disease – reporting and data system. An expert consensus document of the Society of Cardiovascular Computed Tomography (SCCT), the American College of Radiology (ACR) and the North American Society for Cardiovascular Imaging (NASCI). Endorsed by the American College of Cardiology. J Cardiovasc Comput Tomogr. 2016;10:269–81.
42. Wolterink JM, Leiner T, Takx RA, Viergever MA, Išgum I. Automatic coronary calcium scoring in non-contrast-enhanced ECG-triggered cardiac CT with ambiguity detection. IEEE Trans Med Imaging. 2015;34:1867–78.
43. Wolterink JM, Leiner T, de Vos BD, van Hamersvelt RW, Viergever MA, Išgum I. Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. Med Image Anal. 2016;34:123–36.
44. Wolterink JM, Leiner T, de Vos BD, Coatrieux JL, Kelm BM, Kondo S, Salgado RA, Shahzad R, Shu H, Snoeren M, Takx RA, van Vliet LJ, van Walsum T, Willems TP, Yang G, Zheng Y, Viergever MA, Išgum I. An evaluation of automatic coronary artery calcium scoring methods with cardiac CT using the orCaScore framework. Med Phys. 2016;43:2361.
45. Zreik M, van Hamersvelt RW, Wolterink JM, Leiner T, Viergever MA, Išgum I. Automatic detection and characterization of coronary artery plaque and stenosis using a recurrent convolutional neural network in coronary CT angiography. https://arxiv.org/abs/1804.04360
46. Kolossváry M, Karády J, Szilveszter B, Kitslaar P, Hoffmann U, Merkely B, Maurovich-Horvat P. Radiomic features are superior to conventional quantitative computed tomographic metrics to identify coronary plaques with napkin-ring sign. Circ Cardiovasc Imaging. 2017;10:e006843.
47. Mannil M, von Spiczak J, Manka R, Alkadhi H. Texture analysis and machine learning for detecting myocardial infarction in noncontrast low-dose computed tomography: unveiling the invisible. Investig Radiol. 2018;53:338–43.
48. Menke J, Kowalski J. Diagnostic accuracy and utility of coronary CT angiography with consideration of unevaluable results: a systematic review and multivariate Bayesian random-effects meta-analysis with intention to diagnose. Eur Radiol. 2016;26:451–8.
49. Zreik M, Lessmann N, van Hamersvelt RW, Wolterink JM, Voskuil M, Viergever MA, Leiner T, Išgum I. Deep learning analysis of the myocardium in coronary CT angiography for identification of patients with functionally significant coronary artery stenosis. Med Image Anal. 2018;44:72–85.
50. Coenen A, Kim YH, Kruk M, Tesche C, De Geer J, Kurata A, Lubbers ML, Daemen J, Itu L, Rapaka S, Sharma P, Schwemmer C, Persson A, Schoepf UJ, Kepka C, Hyun Yang D, Nieman K. Diagnostic accuracy of a machine-learning approach to coronary computed tomographic angiography-based fractional flow reserve: result from the MACHINE consortium. Circ Cardiovasc Imaging. 2018;11:e007217.
51. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, Andreini D, Budoff MJ, Cademartiri F, Callister TQ, Chang HJ, Chinnaiyan K, Chow BJ, Cury RC, Delago A, Gomez M, Gransar H, Hadamitzky M, Hausleiter J, Hindoyan N, Feuchtner G, Kaufmann PA, Kim YJ, Leipsic J, Lin FY, Maffei E, Marques H, Pontone G, Raff G, Rubinshtein R, Shaw LJ, Stehli J, Villines TC, Dunning A, Min JK, Slomka PJ. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. Eur Heart J. 2017;38:500–7.
52. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol. 2017;69:2657–64.
53. Peng P, Lekadir K, Gooya A, Shao L, Petersen SE, Frangi AF. A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. MAGMA. 2016;29:155–95.
54. Suinesiaputra A, Cowan BR, Al-Agamy AO, Elattar MA, Ayache N, Fahmy AS, Khalifa AM, Medrano-Gracia P, Jolly MP, Kadish AH, Lee DC, Margeta J, Warfield SK, Young AA. A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images. Med Image Anal. 2014;18:50–62.
55. Suinesiaputra A, Sanghvi MM, Aung N, Paiva JM, Zemrak F, Fung K, Lukaschuk E, Lee AM, Carapella V, Kim YJ, Francis J, Piechnik SK, Neubauer S, Greiser A, Jolly MP, Hayes C, Young AA, Petersen SE. Fully-automated left ventricular mass and volume MRI analysis in the UK Biobank population

cohort: evaluation of initial results. Int J Cardiovasc Imaging. 2018;34:281–91.

56. Jaarsma C, Leiner T, Bekkers SC, Crijns HJ, Wildberger JE, Nagel E, Nelemans PJ, Schalla S. Diagnostic performance of noninvasive myocardial perfusion imaging using single-photon emission computed tomography, cardiac magnetic resonance, and positron emission tomography imaging for the detection of obstructive coronary artery disease: a meta-analysis. J Am Coll Cardiol. 2012;59:1719–28.
57. Greenwood JP, Maredia N, Younger JF, Brown JM, Nixon J, Everett CC, Bijsterveld P, Ridgway JP, Radjenovic A, Dickinson CJ, Ball SG, Plein S. Cardiovascular magnetic resonance and single-photon emission computed tomography for diagnosis of coronary heart disease (CE-MARC): a prospective trial. Lancet. 2012;379:453–60.
58. Kellman P, Hansen MS, Nielles-Vallespin S, Nickander J, Themudo R, Ugander M, Xue H. Myocardial perfusion cardiovascular magnetic resonance: optimized dual sequence and reconstruction for quantification. J Cardiovasc Magn Reson. 2017;19:43.
59. Tarroni G, Corsi C, Antkowiak PF, Veronesi F, Kramer CM, Epstein FH, Walter J, Lamberti C, Lang RM, Mor-Avi V, Patel AR. Myocardial perfusion: near-automated evaluation from contrast-enhanced MR images obtained at rest and during vasodilator stress. Radiology. 2012;265:576–83.
60. Dyverfeldt P, Bissell M, Barker AJ, Bolger AF, Carlhäll CJ, Ebbers T, Francios CJ, Frydrychowicz A, Geiger J, Giese D, Hope MD, Kilner PJ, Kozerke S, Myerson S, Neubauer S, Wieben O, Markl M. 4D flow cardiovascular magnetic resonance consensus statement. J Cardiovasc Magn Reson. 2015;17:72.
61. Crandon S, Elbaz MSM, Westenberg JJM, van der Geest RJ, Plein S, Garg P. Clinical applications of intra-cardiac four-dimensional flow cardiovascular magnetic resonance: a systematic review. Int J Cardiol. 2017;249:486–93.
62. Karim R, Bhagirath P, Claus P, James Housden R, Chen Z, Karimaghaloo Z, Sohn HM, Lara Rodríguez L, Vera S, Albà X, Hennemuth A, Peitgen HO, Arbel T, Gonzàlez Ballester MA, Frangi AF, Götte M, Razavi R, Schaeffter T, Rhode K. Evaluation of state-of-the-art segmentation algorithms for left ventricle infarct from late gadolinium enhancement MR images. Med Image Anal. 2016;30:95–107.
63. Rayatzadeh H, Tan A, Chan RH, Patel SJ, Hauser TH, Ngo L, Shaw JL, Hong SN, Zimetbaum P, Buxton AE, Josephson ME, Manning WJ, Nezafat R. Scar heterogeneity on cardiovascular magnetic resonance as a predictor of appropriate implantable cardioverter defibrillator therapy. J Cardiovasc Magn Reson. 2013;15:31.
63. Samad MD, Wehner GJ, Arbabshirani MR, Jing L, Powell AJ, Geva T, Haggerty CM, Fornwalt BK. Predicting deterioration of ventricular function in patients with repaired tetralogy of Fallot using machine learning. Eur Heart J Cardiovasc Imaging. 2018;19:730–8.
65. Yang B, Ying L, Tang J. Artificial neural network enhanced Bayesian PET image reconstruction. IEEE Trans Med Imaging. 2018;37:1297–309.
66. Kim K, Wu D, Gong K, Dutta J, Kim JH, Son YD, Kim HK, El Fakhri G, Li Q. Penalized PET reconstruction using deep learning prior and local linear fitting. IEEE Trans Med Imaging. 2018;37:1478–87.
67. Arsanjani R, Xu Y, Dey D, Vahistha V, Shalev A, Nakanishi R, Hayes S, Fish M, Berman D, Germano G, Slomka PJ. Improved accuracy of myocardial perfusion SPECT for detection of coronary artery disease by machine learning in a large population. J Nucl Cardiol. 2013;20:553–62.
68. Betancur J, Commandeur F, Motlagh M, Sharir T, Einstein AJ, Bokhari S, Fish MB, Ruddy TD, Kaufmann P, Sinusas AJ, Miller EJ, Bateman TM, Dorbala S, Di Carli M, Germano G, Otaki Y, Tamarappoo BK, Dey D, Berman DS, Slomka PJ. Deep learning for prediction of obstructive disease from fast myocardial perfusion SPECT: a multicenter study. JACC Cardiovasc Imaging. 2018; https://doi.org/10.1016/j.jcmg.2018.01.020. S1936-878X(18)30131-1
69. Arsanjani R, Dey D, Khachatryan T, Shalev A, Hayes SW, Fish M, Nakanishi R, Germano G, Berman DS, Slomka P. Prediction of revascularization after myocardial perfusion SPECT by machine learning in a large population. J Nucl Cardiol. 2015;22:877–84.
70. Betancur J, Otaki Y, Motwani M, Fish MB, Lemley M, Dey D, Gransar H, Tamarappoo B, Germano G, Sharir T, Berman DS, Slomka PJ. Prognostic value of combined clinical and myocardial perfusion imaging data using machine learning. JACC Cardiovasc Imaging. 2018;11:1000–9.
71. Haro Alonso D, Wernick MN, Yang Y, Germano G, Berman DS, Slomka P. Prediction of cardiac death after adenosine myocardial perfusion SPECT based on machine learning. J Nucl Cardiol. 2018; https://doi.org/10.1007/s12350-018-1250-7. [Epub ahead of print]
72. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, Gomes AS, Folsom AR, Shea S, Guallar E, Bluemke DA, Lima JAC. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis. Circ Res. 2017;121:1092–101.
73. https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm587890.htm. Accessed July 22 2018.
74. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, Greenland P, Jacob DR Jr, Kronmal R, Liu K, Nelson JC, O'Leary D, Saad MF, Shea S, Szklo M, Tracy RP. Multi-ethnic study of atherosclerosis: objectives and design. Am J Epidemiol. 2002;156:871–81.

75. Fonseca CG, Backhaus M, Bluemke DA, Britten RD, Chung JD, Cowan BR, Dinov ID, Finn JP, Hunter PJ, Kadish AH, Lee DC, Lima JA, Medrano-Gracia P, Shivkumar K, Suinesiaputra A, Tao W, Young AA. The cardiac atlas project—an imaging database for computational modeling and statistical atlases of the heart. Bioinformatics. 2011;27:2288–95.
76. Chu WK, Chandraratna PA, Raeside DE, Brown RE, Poehlman H. Toward the automation of echocardiography. Radiology. 1977;123:795–7.

14 Deep Learning in Breast Cancer Screening

Hugh Harvey, Andreas Heindl, Galvin Khara, Dimitrios Korkinof, Michael O'Neill, Joseph Yearsley, Edith Karpati, Tobias Rijken, Peter Kecskemethy, and Gabor Forrai

14.1 Background

Out of the myriad proposed use-cases for artificial intelligence in radiology, breast cancer screening is perhaps the best known and most researched. Computer aided detection (CAD) systems have been available for over a decade, meaning that the application of more recent deep learning techniques to mammography already has a benchmark against which to compete. For the most part there is also a behavioral change barrier to overcome before deep learning technologies are accepted into clinical practice, made even more difficult by CAD's largely unconvincing performance compared to its promise.

In this chapter we discuss the history of breast cancer screening, the rise and fall of traditional CAD systems, and explore the different deep learning techniques and their associated common challenges.

14.1.1 The Breast Cancer Screening Global Landscape

Breast cancer is currently the most frequent cancer and the most frequent cause of cancer-induced deaths in women in Europe [1]. The favorable results of randomized clinical trials have led to the implementation of regional and national population-based screening programmes for breast cancer in many upper-middle-income countries since the end of the 1980s [2]. The primary aim of a breast screening programme is to reduce mortality from breast cancer through early detection. Early detection of cancer comprises of two strategies: screening and early diagnosis.

- Screening involves the systematic application of a screening test for a specific cancer in an asymptomatic population in order to detect and treat early cancer or pre-cancerous health conditions before they become a threat to the well-being of the individual or the community. Mammography is the cornerstone of breast cancer screening and is widely offered as a public health policy on a routine basis.
- Early diagnosis is based on improved public and professional awareness (particularly at the primary health care level) of signs and symptoms associated with cancer, improved

H. Harvey (✉) · A. Heindl · G. Khara · D. Korkinof ·
M. O'Neill · J. Yearsley · E. Karpati · T. Rijken ·
P. Kecskemethy
Kheiron Medical Technologies, London, UK
e-mail: hugh@kheironmed.com

G. Forrai
European Society of Breast Imaging, Vienna, Austria
e-mail: forrai.gabor@t-online.hu

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_14

> health-care-seeking behavior, prompt clinical assessment, and early referral of suspected cancer cases, such that appropriate diagnostic investigations and treatment can be rapidly instituted leading to improved mortality outcomes. Indeed, all developed European countries have incorporated nationwide organized screening programmes, starting from the 1980s.

Breast cancer screening using mammography has been proven to be the most effective single imaging tool at the population level for detecting breast cancer in its earliest and most treatable stage [3]. However, according to a review by Boyer et al. [4], the limiting factors with the use of mammography are the breast's structure (high density) and the fact that mammography is difficult to interpret, even for experts. As a consequence mistakes during interpretation may occur due to fatigue, lack of attention, failure in detection or interpretation, and can lead to significant inter and intra-observer variation. There are as many missed lesions when they have not been seen (i.e., reading errors) as when they have been incorrectly judged (i.e., decision errors) [5]. The reported rate of missed cancers varies from 16 to 31% [6]. To reduce this rate, double-reading of screening mammograms by two independent experts was introduced in some programs. Blinded double-reading reportedly reduces false negative results and the average radiologist can expect an 8–14% gain in sensitivity and a 4–10% increase in specificity with double-reading pairing [7]. This is not surprising considering that double readers are often more specialized, and read a larger volume of cases per year. Double-reading is now well established in many European countries, whereas in the USA, where double-reading is not mandatory, single-reading programmes with ad-hoc non-invitational screening are more common.

The most common direct harm associated with errors in mammography are false positive test results [8] which cause additional work and costs for health care providers, and emotional stress and worry for patients. Patient harms arise from false negatives, leading to delays in diagnosis and an increase in interval cancers downstream. According to Fletcher [9], false positive test results increase when technology increases sensitivity but decreases specificity. While the use of more sensitive and less specific technology may be appropriate for patients at very high risk of developing cancer (such as those with BRCA mutations or untested women with first-degree relatives with BRCA mutations) use of these tests are not appropriate for general populations at lower risk. In Europe, cancer detection rates are similar to those in the USA [10], but European frequency of false positives is lower [11], which we may assume are due to differences in the medico-legal environment, existence of double reporting programmes, guidelines for appropriate false positive rates, and more standardized training requirements for mammographic readers.

A Brief History of UK Breast Screening

- 1986: In the UK, Professor Sir Patrick Forrest produced the "Forrest report," commissioned by the forward thinking Health Secretary, Kenneth Clarke. Having taken evidence on the effectiveness of breast cancer screening from several international trials (America, Holland, Sweden, Scotland, and the UK), Forrest concluded that the NHS should set up a national breast screening program.
- 1988: This was swiftly incorporated into practice, and by 1988 the NHS had the world's first invitation-based breast cancer screening program. "Forrest units" were set up across the UK for screening women between 50 and 64 years old who were invited every 3 years for a standard mammogram (two compressed views of each breast).
- 2000: With the advent of digital mammography, medical imaging data became available in a format amenable to computational analysis.

(continued)

- 2004: Researchers developed what became known as CAD, using feature-engineered programs to highlight abnormalities on mammograms. These systems use hand-crafted features such as breast density, parenchymal texture, the presence of a mass or microcalcifications to determine whether or not a cancer might be present. They were designed to alert a radiologist to the presence of a specific feature by attempting to mimic an expert's decision-making process by highlighting regions on a mammogram according to recognized characteristics.
- 2012: Ongoing analysis of the screening program proved the benefit of widening the age range for invitation to screening to between 47 and 73 years old.
- 2014: The UK system was now successfully discovering over 50% of the female population's breast cancers (within the target age range) before they became symptomatic. Screening is now the internationally recognized hallmark for best practice.

14.1.2 The Rise and Fall of CAD

14.1.2.1 Rise: The Premise and Promise

The success of screening programs has driven both demand and costs, with an estimated 37 million mammograms now being performed each year in the USA alone [12]. There are consequently not enough human radiologists to keep up with the workload. The situation is even more pressured in European countries where double-reading essentially requires that the number of radiologists per case is double than that of the USA. The shortage of expensive specialized breast radiologists is getting so acute that murmurs of reducing the benchmark of double-reading to single-reading in the UK are being uttered, despite convincing evidence that double-reading is simply better. The cost-inefficiency of double-reading, however, is a tempting target for policy makers looking to trim expenditure.

Initially, CAD was optimistically seen as a tool that would augment the radiologist, helping lower the potential to miss cancers on a mammogram (false negatives) and reducing the frequency of false positives. Ultimately, CAD was positioned as a means to improve the economic outcomes of screening by tackling both of these challenges. It made use of some of the most advanced techniques for the time during its boom in the late 1990s. The most common user interface of CAD is that of overlaid markings on top of a mammography image indicating the areas which the CAD has processed and detected as potentially representing a malignant feature. While there are different systems on the market, they provide broadly similar outputs.

Feature extraction utilizes machine recognition of hand-engineered visual motifs. An early example is ImageChecker M1000 which detected spiculated lesions by identifying radiating lines emerging from a 6-mm center within a 32-mm circle [13]. While this helped the method to be interpretable it also led to significant detection limitations that manifested in a greater number of false positives as it struggled to account for all eventualities.

Once the features are extracted another method is used to decide, based on those features, whether an area is malignant or not. Such discriminators are traditionally rule based, decisions trees, support-vector machines, or multi-layer perceptrons. An example of this could be a simple rule such as "*if* spiculated mass present *then* cancer." These methods have many issues due to their oversimplification of the underlying problem.

14.1.2.2 How Does CAD Perform?

There is a wide body of literature examining the performance of different CAD systems, most commonly ImageChecker (R2, now Hologic) and iCAD SecondLook. These studies used several different methodologies, so direct comparisons between them are difficult. In this section we review the most seminal of these works, noting the differences in methodologies, sample sizes, outcome measurements, and conclusions. The variation in these studies makes for a somewhat heterogeneous overall picture, with some studies strongly advocating for the use of CAD, and others showing no significant benefits [14].

Gilbert et al. [15] conducted a trial to determine whether the performance of a single reader using a CAD (ImageChecker) would match the performance achieved by two readers. 28,204 subjects were screened in this prospective study. Authors reported similar cancer detection rates, sensitivity, specificity, and positive predictive value after single-reading with CAD and after double-reading, and a higher recall rate after single-reading with CAD. Finally, no difference was found in pathological attributes between tumors detected by single-reading with CAD alone and those detected by double-reading alone. The authors concluded that single-reading with CAD could be an alternative to double-reading, especially for the detection of small breast cancers where the double-reading remains the best method, and could improve the rate of detection of cancer from screening mammograms read by a single reader.

A second prospective analysis called CADETII published by the same group [16] was conducted to evaluate the mammographic features of breast cancer that favor lesion detection with single-reading and CAD or with double-reading. Similar results were obtained for patients in whom the predominant radiologic feature was either a mass or a microcalcification. However, authors reported superior performance for double-reading in the detection of cancers that manifested as parenchymal deformities and superior performance for single-reading with CAD in the detection of cancers that manifested as asymmetric densities, suggesting that for more challenging cancer cases, both reading protocols have strengths and weaknesses. However, there was a small but significant relative difference in recall rates of 18% between the two study groups.

Taylor and Potts [17] performed a meta-analysis to estimate impact of CAD and double-reading respectively on odds ratios for cancer detection and recall rates. Meta-analysis included 10 studies comparing single-reading with CAD to single-reading and 17 studies comparing double to single-reading. All studies were published between 1991 and 2008. Despite an evident heterogeneity between the studies, evidences were sufficient to claim that double-reading increases cancer detection rate and that double-reading with arbitration does so while lowering recall rate. However, evidences were insufficient to claim that CAD improves cancer detection rates, while CAD clearly increased recall rate. When comparing CAD and double-reading with arbitration, authors did not find a difference in cancer detection rate, but double-reading with arbitration showed a significantly better recall rate. Based on these findings, authors concluded that the best current evidence shows grounds for preferring double-reading to single-reading with CAD.

Noble et al. [18] aimed to assess the diagnostic performance of a CAD (ImageChecker) for screening mammography in terms of sensitivity, specificity, incremental recall, and cancer diagnosis rates. This meta-analysis was based on the results of three retrospective studies and four prospective studies published between 2001 and 2008. Authors reported strong heterogeneity in the results between the different studies. They supposed that several environmental factors could influence the CAD performances, including accuracy and experience of radiologist, i.e., very accurate radiologists may have a smaller incremental cancer detection rate using CAD than less accurate or less experienced radiologists because they would miss fewer cases of cancer without CAD. Radiologists who are more confident in their interpretation skills may also be less likely to recall healthy women primarily based upon CAD findings. On the other hand, less confident mammographers concerned about false negative readings may recommend the recall of a greater proportion of healthy women when CAD is employed. Based on these findings, it was difficult to draw conclusions on the beneficial impact of using a CAD for screening mammography.

Karssemeijer et al. [19] conducted a prospective study to compare full field digital mammography (FFDM) reading using CAD (ImageChecker) with screen film mammography (SFM) in a population-based breast cancer screening program for initial and subsequent screening examinations. In total, 367,600 screening examinations were performed and results similar to previous studies were reported, i.e., the detection of ductal carcinoma in situ and microcalcification clusters improved with FFDM using CAD, while the recall rate increased. In conclusion, this study supports the use of a CAD to assist a single reader in centralized screening programs for the detection of breast cancers.

Destounis et al. [20] conducted a retrospective study to evaluate the ability of the CAD ImageChecker to identify breast carcinoma in standard mammographic projections based on the analysis of 45 biopsy-proven lesions and 44 screening BIRADS category 1 digital mammography examinations which were used as a comparative normal/control population. CAD demonstrated a lesion/case sensitivity of 87%. The image sensitivity was found to be 69% in the MLO (mediolateral oblique) view, and 78% in the CC (craniocaudal) view. For this evaluation, CAD was able to detect all lesion types across the range of breast densities supporting the use of a CAD to assist in the detection of breast cancers.

van den Biggelaar et al. [21] prospectively aimed to assess the impact of different mammogram reading strategies on the diagnosis of breast cancer in 1048 consecutive patients referred for digital mammography to a hospital (i.e., symptomatic, not a screening population). The following reading strategies were implemented: single-reading by a radiologist with or without CAD (iCAD Second Look), breast technologists employed as pre-readers or double readers. Authors reported that the strategy of double-reading mammograms by a radiologist and a technologist obtained the highest diagnostic yield in this patient population, as compared to the strategy of pre-reading by technologists or the conventional strategy of mammogram reading by the radiologist only. Comparing the findings in the different reading strategies showed that double-reading resulted in a higher sensitivity at the cost of a lower specificity, whereas pre-reading resulted in a higher specificity at the cost of a lower sensitivity. In addition, the results of the present study demonstrated that systematic application of CAD software in a clinical population failed to improve the performance of both radiologist and technologist readers. In conclusion, this study does not support the use of CAD in the detection of breast cancers.

Sohns et al. [22] conducted a retrospective study to assess the clinical usefulness of ImageChecker in the interpretation of early research, benign, and malignant mammograms. 303 patients were analyzed by three single readers with different experience with and without the CAD. Authors reported that the three readers could increase their accuracy by the aid of the CAD system with the strongest benefit for the less experienced reader. They also reported that the increase of accuracy was strongly dependent on the readers' experience. In conclusion, this study supported the use of a CAD in the interpretation of mammograms for the detection of breast cancers especially for less experienced readers.

Murakami et al. [23] assessed, in a retrospective study including 152 patients, the usefulness of the iCAD SecondLook in the detection of breast cancers. Authors reported that the use of a CAD system with digital mammography (DM) could identify 91% of breast cancers with a high sensitivity for cancers manifesting as calcifications (100%) or masses (98%). Of particular interest, authors also reported that sensitivity was maintained for cancers with a histopathology for which the sensitivity of mammography is known to be lower (i.e., invasive lobular carcinomas and small neoplasms). In conclusion, this study supported the use of a CAD as an effective tool for assisting the diagnosis of early breast cancer.

Cole et al. [24] aimed to assess the impact of two CAD systems on the performance of radiologists with digital mammograms. 300 cases were retrospectively reviewed by 14 and

15 radiologists using respectively the iCAD SecondLook and the R2 Image Checker. Authors reported that although both CADs increased area under the curve (AUC) and sensitivity of the readers, the average differences observed were not statistically significant. Cole et al. concluded that radiologists rarely changed their diagnostic decision after the addition of CAD, regardless of which CAD system was used.

A study conducted by Bargalló et al. [25] aimed to assess the impact of shifting from a standard double-reading plus arbitration protocol to a single-reading by experienced radiologists assisted by CAD in a breast cancer screening program. During the 8 years of this prospective study, 47,462 consecutive screening mammograms were reviewed. As main findings, the authors reported an increase of the cancer detection rate in the period when the single reader was assisted by a CAD (iCAD SecondLook), which could be even higher depending on the experience of the radiologist, i.e., specialized breast radiologists performed better than general radiologists. The recall rate was slightly increased during the period when the single reader was assisted by a CAD (iCAD SecondLook). However, this increase could be, according to the authors, related to the absence of arbitration which was responsible for the strong reduction of recall rate in the double reader protocol. In conclusion, this study supported the use of a CAD to assist a single reader in centralized screening programs for the detection of breast cancers.

Lehman et al. [26] aimed to measure performance of digital screening mammography with and without CAD (unidentified manufacturer) in US community practice. Authors compared retrospectively accuracies of digital screening mammography interpreted with ($N = 495{,}818$) and without ($N = 129{,}807$) CAD. Authors did not report any improvement in sensitivity, specificity, recall rates, except for the detection of intraductal carcinoma when the CAD was used and concluded that there was no beneficial impact of using CAD for mammography interpretation. This study did not support the use of CAD in breast cancer detection in a screening setting.

14.1.2.3 So Why Did CAD "Fail"?

In the USA the use of CAD systems earns radiologists extra reimbursement (varying between 15 and 40, but trending downwards due to recent bundling of billing codes). Such reimbursement has undoubtedly incentivized clinical uptake in a payer system. As a result, the use of CAD in US screening has effectively prevented adoption of double-reading (which would be more costly); however, with the performance of CAD systems coming into disrepute, some have questioned if monetarily favoring the use of such a system is economically correct when their accuracy and efficacy is in doubt [27].

As discussed above, many published studies have yielded ambiguous results; however, most found the use of CAD to be associated with a higher sensitivity, but lower specificity. Sanchez et al. [28] observed similar trends in a Spanish screening population, where CAD by itself produced a sensitivity of 84% and a corresponding specificity of 13.2%. Freer and Ulissey [29] found the use of CAD caused a recall rate increase of 1.2% with a 19.5% increase in the number of cancers detected. These studies were on CADs that analyzed screen film mammograms which were digitized, and not FFDM. Others [30] found that the use of CAD on FFDM detected 93% of calcifications, and 92% of masses, at a cost of 2.3 false positive marks per case on average.

Despite heterogeneous results, by 2008 over 70% of screening hospitals in the USA had adopted CAD [31]. This widespread adoption, coupled with a lack of definitive positive evidence with CAD usage in a screening setting, has resulted in some skepticism among radiologists. This all coming at a cost of over $400 million a year in the USA [32]. In essence, while CAD was sometimes shown to improve sensitivity, it often decreased specificity, leading to distraction of radiologists by having to check false positive markings [33], and increasing recall rates. Conversely, in Europe, CAD uptake is under 10%, and human double-reading is far more widespread, with significantly lower recall rates.

Table 14.1 Reported and minimum human-level performance for mammography reading

Source	Sensitivity	Specificity
Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate [26]	87.3%	91.4%
US national performance benchmarks for modern screening digital mammography [34]	87%	89%
Minimally acceptable interpretive performance criteria for screening mammography [35]	75%	88%
Criteria for identifying radiologists with acceptable screening mammography interpretive performance [36]	≥80%	≥85%
Breast cancer surveillance consortium US digital screening mammography [37]	84%	91%

The technical reason for CADs inability to perform at the level of a human reader (Table 14.1) is due to the underlying algorithms' lack of flexibility in predicting the diverse range of abnormalities that arise biologically in the breast. Microcalcifications are morphologically completely different to masses, and each has their own family of subclasses with distinct shapes, sizes, and orientations. Architectural distortions exhibit even more subtlety. Disappointment was inevitable due to this huge level of variation, coupled with the fact that traditional CAD algorithms have to account for all of these structural and contextual differences explicitly in the form of hand-engineered features (which themselves are based on heuristics or mathematical pixel distributions). Deep learning algorithms do not suffer from this problem, as they adaptively learn useful features depending on the task at hand by themselves (at the expense of requiring more computational power and more data to train, with a potential decrease in interpretability). Despite this no system has yet reported both superior stand-alone sensitivity *and* specificity for even the minimum acceptance criteria.

Regardless of whether CAD adoption has resulted in a net positive or negative to a radiologist's workflow, a number of facts cannot be disputed. Traditional CAD by itself is not capable of replicating performance similar to a radiologist, thus a significant portion of a radiologist's time is spent either discarding areas marked by the system due to its high false positive rate, or second guessing a CAD's false negative misses. Any cancers missed or falsely flagged by CAD have significant downstream costs on both the health care system and patient welfare—and in summary make CAD technologically insufficient for making a positive net impact. Thus, the ultimate goal of future software based on deep learning is to detect malignancies and diagnose screening cases at a level that undoubtedly supports radiologists, which is at or beyond the level of an experienced single reader's average performance. This will ensure that when used to support a single-reader setting, diagnostic time is minimized, with a minimum number of false positive marks. Such a system could ensure that sensitivity and specificity is not sacrificed if the second read in a double reader program is carried out by a deep learning system. Perhaps most importantly, the reduction in false positives could lead to significant benefits to health care costs and patient outcomes. This opens up the possibility of implementing cost-effective screening programs in developing countries, where the lack of trained radiologists makes it impossible in the current climate.

14.1.3 A Brief History of Deep Learning for Mammography

The first successful application of deep learning in mammography dates back to 1996 [38]. The authors proposed a patch-based system that is able to detect the presence of a mass in regions of interest (ROI). The decision of inspecting mammograms by extracting small patches was motivated by the limited amount of computational resources available at that time. These early deep learning systems were not designed to detect microcalcifications because the authors referred to previous publications that claimed that existing CAD can detect them reliably enough. Subsequently it took some years of development on the hardware side until researchers got interested again in deep learning-based approaches. Dhungel et al. [39], and Ertosun and Rubin [40] can be accredited with starting off the new wave of deep learning with hybrid approaches, combining traditional machine learning with deep learning. The former suggested a cascaded CNN-based approach followed by a random forest and classical image post-processing. The latter published a two-stage deep learning system where the first classifies whether the image contains a mass, and the second localizes these masses.

In 2015 Carneiro et al. [41] achieved significant improvement for mass and microcalcification detection by using deep learning networks that were pre-trained on ImageNet which is a collection of about 14 million annotated real-life images. Deep neural networks seemed to be able to learn robust high-level features despite the significant differences in image content. Furthermore the authors made use of multiple views (MLO and CC) without pre-registration of the input images. Evaluation was done on the InBreast [42] and DDSM [43] datasets.

In 2017 Kooi et al. [44] published a deep learning-based approach trained on a much larger dataset of 45,000 images. Again, the authors proposed patch-based deep learning approach that focused on the detection of solid malignant lesions including architectural distortions, thus ignoring cysts or fibroadenomata. A small reader study showed that their algorithm had similar patch-level performance as three experienced readers.

In the same year Teare et al. [45] described a complex network that can deal with three classes encountered in mammography (normal, benign, malignant). Their approach encompasses a full image-based analysis where an enhanced input image is derived by a composition of three differently contrasted original input images. In parallel to the full image analysis, the authors suggest using a second patch-based network. Both outputs are later combined using a random forest that computes the final probability of cancer and the location of the suspected abnormality. The study did not provide any details on the proprietary dataset (for example, the hardware manufacturer, the distribution of lesion types, etc.)

Compared to the aforementioned approaches, Kim et al. [46] propose a setup that is based on pure data-driven features from raw mammograms without any lesion annotations. An interesting observation that the authors make is that their study yielded a better sensitivity on samples with a mass rather than calcifications, something which contradicts previous reports utilizing traditional CAD methodologies [47]. In order to overcome vendor-specific image characteristics, the authors applied random perturbation of the pixel intensity. Differences in diagnostic performance on different hardware vendors was attributed to underlying differences in the distribution of malignant cases for each vendor. One limitation of the proposed approach is that benign cases were excluded completely, which is obviously problematic when comparing results to a real-world clinical setting.

14.2 Goals for Automated Systems

Automated systems designed to assist radiologists in mammographic screening have three high-level tasks to achieve.

14.2.1 Recall Decision Support

The goal in screening is to make a decision about whether to recall a patient for further assessment or not. This, in essence, is a high-level binary output from a human radiologist for each and every screening case they read. Recall rates vary between different screening programmes, tending to be higher in the USA than in Europe and very low in Northern Europe. The UK guidelines suggest a recall rate as low as <5–7% is achievable [48], whereas in the USA recall rates of up to 15% have been reported [49].

Assuming an autonomous system can achieve a very high (99%) sensitivity at a suitable recall rate, the possibility of triaging opens up, thus aiding screening units in optimizing their workflow.

A powerful-enough software system learning from the past failures of CAD could be used to act as an effective second reader if its sensitivity and specificity is good enough (most likely requiring performance at or beyond the level of a human reader). Because of the vastly different performance, this setup is qualitatively different from CADs of the past in that they don't directly support a single reader but actually act as a reader themselves, although with vastly different "way of thinking" (capabilities and error profiles) than radiologists, thus complementing radiologists.

In 2016, the DREAM challenge was set up, inviting deep learning researchers to develop systems to detect breast cancers on a proprietary dataset for a monetary grand prize. The input data consisted of around 640,000 images of both breasts and, if available, previous screening exams of the same subject, clinical/demographic information such as race, age, and family history of breast cancer. The winning team (Therapixel, France) attained a specificity of 80.8% at a set sensitivity of 80% (AUC 0.87) [50]. This represented the first public competition to apply deep learning to screening mammography. However, none of the entrants, including the winners, reached close to the performance of single human reading radiologists (Table 14.1). This may have been due to issues in the underlying data, its labeling, limitations in the competition design, or more simply availability of mature deep learning techniques in radiology at this relatively early time for the field.

Teare et al. [45] report a case-wise AUC of 0.92, with a sensitivity of 91%, and specificity of 80.4%. Kim et al. [46] report an AUC of 0.90, with their case-wise decisions being derived from summations of individual lesion detection. Ribli et al. [51] trained a network to segment lesions achieving an AUC of 0.85. They also showed an AUC of 0.96 on the InBreast dataset and 0.3 false positives per image at a sensitivity of 90%. These results show a significant improvement on earlier CAD performance; however, the recall rates for these newer deep learning systems are not yet known, and again, human-level performance in recall decision-making is not yet met. Unfortunately, the fact that these assessments were performed on datasets of limited quality limits the conclusiveness of the results.

14.2.2 Lesion Localization

A binary recall decision is of limited interpretability, and begs the question "How can we know what the decision was based on?". For each recall decision, it would be both useful and reassuring to see the suspicious regions that led to the decision. Ideally the algorithm would not only be able to discriminate between normal and suspicious regions, but would also be able to display where these features are present on the image. This process is called localization.

Traditional CAD systems offered this kind of localization. However, as discussed in Sect. 14.1.2.3 these systems were undiscerning, generating so many false positives as to render the localizations almost meaningless. Deep-CNNs are also capable of localization and top the leader-boards of all the (non-radiological) major localization challenges such as PASCAL VOC, COCO, and ImageNet. It is natural that they be applied to digital mammography too.

Most attempts to perform localization in mammography with deep learning algorithms have taken a patch-based approach [41, 44, 52–54]. At training time, individual patches are sampled and cropped from a full sized image and fed through a CNN to produce class predictions for each patch (for example, malignant mass, benign mass, or background). At test time, the network can be slid incrementally over the image to ensure full prediction coverage. The patch-based approach mitigates the difficulty of fitting full-sized mammograms into memory. However, it also suffers two major drawbacks. Firstly, by cropping patches from the full image we are asking the network to base its classification decision on a small fraction of the available context. Imagine cutting out a section of irregular parenchyma from an image and trying to decide whether it is normal tissue or a mass. To make a reasoned judgment, we need to "take a step back" and study the whole surrounding area. What looks normal in a fibrous parenchyma may be a clear anomaly in a fatty breast. The result is typically a large number of false positives when the patches are re-combined back into the full image. The second major drawback is the inefficiency of the process. A significant amount of redundant computation is performed over the overlapping regions of the input patches.

The current state of the art in deep learning localization does not include patch-based methods. Rather, it is semantic segmentation [55], object detection [56–58], and instance segmentation approaches (Fig. 14.1) that top the leaderboards of the big public dataset challenges. These approaches all consider the full image rather than cropped patches. This allows them to overcome the two major drawbacks of the patch-based approach: the network now sees the full context of the image, and the forward pass computation is highly amortized over overlapping regions of the image.

In semantic segmentation the goal is to classify each individual pixel in an image as belonging to one class or another. In order to do this, the low-resolution encoding of the image produced by the contracting network of the CNN backbone must be gradually expanded back to full image resolution. This can be achieved by appending a number of subsequent layers to the CNN backbone, with the pooling operations of these layers replaced by up-sampling operations. In order to recover the location specific information lost during the pooling operations in the contracting path, high resolution features from the contracting path must be re-combined with the up-sampled feature maps via skip connections. Ronneberger et al. [59] had breakthrough success with the application of such a segmentation network to biomedical images, winning the ISBI cell tracking challenge 2015 by a large margin. Variants of this network are also being applied in digital mammography [60, 61].

Semantic segmentation cannot distinguish between separate *instances* of each class, for example, if there were two masses in the image, the algorithm would not tell you which pixels belonged to the first mass and which to the second, but would simply say "here are all the mass pixels." In object detection, the goal *is* to

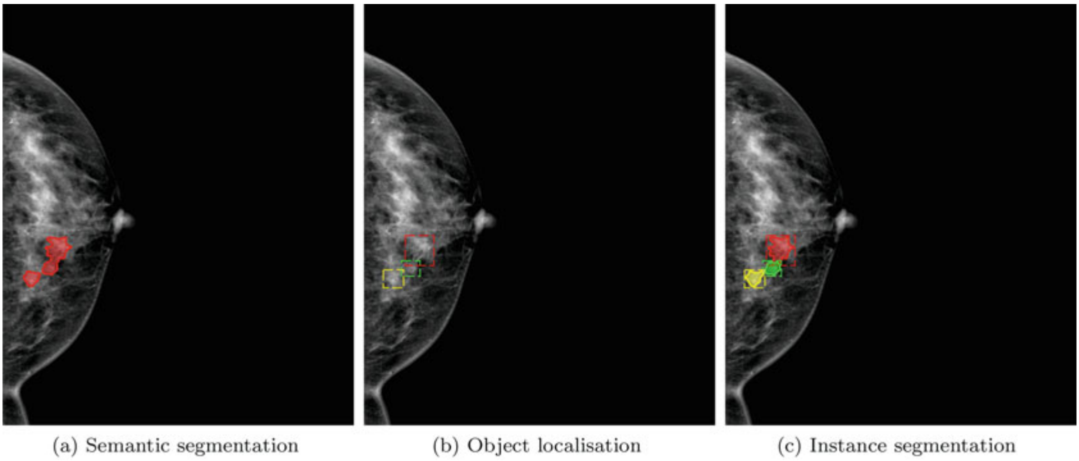

(a) Semantic segmentation (b) Object localisation (c) Instance segmentation

Fig. 14.1 Comparison of the three state-of-the-art localization approaches in deep learning. In (**a**) each pixel is classified as one class or another (here mass vs background). In (**b**) each mass instance is separately identified via bounding boxes. In (**c**) approaches (**a**) and (**b**) are combined by providing bounding boxes and pixel-level labels for each separate mass instance

separately identify each instance of each class. However, unlike semantic segmentation, this is not done on the pixel level. Often, some sort of selective search is used to generate region proposals with a high "objectness" score [56,57,62], and this is followed by a classifier head to predict a class probability for each proposed region, and a regressor head to adjust shapes of the bounding boxes that define each region for more precise localization. Ribli et al. [51] came runner up in the digital mammography DREAM challenge using the Faster-RCNN object detection network [56] (although the challenge was actually based on case-wise recall decision, the high sensitivity and specificity of the localizations meant that the case-wise label could also be reliably inferred).

Instance segmentation combines the pixel-level detail of semantic segmentation with the instance-level aspect of object detection. It classifies each pixel according to both class and instance. In other words, it tells you "here are all the masses, and here are all the pixels that belong to each one." This approach extends object detection by including a semantic segmentation branch in parallel with the classifier and bounding box regression branches. For each detected object, we now get a refined bounding box, a class probability, and a segmentation mask. The state of the art in this task is the Mask-RCNN network [63], the direct descendant of Faster-RCNN, although as of yet there are no reported results of this network on mammography datasets.

14.2.3 Density Stratification and Risk Prediction

It is widely accepted that the density of breast tissue—that is, the proportion of fibroglandular to fatty tissue in the breast—is a strong hindrance to the detection of breast cancer due to the potential for lesions to hide in a high density background [64–66]. It is estimated that 26% of breast cancers in woman under the age of 55 are attributable to breast density over 50% (independent of other risk factors such as age) [67]. A widely used measure of breast density in digital mammography has been the proportion of the mammogram that is opaque, referred to as percent density (PD). In area-based PD, opacity is judged by simple thresholding of image pixel values, while volume-based PD (Volpara) also considers the thickness of the dense tissue by making use of the unthresholded pixel values. Opinions differ as to which approach, area or volume-based PD, is the better. Shepherd [68] concluded that volumetric PD methods are better predictors of breast cancer risk than area-based PD, while others have concluded the opposite [69,70].

While PD is an important risk factor, there is growing evidence to suggest that texture characteristics, which are not necessarily correlated with PD, may also be an indicator [71]. Indeed, a key recent result from the University of Manchester's PROCAS trial, set up with the aim of predicting patient risk at screening, was that texture features could in fact be a more powerful risk indicator than PD methods [72]. Similar conclusions have been drawn previously by others [73–75]. The recognition of texture characteristics as an important risk indicator has led to the adoption of classification systems that classify breasts not only according to the proportion of fibroglandular tissue, but also its distribution (e.g., "scattered" or "heterogeneously dense"). Two such systems are the BI-RADS density scale (Fig. 14.2) and the parenchymal pattern (PP) scale (Fig. 14.3) developed by Professor Tabar [77].

Whether breast cancer risk is best assessed by PD or texture-based approaches, or a combination of the two, is still a matter of ongoing research. What is clear is that a major downside of these methods is that they rely on rigid, hand-crafted features based on simple pixel intensities (and in the case of texture, the gray-level co-occurrence matrix of neighboring pixels or Gaussian features [78]). This introduces a lot of human bias, as well as trial and error to identify which features are the most effective.

With deep learning approaches, instead of needing to decide on a set of features to use a priori, the most salient features to use are learned directly from the data, and are specifically tailored to the task at hand. In addition, the features

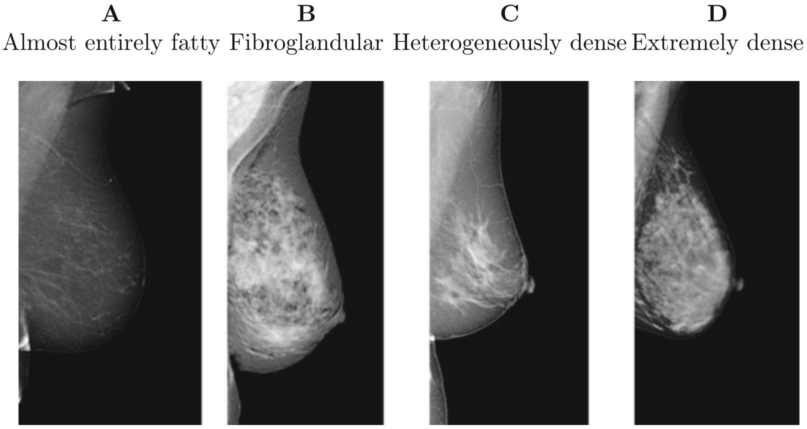


Fig. 14.2 BI-RADS density scale. Reproduced with kind permission from the American College of Radiology [76]

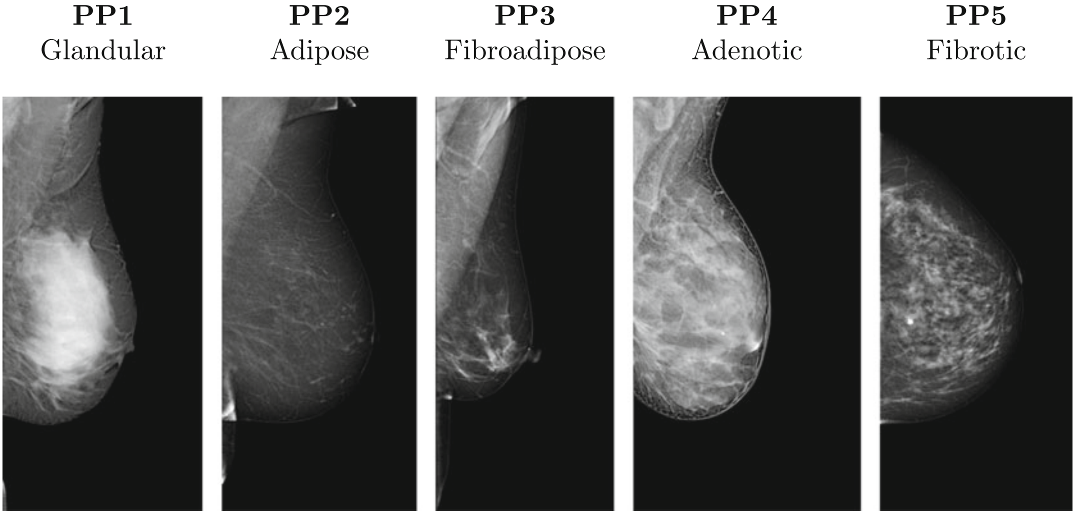


Fig. 14.3 Tabar Parenchymal Patterns (PP). Reproduced with kind permission from Professor Tabar [77]

learned by deep learning algorithms are significantly richer than the crude PD or GCLM-based features. In particular, they are highly specialized for discriminating between different patterns and textures. It is unsurprising therefore that deep learning algorithms are already being applied to great effect in density and PP estimation. For example, Wu et al. [79] trained a CNN to classify breasts according to the BI-RADS density scale.

O'Neill (Kheiron Medical Technologies) presented at RSNA 2017 on how a CNN-based model could be applied equally effectively to BI-RADS and PP classification, and that best results were obtained by jointly training for both classification tasks at once (Fig. 14.4). Both works reported human-level accuracies on these tasks compared with a consensus of radiologists.

These results highlight the potential of deep learning to improve risk assessment in breast cancer screening based on tissue density. Not only are deep neural networks highly adept at learning the types of feature used in traditional PD and texture-based approaches, they also have the flexibility to learn any other (perhaps more subtle) risk indicators present in the mammogram. They offer the potential therefore, of a single unified approach to breast cancer risk assessment that is both consistent and accurate.

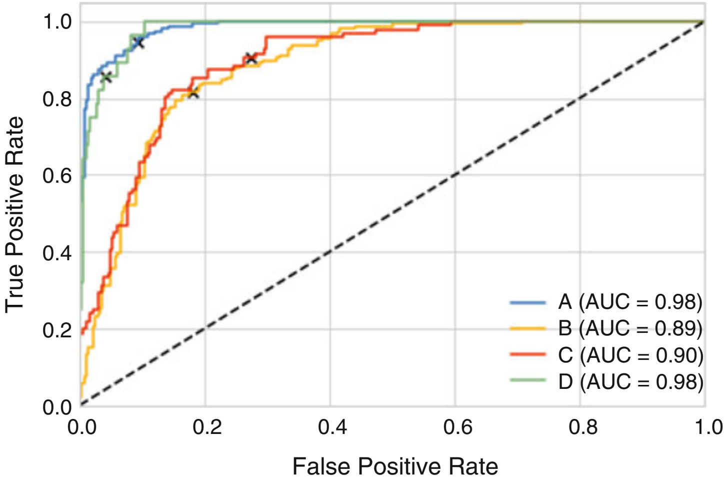


Fig. 14.4 ROC curves for the four density classes, taking ground truth as the multi-center radiologist consensus. The model was trained jointly with BI-RADS and PP labels. The lower AUC for classes B and C is likely due to noisy labels—these classes are the hardest for radiologists to distinguish between

14.3 Deep Learning Challenges Specific to Mammography

14.3.1 Memory Constraints and Image Size

The majority of deep neural networks for image perception tasks in the public domain were designed for images with a maximum size of 299 × 299. This is vastly different to FFDM, which are orders of magnitude greater in height, width, and total pixel count. This increase in image size comes at a hefty design cost when developing such algorithms for mammography. As the amount of RAM available on the majority of high-end GPUs is currently 12 GB, one full resolution image is too large to train on with any CNN architecture that has yielded the state-of-the-art results on the ImageNet challenge over the past 6 years. This problem has traditionally been tackled by down-sampling the image to a smaller resolution, or splitting the image up into smaller constituent patches. Both approaches are highlighted in Fig. 14.5.

Patch-based approaches are by far the most common way to overcome the intractable computational requirements of full resolution digital mammograms. The limitations of these have been discussed in Sect. 14.2.2.

Another approach to solving issues with large image sizes is to down-sample the image to a size which is possible to train on. The positives of this method include preserving the contextual information contained in the image, and not requiring localized pixel-level labels for classification tasks. As long as the down-sampling does not destroy the presence of important visual features, it is the most likely route to successful classification. However, it is not without drawbacks. As mentioned previously, most successful CNN architectures were designed on much smaller image sizes. Researchers must be consider whether the maximum receptive fields (the maximum region of the image which contributes to a neurons activation in the final convolutional layer of a CNN) of these architectures is large enough to span the necessary contextual information. If this receptive field is too small (or too large), a significant amount of network redesign may be necessary, which is particularly difficult in the case of more sophisticated architectures (such as inception and resnets). Also, much smaller batch sizes are possible when training on down-sampled images, which may affect training.

There are other, more sophisticated, ways to get around the trade-off between network depth, image, and batch size. These include advanced methods which look explicitly at the execution of training with deep learning models, with check-pointed memory management. The intermediate results are usually required for the commonly used back propagation method of updating network weights [81]. Another method introduced

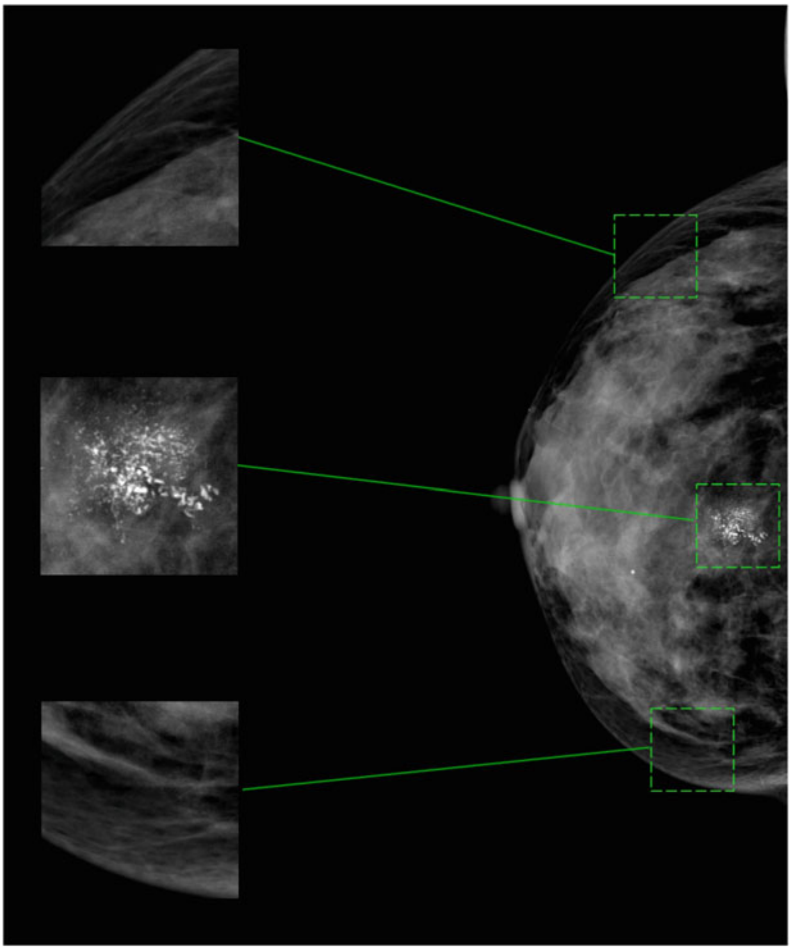


Fig. 14.5 A comparison between downscaling a full mammogram or cropping full resolution patches. If the proportion of down-sampling is too high, important visual features may be lost. Reproduced with kind permission from [80]

(and getting a lot of attention recently) are reversible networks [82], where one can compute the gradient without storing a majority of the network's activations, decreasing memory requirements by replacing activations with simple mathematical operations.

Looking towards the future, as more advanced techniques emerge, and more powerful GPUs are designed, researchers will still grapple with similar memory limitations. A variety of exciting avenues are yet to be explored, including the possibility of considering a patients entire screening history, or genetic information, when making a recall decision. Thus design choices similar to those described above will need to be considered, and such choices are a key factor in determining a models success.

14.3.2 Data Access and Quality

Researchers have access to several public and restricted image databases. However, quantity, quality, and availability of metadata and clinical data vary a lot between those datasets. For example, scanned hard copy films may not be useful for developing state-of-the-art digital mammography algorithms. One of the more popular databases is DDSM which is available to the general public containing more than 10,000 images. Unfortunately, the quality of the digitized films does not match that of FFDM [51] and the provided annotations are not as accurate as they should be for training machine learning systems (e.g., 339 images contain annotations although the masses are not clearly visible [83]). An up-to-date and better curated version of DDSM was published more recently [83]. At the time of writing, only one group has published work on the new release of DDSM [84]. The second most frequently cited database is MIAS, however, compared to DDSM it lacks samples. Furthermore, offering only 8-bit images is no longer state of the art, therefore we can only assume that this dataset will not be useful for future deep learning projects. The InBreast dataset is also often used as a benchmark as it consists of annotated FFDM images. However, with 115 cases it is rather small, cannot be considered representative of real-world inputs, and is not suitable to assess the performance of algorithms in real-world settings.

There are many other mammography datasets, with varying volume and quality. Table 14.2 summarizes the most popular of these publicly available data sources.

14.3.3 Data Issues During Training

It is very rarely possible to collect a dataset that is perfectly balanced with respect to different classes and features, completely unbiased and plentiful. Even where this is possible, it can be very expensive and time-consuming. More often than not, researchers need to carefully consider the imperfections of the datasets in order to achieve the desired results.

14.3.3.1 Dataset Imbalance

It is frequently the case in medical imaging that the class we are most interested in accurately predicting is also the least frequent one. For example, the prevalence of breast cancer in a screening population is between 0.6 and 1.0%. Assuming a dataset consists of standard views (CC and MLO) for each breast and that observing malignancies in both sides is relatively rare, it is possible that as many as 99.7% of the images will be benign. Naturally, developers wish to take advantage of all the available images, but severe class imbalance causes problems during model training. Some of the main issues can be identified as follows.

Table 14.2 Commonly used mammography datasets for deep learning

Name	Origin	Year	No. of cases	No. of images	Access
MIAS	UK	1994	161	322	Public
OPTIMAM	UK	2008	9559	154,078	On request
DDSM and CBIS-DDSM	USA	1999	2620	10,480	Public
Nijmegen	Netherlands	1998	21	40	On request
Trueta	Spain	2008	89	320	On request
IRMA	Germany	2008	Unknown	10,509	Public
MIRAcle	Greece	2009	196	204	Unknown
LLNL	USA	Unknown	50	198	Cost
Malaga	Spain	Unknown	35	Unknown	Unknown
NDMA	USA	Unknown	Unknown	1,000,000	On request
BancoWeb	Brazil	2010	320	1400	Public
Inbreast	Portugal	2012	115	410	On request
BCDR-F0X	Portugal	2012	1010	3703	On request
BCDR-D0X	Portugal	2012	724	3612	On request
SNUBH	Korea	2015	Unknown	49	Public

Insufficient Data in Minority Class

The data points in the minority class may be insufficient for training a model with the desired capacity. The straightforward solution is to simply acquire more data. If that is not an option, we may resort to transforming our existing data in a plausible way (i.e., by adding some noise or image rotation) or generating realistic synthetic data. Needless to say, none of the alternatives can be a perfect substitute for high-quality real data.

Account for Class Imbalance During Training

It is necessary to account for class imbalance, otherwise training may fail completely. During supervised learning, at every step of the iterative training process, one can tweak the model parameters towards minimizing a loss function indicative of the model performance. In most cases, this function needs to be differentiable. Examples of loss functions are the error rate, entropy, or mean square error. These metrics are adversely affected by data imbalance. Consider the case of a model always predicting that a breast screening case is benign. This model will be correct, at least, 99% of the time in a screening population, but of course such a model would be of no practical use. There are some simple solutions for re-balancing prevalence, including over- and under-sampling of the classes, as well as differently transforming classes.

Possible solutions to these problems include (either individually or in combination):

- Oversampling the minority class
- Transforming the minority class in a plausible manner [85]
- Under-sampling the majority class.

14.3.3.2 Dataset Bias

Detecting biases in the datasets is crucial for any machine learning method, as any discrepancy between the training data and reality will most certainly be reflected in the model performance. Imagine a scenario in which most of the benign cases we have in our possession come from one type of scanner and most of the malignant cases from another. This undesired pattern can be easily picked up and the resulting model may learn to discriminate between the two scanners, rather than detecting malignancies. Another example is if the training dataset comes from mostly *symptomatic* cases, but we wish to use the resulting model for breast *screening* instead. The two clinical settings are quite different and the trained model may perform poorly when applied to a different clinical setting than the data was acquired from.

14.3.3.3 Under-Fitting, Overfitting, and Generalization

Under-fitting is when both training and validation loss are below their minimum value. This can be caused either by using a model too simple for the task (underspecified) or by not training the model long enough.

Over-fitting is a bit more subtle and occurs when the validation loss is higher than its optimal value. In this case, the model fits well on the training data and the training loss is low, but fails to adequately generalize that knowledge to unseen data. It is usually difficult to diagnose or fix. Over-fitting may be caused when:

- Using a model more complex than necessary, in combination with the finite amount of training data.
- Training for too long and presenting the same training data multiple times to the model.

The effects of both under-fitting and over-fitting are illustrated in Fig. 14.6.

When it comes to mammography and the publicly available datasets, it is important to ascertain the underlying biases in the data, the proportion of cases from different manufactures, and the clinical setting the data was derived from. Over-fitting to one specific hardware manufacturer is a real risk with a non-heterogeneous dataset, as is overfitting a model to a symptomatic cohort of patients. The latter usually present with much larger and more obvious changes on their imaging. More subtle malignancies, as usually found in a screening setting, may be overlooked by an overfitted model to symptomatic patient data.

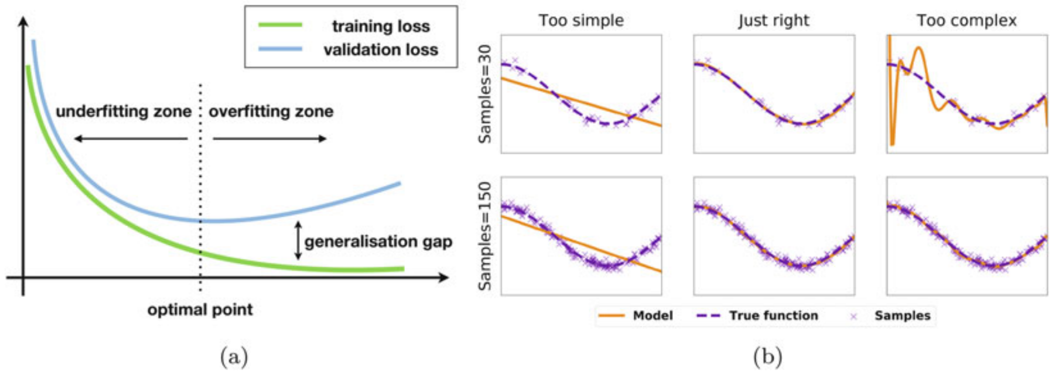


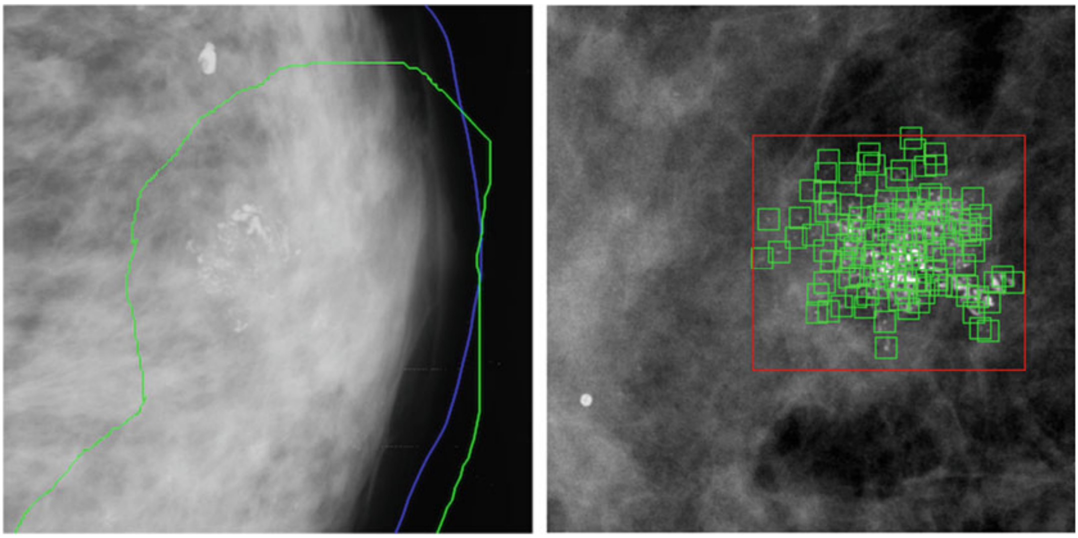
Fig. 14.6 (**a**) Illustration of training, validation loss, areas of under-fitting and over-fitting. The effect is similar with the x-axis representing model capacity or number of iterations. (**b**) The effect of model capacity with the amount of training data

Finally, in order for a deep learning algorithm to be useful in a wide variety of clinical practice, care must be taken to ensure generalizability of the training data as much as possible, to allow for a more generalizable end result model.

14.3.4 Data Labeling

A key ingredient in the success of any machine learning algorithm is how well a dataset is labeled. In real-world applications at the current level of the field at the time of writing this book, no amount of sophisticated research techniques, architectures, and computational resources can make up for poorly labeled data. This general principle is colloquially described as "garbage in, garbage out" within the community. This is an even greater challenge in medical imaging, as a "ground truth" can be difficult to establish due to significant levels of inter and intra-radiologist variability, coupled with sometimes obscure definition metrics. A perfect example of this is breast density. Radiologists disagree not only among each other, but also with themselves, i.e., might provide widely differing opinion at different repeat assessment [86, 87]. These inconsistencies inevitably lead to noisier target labels, which can make naive training a challenge. However, a good consensus requires multiple radiologists to label each image, which quickly becomes prohibitively expensive. A similar problem exists for the BIRADS scheme for malignancy assessment. The subjective differences between a case being labeled BIRADS 4 (a 30% PPV for malignancy) and BIRADS 5 (a 95% PPV for malignancy) again lead to significant variability among readers (especially when considering difficult features such as architectural distortions and asymmetries) [88]. The strongest possible marker for malignancy in this case is to use biopsy-proven follow-up results. However, this data may not always be available.

The problem is further exacerbated if pixel-level labeling is required (as is the case for patch-based and segmentation approaches). This can be especially difficult when combining datasets from various sources, especially those in the public domain. Figure 14.7a highlights a badly annotated calcification in the publicly available DDSM database; here the annotation passes outside of the breast region, and has large areas where there are no calcifications present. Another problem arises in how specific anomalies were labeled; this is particularly tricky for calcifications, where precise hand annotation would be prohibitively time-consuming, whereas coarser region of interest annotations may suffer from a low level signal to noise ratio. Figure 14.7b highlights the ideal scenario where each individual microcalcification is labeled, as well as the overall cluster.



(a) A badly annotated example [89]

(b) A well annotated example [80]

Fig. 14.7 Images from various mammography databases. (**a**) The blue contour highlights the breast edge, and the green contours are the lesion annotations. (**b**) The green boxes are individually labeled calcifications, the red box is the coarser cluster annotation

14.3.5 Principled Uncertainties

The most widely used deep learning models have an important shortcoming: they lack an underlying mechanism to provide uncertainty information about the predictions they make. Instead they often output point estimates between 0 and 1, which are often taken blindly as a measure of confidence. There have already been a few high profile cases where blindly trusting decisions made by deep learning algorithms has had disastrous consequences. For example in 2016, and then again in 2018, there were fatalities due to mistakes made by the perception system of autonomous vehicles. In health care a wrong decision can be a matter of life or death, and so being able to place trust in the decisions of deep learning models applied to such an industry is of critical importance.

Normally, a statistical model would output an entire predictive distribution as opposed to a point estimate. The spread of this distribution would tell us how confident a model is in its prediction, and consequently, how much we should trust it: a narrow spread of values would indicate a high confidence, a broad spread the opposite. However, obtaining the exact predictive distribution of a deep learning model is an intractable problem due to their size and complexity. There have been significant recent attempts to approximate the predictive distribution [90, 91], but solving this problem is still an active area of research.

If the outputs of a deep learning model could be calibrated to a meaningful scale, such as the probability of malignancy, then we would be safe in interpreting them as a measure of confidence: an output of 0.7 in a binary malignant-or-not mammography classification task would mean 70% chance of the scan containing a cancer, allowing a well-reasoned recall decision to made. Unfortunately the outputs of deep learning algorithms are notoriously un-calibrated [92] (Fig. 14.8); improving this calibration is once again an active area of research [92, 93].

14.3.6 Interpretability

With deep learning being used more widely in practical applications, there has been a lot of scrutiny on the underlying algorithms and how

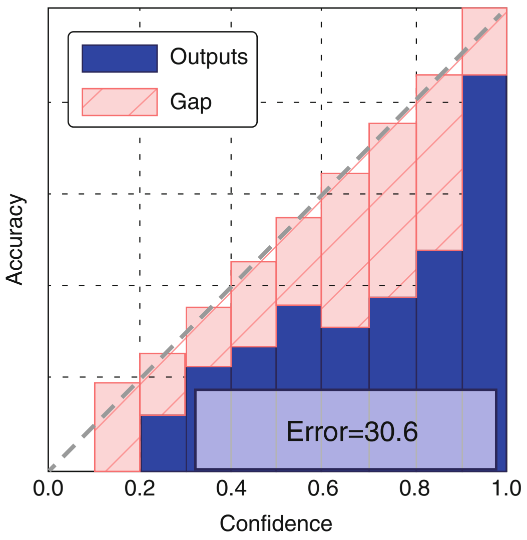


Fig. 14.8 Calibration plots for 110-layer ResNet on CIFAR-100. Confidence is the output of the network. The heights of the blue bars give the actual accuracy achieved by thresholding at the corresponding confidence. The shaded red bars show the discrepancy between the ResNet and a perfectly calibrated model; the ResNet is overconfident in its predictions. Applied to breast screening this may correspond to an excessively high recall rate. Reproduced with kind permission from [92]

automated decisions are reached. Regulatory bodies in the USA, EU, and other countries have already set some requirements for explainability of automated decisions. Researchers proposing deep learning methods for breast image analysis are also making efforts to achieve some level of interpretability of the proposed algorithms [54]. Arguably, this is both due to the anticipation that interpretability will be a requirement for regulatory approval and also for reassuring users (doctors and patients) that the algorithms perform as intended. However, some have argued that CAD systems should not give any localizing information to radiologists at all, and instead simply allow them to review any images deemed suspicious by the system and marked for recall [94]. This allows for a direct “machine read” output to be plugged into existing systems as an independent second or third reader, and completely avoids introducing any anchoring bias to the human-reading process.

Theoretical Perspective From a purely theoretical standpoint, a deep neural network has a deterministic and fully interpretable behavior. It is deterministic, in the sense that the same input will result in the same output every time. It is fully interpretable, in the sense that we can trace the final decision back to each activated neuron and all activations can be traced back to the pixels that contributed. We can go even further and visualize the areas of the image that contribute more in the decision, as presented in [95]. However, this mode of interpretation does not necessarily make sense from a human perspective.

Necessity for Interpretability The main question is whether interpretability is really something necessary to have. There are undoubtedly cases where it adds value to a system and others when it arguably does not. Let us assume we wish to build system to assist junior breast radiologists during their training. In that case, being able to explain why the system flagged a malignancy in an image is very valuable. However, assuming we wish to deploy a deep neural network for automated breast screening in a country with no breast screening program—would interpretability add any value in that case? More importantly, if faced with the decision, should we choose an interpretable system with lower sensitivity over a “black-box” one?

Interpretability Through Supervision Even in cases when interpretability cannot be naturally achieved, it can still be learned. For instance, networks can be trained to attend to and base their decisions on the regions of a mammogram that a human radiologist has indicated as important. We can even train networks to generate textual explanations of their decision, learning that skill from radiologists’ reports.

There are, however, a several issues associated with doing this:

- The task to be learned becomes significantly more complex and difficult to learn.
- Annotations become even more costly and time-consuming.

- There may be inconsistencies between annotators, due to the subjective nature of the task.

Finally, there is no guarantee whatsoever that the network trained to explain itself will outperform the one trained on a much simpler task, and the chances are that it will not.

14.4 Future Directions

14.4.1 Generative Adversarial Networks (GANs)

The ability to train generative models that are able to synthesize images indistinguishable from real ones is the ultimate modeling objective. It implies that the underlying mechanism or distribution responsible for generating the observed images has been successfully captured. Of course, capturing the mechanism responsible for producing gigapixel breast imaging data is very difficult. Nevertheless, it is convenient that generative models do not necessarily require labeled data to train.

A framework for training generative models in an adversarial manner was introduced by the seminal paper of Goodfellow et al. [96] and has signified a leap forward towards effectively training models for image data generation with high fidelity. It is based on a simple but powerful idea that a generator neural network is trained to produce realistic examples and the discriminator is trained to be able to discern between real and fake ones (a “critic”). The two networks form an adversarial relationship and gradually improve one-another through competition, much like two opponents in a game (Fig. 14.9).

GANs are not the only approach to generative models, but are arguably currently the most successful one. Alternative methods include:

- Variational auto-encoders (VAEs) [97]: Without adversarial loss, these methods use perceptual image similarity losses and tend to produce blurrier, less sharp images.
- Auto-regressive models (pixel RNNs) [98]: A recurrent neural network is an auto-regressive model that can be used to generate images sequentially, one pixel after the other. It has shown promise, but has not scaled so far for higher resolutions.

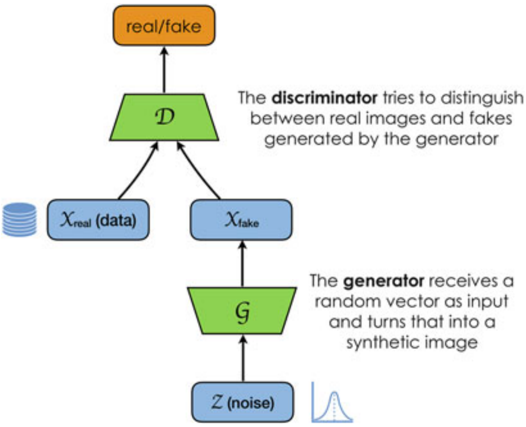

real/fake
$\mathcal{D}$
The **discriminator** tries to distinguish between real images and fakes generated by the generator
X_{real} (data)
X_{fake}
$\mathcal{G}$
The **generator** receives a random vector as input and turns that into a synthetic image
Z (noise)

Fig. 14.9 The discriminator is trained to distinguish between real and synthetic images. The generator attempts to produce realistic images indistinguishable by the discriminator. The two networks gradually improve one-another through this competition. Learning can only take place at the equilibrium between the two adversaries

Using adversarial training on the contrary has been demonstrated to generate sharp images. Training these models, however, is notoriously difficult due to severe instability that manifests itself when the equilibrium between generator and discriminator is lost. For that reason, a great deal of contemporary research is focused towards stabilizing the training and improving our theoretical understanding.

The significance of GANs goes far beyond generating realistic images of faces, furniture, or natural scenes and, as we previously mentioned, stems from modeling the underlying data distribution. There are several applications of high significance for the medical imaging community and, by extension, breast imaging. In Fig. 14.10 we provide examples of some early work on whole mammogram synthesis.

Synthetic Data Generation The expectation of GANs is that well-trained generative models could be used to synthesize an unlimited amount of high-quality images that can be used to

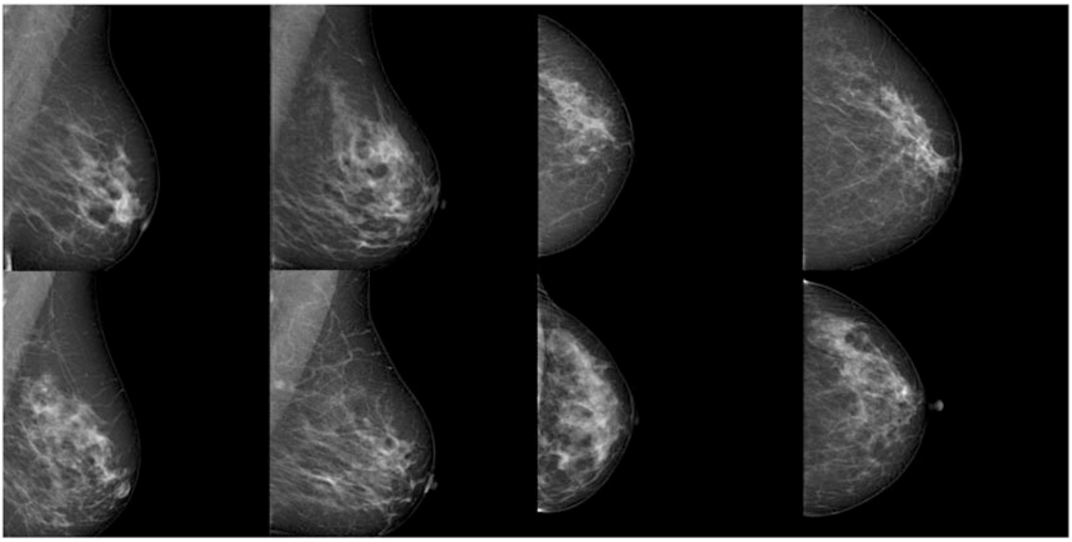


Fig. 14.10 One of the two rows of images in this figure consists of real mammograms and the other of synthetically generated ones. Can you tell which is which?

improve downstream detection and classification models.

Promising examples of using GAN-generated synthetic data are recently emerging in the literature. For instance, Salehinejad et al. [99] generated X-ray images and Costa et el. [100] retinal images with the accompanying vessel segmentation. Our group has also published early work on synthesis of high resolution mammograms [101] (Fig. 14.10). However, more evidence and clinical evaluation is required before there can be wide adoption of such methods.

Semi-supervised Learning Semi-supervised learning is a very effective way to leverage unlabeled data to increase model performance or reduce the requirement for labeled examples.

The concept is closely related to multi-task learning, where jointly modeling multiple tasks is beneficial to each individual task as well. In the semi-supervised case, an additional benefit is that at least one of the tasks learned does not require labeled data. The benefit could be superior in performance for the same amount of labeled data, or a more graceful degradation when reducing the amount of labeled data. This may be of particular use in mammography where small datasets are publicly available, but large amounts of labeled data aren't as readily accessible.

GANs have been particularly useful in semi-supervised learning and several studies have shown the aforementioned benefits in practice, both in modeling natural [102] and medical images [103].

Domain Adaptation A common problem in medical imaging is being able to transfer a trained model to a different modality, manufacturer, or other domain where labeled data are scarce or unavailable. Generative adversarial networks have been successfully used in medical imaging to do so. For example, Kamnitsas et al. [104] used GANs for domain transfer in brain CT semantic segmentation and Wolterink et al. [105] used them to transfer from low to regular-dose CT. Future work on tomosynthesis imaging (see Sect. 14.4.3) may benefit from the use of GANs for domain adaptation.

14.4.2 Active Learning and Regulation

Current regulatory processes do not allow for active learning using deep learning models. A "build and freeze" framework is the current standard, requiring developers to validate their model on a rigid dataset, report the results, and then apply for regulatory approval.

In the future it might well be possible that models implemented in hospitals make use of active learning, whereby networks continuously learn from new clinical data in a live setting. In a symbiotic system, a biopsied mass could be used to provide a data label, and once this label becomes available the image could be added to a liquid training dataset. A positive biopsy result would create a positive label, along with metadata including phenotyping and genomics of the malignancy subtype. A negative result would not be assigned a data label until a set amount of time has passed, for example 2 years, therefore giving more confidence to the negative label. This should allow for the system to continuously improve, especially with the prospect of having a global network learning from thousands of scans a day to help radiologists.

There are several barriers to overcome before a constantly learning system could be deployed; patient consent, validation of a model which is continuously updating, and overcoming variation between clinical sites based on their local data. However, it is up to the regulatory bodies to change their practice before symbiotic constantly learning systems may even be feasible.

14.4.3 Tomosynthesis

While the vast majority of screening programs across the globe currently employ 2D mammography, the rising use of digital breast tomosynthesis (DBT) likely signals the direction for the future of these programs. DBT is a tomography technique in which numerous low dose X-ray images are acquired in an arc around the compressed breast. A 3D reconstruction is formed from the various projections, in a similar fashion to CT and MRI scans. DBT is also capable of constructing a 2D synthetic image, by superimposing all of the slices in a manner that resembles a traditional 2D mammogram. The primary advantage of DBT over traditional mammography is its application to dense breast tissue. Particularly dense tissue is capable of obscuring the presence of certain types of lesions on 2D mammography scans. These lesions are more easily resolved when considering different angled slices from DBT. DBT thus enhances the morphological properties of abnormal tissue, which should yield significantly better detection rates, while delivering an X-ray dosage only slightly above that of conventional 2D mammography [106] (and well within recommended safety guidelines). The increase in resolution can also help guide biopsies by providing a more accurate target region.

However, as DBT is a relatively new technique (when considering levels of adoption), there is significantly less literature on it compared to traditional FFDM. Initial prospective studies showed either superiority, or non-inferiority when compared to FFDM, but these were all small in scale, and are summarized in the review by Vedantham et al. [107]. The most significant retrospective study was conducted by Gilbert et al. [108], with DBT showing moderate increases in performance to 2D mammography (AUC increased from 0.84 to 0.88), especially in the case of dense tissue (AUC increased from 0.83 to 0.87). The most significant criticisms of the modality come from a resource perspective, with significant infrastructure updates and training procedures required. As each scan consists of tens of slices, the time of reading also increases [109], thus hospitals which are already operating at capacity may not easily deal with the increased workload.

This increase in cognitive workload makes it a perfect candidate for assistance from machine learning algorithms. Past research on CAD use in DBT is sparse, with results reproducing the same limitations as 2D CAD systems—a prohibitively high number of false positives at adequate sensitivity levels. This has been shown in studies on masses [110], calcifications [111,112], and both [113]. Thus the development of the next generation of intelligent algorithms, capable of constructively aiding a radiologist, will be critical in facilitating DBT adoption, especially in countries where radiologists are already overworked. Unfortunately, the increase in overall workload from DBT cases also translates into extra engineering and labeling challenges. All of the memory constraints that apply to traditional four image 2D cases need now to account for cases that can have a multitude of images. Also, accurate pixel-level labels will be crucial with current techniques, which translates to costly annotation demands. Despite these hurdles, the possibility of adopting DBT as a standard for future screening programs promises an exciting future in which patient outcomes improve. Deep learning algorithms could play a critical role in realizing this future.

14.4.4 Genomics

The biggest challenge in fighting cancer is the heterogeneity of the disease. Progress on various scientific fields pushed further the understanding of the complex biological processes of invasive breast cancer. However, linking molecular data with radiological imaging data is not trivial. An interesting paper analyzing the correlation between breast cancer molecular subtypes and mammographic appearance was published by Killelea et al. [114]. Their retrospective analysis revealed characteristic associations between the appearance of the tumors on the mammographic image and the molecular profile. Architectural distortions were associated with luminal-type cancers whereas calcifications with or without mass are correlated with HER2-positive cancers. Triple negative cancer was found to be associated with a noncalcified mass. Those promising findings raise the question on what applying deep learning to paired molecular and mammographic data could reveal. Subtle features that are not captured in the high-level statistical analysis by Killelea et al. [114] may give new insights into breast cancer development and may reveal image-based biomarkers that could potentially replace expensive sequencing in the future. A novel way of analyzing the DNA using deep learning was published by Nguyen et al. [115]. The authors propose to use deep CNNs for DNA sequence analysis. They keep the sequential form of the input by sliding a window of a fixed size over the DNA sequence and encode the resulting words as binary 2D matrix. The results are promising but the authors have chosen their hyper-parameters empirically (word size, region size, network architecture) so further work is required to get a better understanding how those parameters influence the analysis. Yin et al. [116] take an image representation of the DNA sequence as input to a CNN and predict key determinants of chromatin structure. Their approach is able to detect interactions between distal elements in the DNA sequence as well as the presence and absence of splice-junctions. Compared to Nguyen et al. [115], the authors added residual connections to reuse the learned features as well as larger convolution filters.

A popular term in literature is "*radiogenomics*" [117] which refers to the relationship between imaging phenotypes and tumor genetics by image-based surrogates for genetic testing [118]. Commercial genetic tests, such as OncotypeDx (Genomic Health Inc., San Francisco), which is used to predict recurrence and therapeutic response, are currently being explored with respect to radiogenomic associations. However, the majority of publications still use traditional machine learning with hand-crafted features to find associations between genetics and image-derived features [118, 119]. The promised future of radiogenomics will require linking massive mammography and genetic datasets, something that is yet to be achieved.

14.5 Summary

It is not surprising to see a flurry of deep learning activity in the mammography sector, especially in Europe, where several countries hold robust breast nationwide screening databases, with every mammogram result, biopsy, and surgical outcome linked to every screening event. Early research in deep learning has shown both sensitivity and specificity of these algorithms approaching that of single human readers. Over the next couple of years we will undoubtedly see deep learning algorithms entering into screening settings. This will of course necessitate robust clinical trials both retrospectively to benchmark performance and prospectively to ensure that algorithmic performance is maintained in a real-world clinical setting. The holy grail will be to prove conclusively that deep learning systems can accurately make recall decisions as well as, or better than, human double-reading, while providing highly explainable and interpretable results when needed. However, radiologists are unlikely to hand over the reigns just yet, and may instead prefer single-reader programs supported by deep learning, effectively halving the workload for the already overstretched double-reading radiologists.

Deep learning technology could also potentially improve consistency and accuracy of existing single-reader programs, such as those in the USA, as well as provide an immense new resource to countries yet to implement a screening programme at all. The potential for deep learning-support in national screening, as well as for underdeveloped health care systems to leapfrog into the deep learning era, may therefore be just a few years away.

It is interesting to note that despite the advances of traditional CAD, the European guidelines for quality assurance in breast cancer screening and diagnosis (and their following supplements) do *not* address evaluation of processing algorithms and CAD [120]. There are however consolidated standards focusing on different topics to ensure the technical image quality of mammograms used for screening and assessment is sufficient to achieve the objectives of cancer detection. Perhaps, with the advent of deep learning these guidelines will eventually be updated to include CAD usage, especially if deep learning systems are eventually proven to demonstrate stand-alone sensitivity and specificity above that of single human readers, while simultaneously reducing recall rates and reducing the occurrence of interval cancers (cancers that present in between screening intervals).

To reach this goal, several hurdles must be overcome. First, larger more accurately labeled datasets are required, both for algorithmic training and validation. GANs may hold the potential to unlock vast amounts of synthetic training data, although their performance at present is not sufficient to provide robust comparison against real-world data. DBT may also herald a new source of "big data," simply by providing more images per case. It is certainly within the sights of researchers to utilize domain adaptation techniques to apply 2D mammography algorithms to 3D datasets. Finally, the era of radiogenomics, much anticipated but limited by data availability at scale, will only come of age once genomic testing in breast cancer becomes standard practice.

14.6 Take Home Points

- Breast mammography (2D FFDM) is seen as a key modality ripe for deep learning.
- Deep learning is most likely to act as a second reader in screening programs.
- Deep learning has the potential to improve accuracy and consistency of screening programs.
- As for any medical imaging analysis, access to large labeled datasets remains a challenge.
- Generative adversarial networks may assist in data augmentation via image synthesis.
- 3D tomosynthesis and radiogenomics are the next area of research for deep learning tools.

References

1. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JWW, Comber H, Forman D, Bray F. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. Eur J Cancer. 2013;49(6):1374–1403.
2. Tabár L, Gad A, Holmberg LH, Ljungquist U, Fagerberg CJG, Baldetorp L, Gröntoft O, Lundström B, Månson JC, Eklund G, Day NE, Pettersson F. Reduction in mortality from breast cancer after mass screening with mammography: randomised trial from the breast cancer screening working group of the Swedish National Board of Health and Welfare. Lancet. 1985;325(8433):829–32.
3. Lee CH, David Dershaw D, Kopans D, Evans P, Monsees B, Monticciolo D, James Brenner R, Bassett L, Berg W, Feig S, Hendrick E, Mendelson E, D'Orsi C, Sickles E, Burhenne LW. Breast

cancer screening with imaging: recommendations from the society of breast imaging and the ACR on the use of mammography, breast MRI, breast ultrasound, and other technologies for the detection of clinically occult breast cancer. J Am Coll Radiol. 2010;7(1):18–27.

4. Boyer B, Balleyguier C, Granat O, Pharaboz C. CAD in questions/answers: review of the literature. Eur J Radiol. 2009;69(1):24–33.
5. Duijm LEM, Louwman MWJ, Groenewoud JH, Van De Poll-Franse LV, Fracheboud J, Coebergh JW. Inter-observer variability in mammography screening and effect of type and number of readers on screening outcome. Br J Cancer. 2009;100(6): 901–7.
6. Dinitto P, Logan-young W, Bonaccio E, Zuley ML, Willison KM. Breast imaging can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience 1. Radiology. 2004;232(2):578–84.
7. Beam CA, Sullivan DC, Layde PM. Effect of human variability on independent double reading in screening mammography. Acad Radiol. 1996;3(11): 891–7.
8. Tice JA, Kerlikowske K. Screening and prevention of breast cancer in primary care. Prim Care. 2009;36(3):533–58.
9. Fletcher SW. Breast cancer screening: a 35-year perspective. Epidemiol Rev. 2011;33(1):165–75.
10. Hofvind S, Geller BM, Skelly J, Vacek PM. Sensitivity and specificity of mammographic screening as practised in Vermont and Norway. Br J Radiol. 2012;85(1020):e1226–32.
11. Domingo L, Hofvind S, Hubbard RA, Román M, Benkeser D, Sala M, Castells X. Cross-national comparison of screening mammography accuracy measures in U.S., Norway, and Spain. Eur Radiol. 2016;26(8):2520–8.
12. Langreth R. Too many mammograms. Forbes; 2009.
13. Taylor P, Champness J, Given-Wilson R, Johnston K, Potts H. Impact of computer-aided detection prompts on the sensitivity and specificity of screening mammography. Health Technol Assess. 2005;9(6):iii, 1–58.
14. Philpotts LE. Can computer-aided detection be detrimental to mammographic interpretation? Radiology. 2009;253(1):17–22.
15. Gilbert FJ, Astley SM, Gillan MGC, Agbaje OF, Wallis MG, James J, Boggis CRM, Duffy SW. Single reading with computer-aided detection for screening mammography. N Engl J Med. 2008;359(16):1675–84.
16. Gilbert FJ, Astley SM, Gillan MG, Agbaje OF, Wallis MG, James J, Boggis CR, Duffy SW. CADET II: a prospective trial of computer-aided detection (CAD) in the UK Breast Screening Programme. J Clin Oncol. 2008;26(15 suppl):508.
17. Taylor P, Potts HWW. Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. Eur J Cancer. 2008;44(6):798–807.
18. Noble M, Bruening W, Uhl S, Schoelles K. Computer-aided detection mammography for breast cancer screening: systematic review and meta-analysis. Arch Gynecol Obstet. 2009;279(6):881–90.
19. Karssemeijer N, Bluekens AM, Beijerinck D, Deurenberg JJ, Beekman M, Visser R, van Engen R, Bartels-Kortland A, Broeders MJ. Breast cancer screening results 5 years after introduction of digital mammography in a population-based screening program. Radiology. 2009;253(2):353–8.
20. Destounis S, Hanson S, Morgan R, Murphy P, Somerville P, Seifert P, Andolina V, Arieno A, Skolny M, Logan-Young W. Computer-aided detection of breast carcinoma in standard mammographic projections with digital mammography. Int J Comput Assist Radiol Surg. 2009;4(4):331–6.
21. van den Biggelaar FJHM, Kessels AGH, Van Engelshoven JMA, Flobbe K. Strategies for digital mammography interpretation in a clinical patient population. Int J Cancer. 2009;125(12):2923–9.
22. Sohns C, Angic B, Sossalla S, Konietschke F, Obenauer S. Computer-assisted diagnosis in full-field digital mammography-results in dependence of readers experiences. Breast J. 2010;16(5):490–7.
23. Murakami R, Kumita S, Tani H, Yoshida T, Sugizaki K, Kuwako T, Kiriyama T, Hakozaki K, Okazaki E, Yanagihara K, Iida S, Haga S, Tsuchiya S. Detection of breast cancer with a computer-aided detection applied to full-field digital mammography. J Digit Imaging. 2013;26(4):768–73.
24. Cole EB, Zhang Z, Marques HS, Edward Hendrick R, Yaffe MJ, Pisano ED. Impact of computer-aided detection systems on radiologist accuracy with digital mammography. Am J Roentgenol. 2014;203(4):909–16.
25. Bargalló X, Santamaría G, Del Amo M, Arguis P, Ríos J, Grau J, Burrel M, Cores E, Velasco M. Single reading with computer-aided detection performed by selected radiologists in a breast cancer screening program. Eur J Radiol. 2014;83(11):2019–23.
26. Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med. 2015;175(11):1828.
27. Berry DA. Computer-assisted detection and screening mammography: where's the beef? J Natl Cancer Inst. 2011;103(15):1139–41.
28. Sanchez Gómez S, Torres Tabanera M, Vega Bolivar A, Sainz Miranda M, Baroja Mazo A, Ruiz Diaz M, Martinez Miravete P, Lag Asturiano E, Muñoz

Cacho P, Delgado Macias T. Impact of a CAD system in a screen-film mammography screening program: a prospective study. Eur J Radiol. 2011;80(3):e317–21.

29. Freer TW, Ulissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. Radiology. 2001;220(3):781–6.
30. The JS, Schilling KJ, Hoffmeister JW, Friedmann E, McGinnis R, Holcomb RG. Detection of breast cancer with full-field digital mammography and computer-aided detection. Am J Roentgenol. 2009;192(2):337–40.
31. Rao VM, Levin DC, Parker L, Cavanaugh B, Frangos AJ, Sunshine JH. How widely is computer-aided detection used in screening and diagnostic mammography? J Am Coll Radiol. 2010;7(10):802–5.
32. Onega T, Aiello Bowles EJ, Miglioretti DL, Carney PA, Geller BM, Yankaskas BC, Kerlikowske K, Sickles EA, Elmore JG. Radiologists' perceptions of computer aided detection versus double reading for mammography interpretation. Acad Radiol. 2010;17(10):1217–26.
33. Kohli A, Jha S. Why CAD failed in mammography. J Am Coll Radiol. 2018;15(3 Pt B):535–7.
34. Lehman CD, Arao RF, Sprague BL, Lee JM, Buist DSM, Kerlikowske K, Henderson LM, Onega T, Tosteson ANA, Rauscher GH, Miglioretti DL. National performance benchmarks for modern screening digital mammography: update from the breast cancer surveillance consortium. Radiology. 2017;283(1):49–58.
35. Carney PA, Sickles EA, Monsees BS, Bassett LW, James Brenner R, Feig SA, Smith RA, Rosenberg RD, Andrew Bogart T, Browning S, Barry JW, Kelly MM, Tran KA, Miglioretti DL. Identifying minimally acceptable interpretive performance criteria for screening mammography. Radiology. 2010;255(2):354–61.
36. Miglioretti DL, Ichikawa L, Smith RA, Bassett LW, Feig SA, Monsees B, Parikh JR, Rosenberg RD, Sickles EA, Carney PA. Criteria for identifying radiologists with acceptable screening mammography interpretive performance on basis of multiple performance measures. Am J Roentgenol. 2015;204(4):W486–91.
37. Myers ER, Moorman P, Gierisch JM, Havrilesky LJ, Grimm LJ, Ghate S, Davidson B, Mongtomery RC, Crowley MJ, McCrory DC, Kendrick A, Sanders GD. Benefits and harms of breast cancer screening: a systematic review. J Am Med Assoc. 2015;314:1615–34.
38. Sahiner B, Chan HP, Petrick N, Wei D, Helvie MA, Adler DD, Goodsitt MM. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. IEEE Trans Med Imaging. 1996;15(5):598–610.
39. Dhungel N, Carneiro G, Bradley AP. Automated mass detection from mammograms using deep learning and random forest. In: International conference on digital image computing: techniques and applications; 2015. p. 1–8.
40. Ertosun MG, Rubin DL. Probabilistic visual search for masses within mammography images using deep learning. In: IEEE international conference on bioinformatics and biomedicine; 2015. p. 1310–5.
41. Carneiro G, Nascimento J, Bradley AP. Unregistered multiview mammogram analysis with pretrained deep learning models. In: Proceedings of the 18th international conference on medical image computing and computer-assisted intervention. Lecture notes in computer science. Vol 9351. Cham: Springer; 2015. p. 652–60.
42. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: toward a full-field digital mammographic database. Acad Radiol. 2012;19(2):236–48.
43. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The cancer imaging archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26(6):1045–57.
44. Kooi T, Litjens G, van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, den Heeten A, Karssemeijer N. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal. 2017;35:303–12.
45. Teare P, Fishman M, Benzaquen O, Toledano E, Elnekave E. Malignancy detection on mammography using dual deep convolutional neural networks and genetically discovered false color input enhancement. J Digit Imaging. 2017;30(4): 499–505.
46. Kim E-K, Kim H-E, Han K, Kang BJ, Sohn Y-M, Woo OH, Lee CW. Applying data-driven imaging biomarker in mammography for breast cancer screening: preliminary study. Sci Rep. 2018;8(1):2762.
47. Elter M, Horsch A. CADx of mammographic masses and clustered microcalcifications: a review. Med Phys. 2009;36(6):2052–68.
48. Breast screening: consolidated programme standards - GOV.UK; 2017.
49. Rothschild J, Lourenco AP, Mainiero MB. Screening mammography recall rate: does practice site matter? Radiology. 2013;269(2):348–53.
50. Sage Bionetworks. The Digital Mammography DREAM Challenge; 2016.
51. Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detecting and classifying lesions in mammograms with deep learning. Sci Rep. 2018;8(1):4165.
52. Dhungel N, Carneiro G, Bradley AP. The automated learning of deep features for breast mass classification from mammograms. In: Interna-

tional conference on medical image computing and computer-assisted intervention. Cham: Springer; 2016. p. 106–14.

53. Arevalo J, Gonzalez FA, Ramos-Pollan R, Oliveira JL, Lopez MAG. Convolutional neural networks for mammography mass lesion classification. In: IEEE Engineering in Medicine and Biology Society (EMBC). Washington: IEEE; 2015. p. 797–800.
54. Lévy D, Jain A. Breast mass classification from mammograms using deep convolutional neural networks; 2016. arxiv:1612.00542.
55. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell. 2018;40(4):834–48.
56. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks; 2016. arxiv:1506.01497.
57. Li Y, He K, Sun J. R-fcn: object detection via region-based fully convolutional networks. In: Advances in neural information processing systems; 2016.
58. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC. SSD: single shot multibox detector; 2016. arxiv:1512.02325.
59. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention – MICCAI 2015; 2015. p. 234–41.
60. Zhu W, Xiang X, Tran TD, Xie X. Adversarial deep structural networks for mammographic mass segmentation; 2017. arxiv:1612.05970.
61. de Moor T, Rodriguez-Ruiz A, Mérida AG, Mann R, Teuwen J. Automated soft tissue lesion detection and segmentation in digital mammography using a u-net deep learning network; 2018. arxiv:1802.06865.
62. Uijlings JRR, van de Sande KEA, Gevers T, Smeulders AWM. Selective search for object recognition. Int J Comput Vis. 2013;104(2):154–71.
63. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN; 2017. arxiv:1703.06870.
64. Assi V, Warwick J, Cuzick J, Duffy SW. Clinical and epidemiological issues in mammographic density. Nat Rev Clin Oncol. 2012;9(1):33–40.
65. Colin C, Schott-Pethelaz A-M. Mammographic density as a risk factor: to go out of a 30-year fog. Acta Radiol. 2017;58(6):NP1.
66. Colin C. Mammographic density: is there a public health significance linked to published relative risk data? Radiology. 2017;284(3):918–9.
67. Martin LJ, Melnichouk O, Guo H, Chiarelli AM, Hislop TG, Yaffe MJ, Minkin S, Hopper JL, Boyd NF. Family history, mammographic density, and risk of breast cancer. Cancer Epidemiol Biomarkers Prev. 2010;19(2):456–63.
68. Shepherd JA, Kerlikowske K, Ma L, Duewer F, Fan B, Wang J, Malkov S, Vittinghoff E, Cummings SR. Volume of mammographic density and risk of breast cancer. Cancer Epidemiol Biomarkers Prev. 2011;20(7):1473–82.
69. Boyd N, Martin L, Gunasekara A, Melnichouk O, Maudsley G, Peressotti C, Yaffe M, Minkin S. Mammographic density and breast cancer risk: evaluation of a novel method of measuring breast tissue volumes. Cancer Epidemiol Biomarkers Prev. 2009;18(6):1754–62.
70. Aitken Z, McCormack VA, Highnam RP, Martin L, Gunasekara A, Melnichouk O, Mawdsley G, Peressotti C, Yaffe M, Boyd NF, dos Santos Silva I. Screen-film mammographic density and breast cancer risk: a comparison of the volumetric standard mammogram form and the interactive threshold measurement methods. Cancer Epidemiol Biomarkers Prev. 2010;19(2):418–28.
71. Gastounioti A, Conant EF, Kontos D. Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment. Breast Cancer Res. 2016;18(1):91.
72. Astley SM, Harkness EF, Sergeant JC, Warwick J, Stavrinos P, Warren R, Wilson M, Beetles U, Gadde S, Lim Y, Jain A, Bundred S, Barr N, Reece V, Brentnall AR, Cuzick J, Howell T, Evans DG. A comparison of five methods of measuring mammographic density: a case-control study. Breast Cancer Res. 2018;20(1):10.
73. Manduca A, Carston MJ, Heine JJ, Scott CG, Pankratz VS, Brandt KR, Sellers TA, Vachon CM, Cerhan JR. Texture features from mammographic images and risk of breast cancer. Cancer Epidemiol Biomarkers Prev. 2009;18(3):837–45.
74. Li J, Szekely L, Eriksson L, Heddson B, Sundbom A, Czene K, Hall P, Humphreys K. High-throughput mammographic-density measurement: a tool for risk prediction of breast cancer. Breast Cancer Res. 2012;14(4):R114.
75. Häberle L, Wagner F, Fasching PA, Jud SM, Heusinger K, Loehberg CR, Hein A, Bayer CM, Hack CC, Lux MP, Binder K, Elter M, Münzenmayer C, Schulz-Wendtland R, Meier-Meitinger M, Adamietz BR, Uder M, Beckmann MW, Wittenberg T. Characterizing mammographic images by using generic texture features. Breast Cancer Res. 2012;14(2):R59.
76. Bott R. ACR BI-RADS atlas. In: Igarss 2014; 2014.
77. Gram IT, Funkhouser E, Tabár L. The Tabar classification of mammographic parenchymal patterns. Eur J Radiol. 1997;24:131–6.
78. Petersen K, Nielsen M, Diao P, Karssemeijer N, Lillholm M. Breast tissue segmentation and mammographic risk scoring using deep learning. In: International workshop on breast imaging. Lecture notes in computer science. Vol 8539. Cham: Springer; 2014. p. 88–94.

79. Wu N, Geras KJ, Shen Y, Su J, Gene Kim S, Kim E, Wolfson S, Moy L, Cho K. Breast density classification with deep convolutional neural networks; 2017. arxiv:1711.03674.
80. Shin SY, Lee S, Yun ID, Jung HY, Heo YS, Kim SM, Lee SM. A novel cascade classifier for automatic microcalcification detection. Public Libr Sci. 2015;10(12):e0143725.
81. Chen T, Xu B, Zhang C, Guestrin C. Training deep nets with sublinear memory cost; 2016. arxiv:1604.06174.
82. Gomez AN, Ren M, Urtasun R, Grosse RB. The reversible residual network: backpropagation without storing activations; 2017. arxiv:1707.04585.
83. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. Data descriptor: a curated mammography data set for use in computer-aided detection and diagnosis research. Sci Data. 2017;4:170177.
84. Xi P, Shu C, Goubran R. Abnormality detection in mammography using deep convolutional neural networks; 2018. arxiv:1803.01906.
85. Chawla N, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
86. Keller BM, Nathan DL, Gavenonis SC, Chen J, Conant EF, Kontos D. Reader variability in breast density estimation from full-field digital mammograms: the effect of image postprocessing on relative and absolute measures. Acad Radiol. 2013;20(5):560–8.
87. Redondo A, Comas M, Macià F, Ferrer F, Murta-Nascimento C, Maristany MT, Molins E, Sala M, Castells X. Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms. Br J Radiol. 2012;85(1019):1465–70.
88. Lee AY, Wisner DJ, Aminololama-Shakeri S, Arasu VA, Feig SA, Hargreaves J, Ojeda-Fournier H, Bassett LW, Wells CJ, De Guzman J, Flowers CI, Campbell JE, Elson SL, Retallack H, Joe BN. Interreader variability in the use of BI-RADS descriptors for suspicious findings on diagnostic mammography: a multi-institution study of 10 academic radiologists. Acad Radiol. 2017;24(1):60–6.
89. Heath M, Bowyer K, Kopans D, Kegelmeyer P, Moore R, Chang K, Munishkumaran S. Current status of the digital database for screening mammography. In: Digital mammography. Dordrecht: Springer; 1998. p. 457–60.
90. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning; 2015. arxiv:1506.02142.
91. Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision?; 2017. arxiv:1703.04977.
92. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks; 2017. arxiv:1706.04599.
93. Cobb AD, Roberts SJ, Gal Y. Loss-calibrated approximate inference in Bayesian neural networks; 2018. arxiv:1805.03901.
94. Nishikawa RM, Bae KT. Importance of better human-computer interaction in the era of deep learning: mammography computer-aided diagnosis as a use case. J Am Coll Radiol. 2018;15(1): 49–52.
95. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps; 2013. arxiv:1312.6034.
96. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Advances in neural information processing systems; 2014. p. 2672–80.
97. Kingma DP, Welling M. Auto-encoding variational Bayes. In: International conference on learning representations; 2014.
98. van den Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks. In: International conference on machine learning. Vol 48; 2016. p. 1747–56.
99. Salehinejad H, Valaee S, Dowdell T, Colak E, Barfett J. Generalization of deep neural networks for chest pathology classification in X-rays using generative adversarial networks. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2018.
100. Costa P, Galdran A, Meyer MI, Niemeijer M, Abramoff M, Mendonca AM, Campilho A. End-to-end adversarial retinal image synthesis. IEEE Trans Med Imaging. 2018;37(3):781–91.
101. Korkinof D, Rijken T, O'Neill M, Yearsley J, Harvey H, Glocker B. High-resolution mammogram synthesis using progressive generative adversarial networks; 2018. arxiv:1807.03401.
102. Adiwardana D, et al. Using generative models for semi-supervised learning. In: Medical image computing and computer-assisted intervention – MICCAI 2016; 2016. p. 106–14.
103. Lahiri A, Ayush K, Biswas PK, Mitra P. Generative adversarial learning for reducing manual annotation in semantic segmentation on large scale microscopy images: automated vessel segmentation in retinal fundus image as test case. In: IEEE Computer Society conference on computer vision and pattern recognition workshops, July 2017; 2017. p. 794–800.
104. Kamnitsas K, Baumgartner C, Ledig C, Newcombe V, Simpson J, Kane A, Menon D, Nori A, Criminisi A, Rueckert D, Glocker B. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Lecture notes in computer science. Vol 10265. Cham: Springer; 2017. p. 597–609.

105. Wolterink JM, Leiner T, Viergever MA, Isgum I. Generative adversarial networks for noise reduction in low-dose CT. IEEE Trans Med Imaging. 2017;36(12):2536–45.
106. Gennaro G, Bernardi D, Houssami N. Radiation dose with digital breast tomosynthesis compared to digital mammography: per-view analysis. Eur Radiol. 2018;28(2):573–81.
107. Vedantham S, Karellas A, Vijayaraghavan GR, Kopans DB. Digital breast tomosynthesis: state of the art. Radiology. 2015;277(3):663–84.
108. Gilbert FJ, Tucker L, Gillan MGC, Willsher P, Cooke J, Duncan KA, Michell MJ, Dobson HM, Lim YY, Suaris T, Astley SM, Morrish O, Young KC, Duffy SW. Accuracy of digital breast tomosynthesis for depicting breast cancer subgroups in a UK retrospective reading study (TOMMY trial). Radiology. 2015;277(3):697–706.
109. Connor SJ, Lim YY, Tate C, Entwistle H, Morris J, Whiteside S, Sergeant J, Wilson M, Beetles U, Boggis C, Gilbert F, Astley S. A comparison of reading times in full-field digital mammography and digital breast tomosynthesis. Breast Cancer Res. 2012;14(S1):P26.
110. Chan HP, Wei J, Zhang Y, Helvie MA, Moore RH, Sahiner B, Hadjiiski L, Kopans DB. Computer-aided detection of masses in digital tomosynthesis mammography: comparison of three approaches. Med Phys. 2008;35(9):4087–95.
111. Sahiner B, Chan HP, Hadjiiski LM, Helvie MA, Wei J, Zhou C, Lu Y. Computer-aided detection of clustered microcalcifications in digital breast tomosynthesis: a 3D approach. Med Phys. 2011;39(1):28–39.
112. Samala RK, Chan HP, Lu Y, Hadjiiski L, Wei J, Sahiner B, Helvie MA. Computer-aided detection of clustered microcalcifications in multiscale bilateral filtering regularized reconstructed digital breast tomosynthesis volume. Med Phys. 2014;41(2):021901.
113. Morra L, Sacchetto D, Durando M, Agliozzo S, Carbonaro LA, Delsanto S, Pesce B, Persano D, Mariscotti G, Marra V, Fonio P, Bert A. Breast cancer: computer-aided detection with digital breast tomosynthesis. Radiology. 2015;277(1): 56–63.
114. Killelea BK, Chagpar AB, Bishop J, Horowitz NR, Christy C, Tsangaris T, Raghu M, Lannin DR. Is there a correlation between breast cancer molecular subtype using receptors as surrogates and mammographic appearance? Ann Surg Oncol. 2013;20(10):3247–53.
115. Nguyen NG, Tran VA, Ngo DL, Phan D, Lumbanraja FR, Faisal MR, Abapihi B, Kubo M, Satou K. DNA sequence classification by convolutional neural network. J Biomed Sci Eng. 2016;9(9):280–6.
116. Yin B, Balvert M, Zambrano D, Sander M, Wiskunde C. An image representation based convolutional network for DNA classification; 2018. arxiv:1806.04931.
117. Rutman AM, Kuo MD. Radiogenomics: creating a link between molecular diagnostics and diagnostic imaging. Eur J Radiol. 2009;70(2):232–41.
118. Grimm LJ. Breast MRI radiogenomics: current status and research implications. J Magn Reson Imaging. 2016;43(6):1269–78.
119. Incoronato M, Aiello M, Infante T, Cavaliere C, Grimaldi AM, Mirabelli P, Monti S, Salvatore M. Radiogenomic analysis of oncological data: a technical survey. Int J Mol Sci. 2017;18(4):pii: E805.
120. Perry N. European guidelines for quality assurance in breast cancer screening and diagnosis. Ann Oncol. 2006;12(4):295–9.

15 Neurological Diseases

Nathaniel Swinburne and Andrei Holodny

15.1 Introduction

Neuroradiology has often been at the forefront of radiological imaging advances, such as the advent of diffusion-weighted MRI [1], due to the high stakes associated with diseases of the brain and spine as well as pragmatic factors such as the small field of view required for brain imaging and the sparing of the brain from respiratory motion artifact. With advances in computer vision in recent years, much interest has centered on the application of these technologies to neuroimaging; however, this presents a challenge due to the cross-sectional and, in the case of MRI, multiparametric nature of brain and spine imaging. The hardware demands associated with training deep learning networks using large numbers of three-dimensional image volumes are significant [2], although newer techniques [3] in combination with the availability of increasingly powerful GPU chips are beginning to overcome these challenges. AI applications to neuroimaging involve all aspects of image acquisition and interpretation and include study protocoling, image reconstruction, segmentation, and detection of disease processes (i.e., image classification).

15.2 Preprocessing of Brain Imaging

When utilizing supervised training for any task, the quality of the labeled training data has a profound impact on the success of the trained network. Accordingly, brain imaging data typically undergo several preprocessing steps before being utilized in AI applications. These steps include brain extraction (i.e., skull stripping), histogram normalization, and coregistration.

For many brain imaging AI applications, the removal of non-brain tissues from imaging data, including the skull, orbital contents, and soft tissues of the head and neck, leads to better performance [4–6]. The most commonly used tools for these tasks include the FMRIB Software Library (FSL) Brain Extraction Tool (BET) [7–9] and BET 2 [10], Brain Surface Extractor (BSE) [11], FreeSurfer [12], Robust Learning-based Brain Extraction System (ROBEX) [13], and Brain Extraction based on nonlocal Segmentation Technique (BEaST) [14]. For pediatric brain imaging, Learning Algorithm for Brain Extraction and Labeling (LABEL) has shown superior brain extraction performance as compared with several other commonly used tools [15]. Newer

N. Swinburne (✉) · A. Holodny
Neuroradiology Service, Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY, USA
e-mail: swinburn@mskcc.org

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_15

approaches for brain extraction that have utilized 3D convolutional neural networks (CNNs) have demonstrated superiority when used specifically for brain tumor studies [16] and have outperformed several older conventional non-CNN approaches [17].

Many big data applications utilize MR images acquired from multiple centers and scanners, which introduces challenges related to source heterogeneity. For example, MR imaging is prone to various artifacts that may degrade the performance of AI applications. Variations in image intensity that occur due to inhomogeneities of MRI field strength, certain image acquisition artifacts, and patient motion may be addressed with bias field correction [18]. Commonly used tools for bias correction include nonparametric nonuniform intensity normalization (N3) [19] and N4ITK [20]. Another issue unique to MR imaging not encountered when using radiographs or CT is that variations in MRI scanner hardware and sequence designs frequently result in differences in image intensities for a given tissue class. Image histogram normalization is a common technique for standardizing these intensities across a heterogeneously acquired dataset. The most common methods include creating and applying an average histogram for the dataset [21] or matching individual images' histograms to that of a chosen reference image [22].

For many AI applications, it is desirable to coregister brain images from different patients (and sequence acquisitions, when using MRI) to a standard geometry, commonly the Montreal Neurological Institute (MNI) space. Many software tools exist for coregistration, such as FMRIB's Linear Image Registration Tool (FLIRT) [23, 24] and Non-linear Image Registration Tool (FNIRT) [25], Advanced Neuroimaging Tools (ANTs) [26], and FreeSurfer. A newer CNN-based approach dubbed Quicksilver has shown promising results and may outperform traditional methods [27].

Data augmentation is a technique for artificially increasing the number of training samples used in situations where large volumes of labeled data are unavailable [28]. Data augmentation has been described for mitigating the risk of overfitting of deep networks and as a method of handling class imbalance by increasing the proportion of the minority (often disease-positive) class. Pereira et al. performed augmentation using image rotation and reported a tumor segmentation mean performance gain of 2.6% [29]. Akkus et al. achieved an 8.8% accuracy gain for classifying 1p/19q mutation status in low-grade gliomas after augmentation by image rotation, translation, and flipping [30].

15.3 Applications

Applications of AI to neuroimaging address all stages of image acquisition and interpretation and approach both specific and complex tasks.

15.3.1 Protocoling, Acquisition, and Image Construction

Once an imaging study is ordered by a referring clinician an imaging protocol must be assigned that is appropriate for the indication and the patient's medical history. Given the importance of cross-sectional imaging in neuroradiology, protocoling may be a complicated task (particularly in the case of MRI) and is typically performed by the radiologist, interrupting workflow [31] and in so doing potentially contributing to diagnostic errors [32]. In addition to unburdening the radiologist, automated protocolling has the potential to increase MR scanner throughput by including only the sequences pertinent to the given patient. Expanding on previous work applying AI to radiological protocoling [33], Brown and Marotta used natural language processing (NLP) to extract labeled data from radiology information system records, which were then used to train a gradient boost machine to generate custom MRI brain protocols with high accuracy [34].

Once MR data is obtained from the scanner it must first be processed into images for the radiologist to review. This initial raw data is processed by a series of modules that require expert oversight to mitigate image noise and

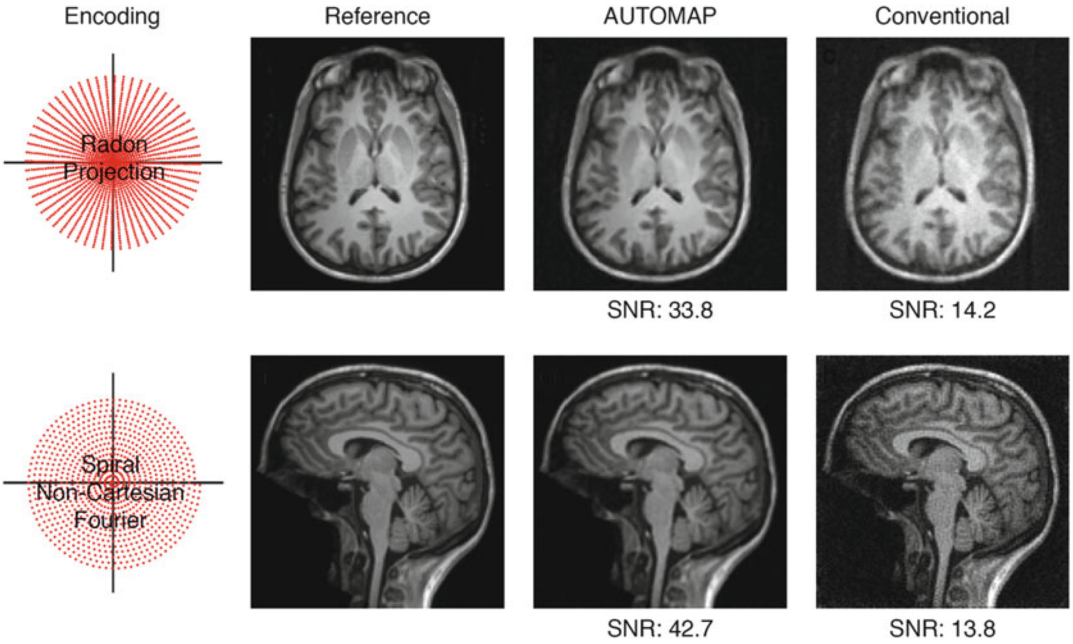


Fig. 15.1 Axial and sagittal MR image reconstructions performed using AUTOMAP (middle column) and using conventional methods (right column), with the ground truth images (left column) included for reference. AUTOMAP, which employs deep learning, results in improved signal-to-noise. Reprinted by permission from Springer Nature: *Nature*, "Image reconstruction by domain-transform manifold learning," Zhu et al. [37]

other artifacts, adding time and introducing variance to the image acquisition process. Building on previous deep learning approaches for shortening MR acquisition times through undersampling [35, 36], a network trained on brain MRI called Automated Transform by Manifold Approximation (AUTOMAP) performs image reconstruction rapidly and with less artifact than conventional methods [37] (Fig. 15.1). Since AUTOMAP is implemented as a feed-forward system it completes image reconstruction almost instantly, enabling acquisition issues to be identified and addressed immediately, potentially reducing the need for patient callbacks.

Deep learning also shows promise for increasing the accessibility of specialized neuroimaging studies by shortening the acquisition time or enabling the generation of entire simulated imaging modalities. For example, diffusion tensor imaging (DTI), which provides information about white matter anatomy in the brain and spine, may be challenging to obtain on young or very sick patients due to the acquisition time and degree of patient cooperation required. Applying deep learning to DTI can achieve a 12-fold reduction in acquisition time by predicting DTI parameters from fewer data points than conventionally utilized [38]. Similarly, a reduction in acquisition time for arterial spin labeling perfusion imaging was achieved using a trained CNN to predict the final perfusion maps from fewer subtraction images [39].

Seven Tesla MR scanners can reveal a level of detail far beyond that of 1.5 or 3 T scanners [40]; however, 7 T magnets are generally confined to academic imaging centers and may be less tolerated by patients due to the high magnetic field strength [41]. By performing canonical correlation analysis on 3 T and 7 T brain MRI from the same patients, Bahrami et al. [42] were able to artificially generate simulated 7 T images using 3 T images for test patients. Furthermore, these simulated 7 T images had superior performance in subsequent segmentation tasks.

Recognizing that at their essence all radiological imaging modalities represent a type of anatomical abstraction, the ability to synthetically generate another MRI sequence, or imag-

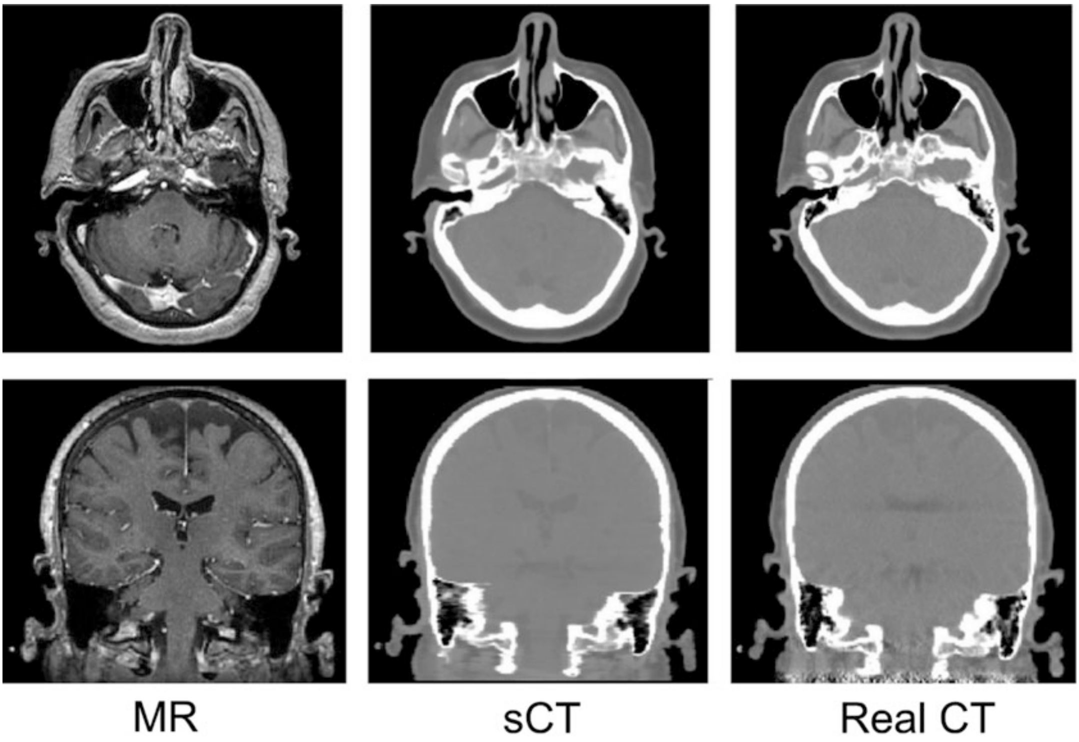


Fig. 15.2 Using a single MRI brain sequence as input (contrast-enhanced T1 gradient echo; left column), a trained CNN can generate synthetic CT (sCT) head images (middle column). Ground truth CT images (right column) are presented for comparison. Reprinted by permission from John Wiley and Sons: *Medical Physics*, “MR-based synthetic CT generation using a deep convolutional neural network method,” Xiao Han [45]

ing modality entirely, presents an intriguing target for AI. Using deep learning, brain MRI T1 images can be generated from T2 images and vice versa [43]. PET–MRI, which holds several advantages over PET–CT, including superior soft tissue contrast, has the disadvantage that in the absence of a CT acquisition it does not readily allow for attenuation correction of the PET images. However, supervised training of a deep network has enabled the generation of synthetic CT head images from contrast-enhanced gradient echo brain MRI, and these synthesized images achieve greater accuracy than existing methods when used to perform attenuation correction on the accompanying PET images [44]. A similar approach was used to train a CNN to utilize a single T1 sequence to generate synthetic CT images with greater speed and lower error rates than conventional methods (Fig. 15.2) [45].

15.3.2 Segmentation

Accurate, fast segmentation of brain imaging, which can be broadly divided into either anatomical (e.g., subcortical structure) or lesion (pathology-specific) segmentation is an important prerequisite step for a number of clinical and research tasks including monitoring progression of white matter [46, 47] and neurodegenerative diseases [48, 49] and assessing tumor treatment response [50]. However, since manual segmentation is tedious, time consuming, and subject to inter- and intra-observer variance, there is great interest in developing AI solutions. To facilitate the comparison of segmentation algorithms, several open competitions exist featuring public datasets and standardized evaluation methodology, several of which are described in this section.

Anatomical brain imaging segmentation entails the delineation of either basic tissue components (e.g., gray matter, white matter, and cerebrospinal fluid) or atlas-based substructures. For the former, commonly utilized brain tissue segmentation datasets include the Medical Image and Statistical Interpretation Lab (MICCAI) 2012 Multi-Atlas Labelling Challenge [51] and the Internet Brain Segmentation Repository (IBSR). Two more specialized MICCAI challenges exist, MRBrainS13 [52], which contains brain MRIs from adults aged 65–80, and NeoBrainS12, which is comprised of neonatal brain MRIs.

The most common brain lesion segmentation tasks addressed by AI are tumor and multiple sclerosis (MS) lesion segmentation. The MICCAI Brain Tumor Segmentation (BRATS) challenges have occurred annually since 2012, with the datasets growing in number over the years to include 243 preoperative glioma multimodal brain MRIs in the 2018 challenge [53, 54]. The winner of the BRATS 2017 segmentation challenge, as determined by the best overall Dice scores and Hausdorff distances for complete tumor, core tumor, and enhancing tumor segmentation, employed an ensemble CNN comprising several existing architectures under the principle that through a majority voting system the ensemble can derive the strengths of its best performing individual networks, resulting in greater generalizability for the performance of other tasks [55].

Additional deep learning segmentation applications target stroke (described subsequently), multiple sclerosis [56, 57], and cerebral small vessel disease (leukoaraiosis) [58] lesions. Anatomical Tracings of Lesions After Stroke (ATLAS-1) is a publicly available annotated dataset containing over 300 brain MRIs with acute infarcts [59]. For MS lesion segmentation, the major public datasets are MICCAI 2008 [60], International Symposium on Biomedical Imaging (ISBI) 2015 [61], and MS Lesion Segmentation Challenge (MSSEG) 2016 [62].

Due to the limited numbers of training and test subjects generally available within existing public annotated datasets, several of the best performing networks for various segmentation tasks have pooled multiple public datasets, supplemented with their own data, or employed data augmentation techniques [63–66]. A study by AlBadawy et al. demonstrated the importance of such measures, finding that the source(s) of tumor segmentation training data held a significant impact on the resulting performance during network validation (Fig. 15.3) [67].

15.3.3 Stroke

Stroke represents a major cause of morbidity and mortality worldwide. For example, in the United States stroke afflicts an estimated 795,000 people each year [68], accounting for 1 in every 20 deaths [69]. With over 1.9 million neurons lost each minute in the setting of an acute stroke [70], it is critical to quickly diagnose and triage stroke patients.

The Alberta Stroke Program Early Computed Tomography Score (ASPECTS) is a validated and widely used method for triaging patients with suspected anterior circulation acute stroke. ASPECTS divides the middle cerebral artery territories into ten regions of interest bilaterally [71]. The resulting score obtained from a patient's non-contrast-enhanced CT head correlates with functional outcomes and helps guide management. e-ASPECTS, a ML-based software tool with CE-mark approval for use in Europe, has demonstrated non-inferiority (10% threshold for sensitivity and specificity) for ASPECT scoring as compared with neuroradiologists from multiple stroke centers [72]. Deep learning networks have also achieved high accuracy at quantifying infarct volumes using DWI [73] and FLAIR [74] MR sequences.

Once a patient is diagnosed with an acute stroke, there is a need to quantify the volume of infarcted (unsalvageable) tissue and the ischemic but not yet infarcted (salvageable) tissue. This latter salvageable tissue is referred to as the ischemic penumbra. Quantification of the infarct core and ischemic penumbra is generally performed with either CT or MR brain perfusion. In the latter approach, the diffusion-perfusion mismatch is used to guide thrombolysis

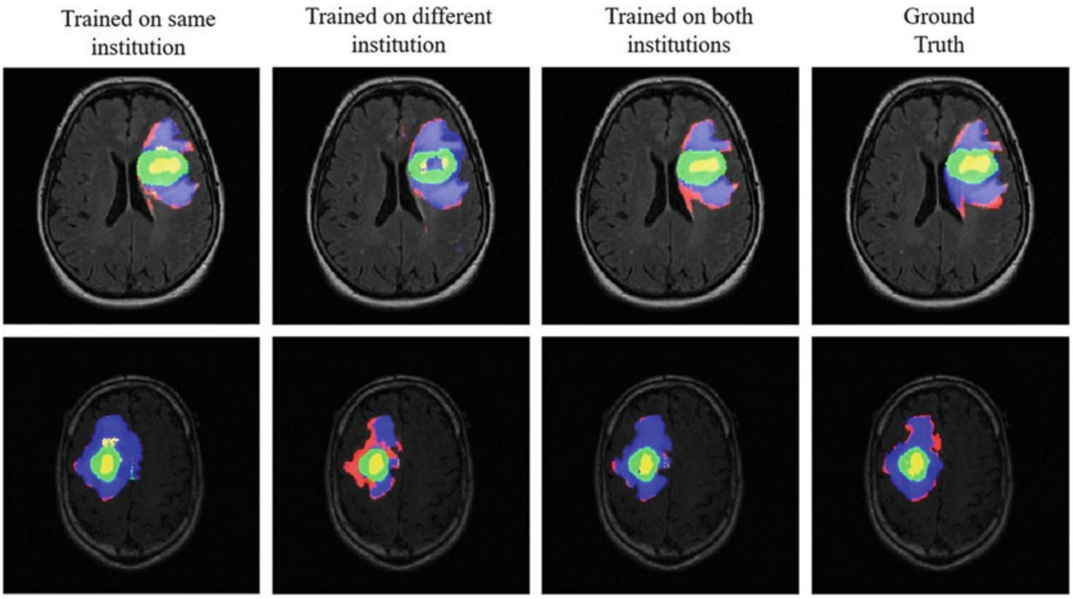

Trained on same institution
Trained on different institution
Trained on both institutions
Ground Truth

Fig. 15.3 Two example brain tumor segmentations generated by separate models trained on data from the same, different, or both institutions. Accuracy was greater when the model was trained with data from the same or both institutions as compared with a model trained only using data from a different institution. The enhancing region (Class 2) is segmented in green, necrotic region (Class 3) in yellow, area of T1 abnormality excluding the enhancing and necrotic regions (Class 4) in red, and the area of FLAIR signal abnormality excluding classes 2–4 (Class 5) in blue. Reprinted by permission from John Wiley and Sons: *Medical Physics*, "Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing," AlBadawy et al. [67]

and thrombectomy decision-making [75]. Using acute DWI and perfusion imaging in concert with follow-up T2/FLAIR as training data, Nielsen et al. developed a deep CNN to distinguish infarcted tissue from the ischemic penumbra using only acute MR perfusion data. They achieved an AUC of 0.88 for diagnosing the final infarct volume and demonstrated an ability to predict the effect of thrombolysis treatment [76]. Additional studies have investigated the prediction of long-term language [77, 78] and motor [79] outcomes using ML evaluation of stroke territory volumes and locations.

15.3.4 Tumor Classification

The ability to classify brain tumor type and World Health Organization grade using MRI has long been a goal of machine learning research. As early as 1998, Poptani et al. used an artificial neural network to differentiate normal brain MR spectroscopy studies from those with infectious and neoplastic diseases, achieving diagnostic accuracies of 73% and 98% for low- and high-grade gliomas, respectively [80]. More recent work has commonly employed support vector machines (SVMs) for tumor classification tasks, perhaps due to evidence that SVMs may perform better than neural networks with small training datasets [81]. In 2008, Emblem et al. applied a SVM approach to the differentiation of low- and high-grade gliomas using MR perfusion imaging, achieving true positive and true negative rates of 0.76 and 0.82, respectively [82]. Subsequent efforts have shown promising results for differentiating among glioma grades and other tumor classes using SVM analysis of conventional MRI without [83] or with [84, 85] the addition of perfusion MRI. Survival of patients with glioblastoma can also be predicted using SVM analysis of features derived from MR perfusion [86], conventional [87], and combined conventional, DTI, and perfusion [88] imaging features. SVM

[88] and other [89] machine learning techniques have also been employed in radiomics research to investigate imaging markers for prediction of tumor molecular subtypes.

Differentiating glioblastoma, primary central nervous system lymphoma, and solitary brain metastasis is a common neuroradiological challenge due to the relatively high prevalence of these tumor classes and the potential for overlapping imaging characteristics. A multilayer perceptron trained using MR perfusion and permeability imaging was able to differentiate these tumor classes with high accuracy (AUC 0.77) comparable to that of neuroradiologists [90].

In the setting of chemoradiation therapy for glioblastoma, differentiating viable tumor from treatment-related necrosis (pseudoprogression) on follow-up brain imaging is a common challenge in clinical neuro-oncology [91]. The application of SVMs to differentiating these entities has shown high accuracy using MR conventional imaging in combination with either perfusion [92] or permeability [93] data. A study evaluating the use of only conventional MRI sequences found that the best SVM accuracy was obtained using the FLAIR sequence (AUC 0.79), which achieved better accuracy than the neuroradiologist reviewers involved in the study [94].

15.3.5 Disease Detection

Applications of AI for neuroimaging disease detection exist within a spectrum of task complexity. On one end, there are applications that perform identification of a specific disease process, which often result in a binary classification (i.e., "normal" vs. "disease"). For example, several applications have been described for differentiating normal brain MRIs from those containing epileptogenic foci [95–97]. On the other end of the spectrum are broader surveillance applications designed to diagnose multiple critical pathologies, which one may envision as ultimately integrating within a real-world clinical radiology workflow. This latter, nascent category has been the source of much excitement [98–101].

In light of the importance and urgency of diagnosing intracranial hemorrhage, a disease process requiring neurosurgical evaluation and representing a contraindication for thrombolysis in the setting of acute stroke, the use of AI for identification of hemorrhage on head CT has been investigated in several studies. Whereas earlier attempts demonstrated promising results employing preprocessing algorithms heavily tailored for isolating hemorrhage [102–104], more recent efforts have investigated whether existing deep CNNs that have shown success at identifying everyday (nonmedical) images could be applied to head CTs. Desai et al. [105] compared two existing 2D deep CNNs for the identification of basal ganglia hemorrhage and found that GoogLeNet [106] outperformed AlexNet [28], noting that data augmentation and pre-training with the ImageNet repository [107] of everyday images improved diagnostic performance (AUC 1.0 for the best performing network). Transfer learning was similarly employed by Phong et al. [108], who achieved comparably high accuracies for identifying intracranial hemorrhage.

A study by Arbabshirani et al. [109] using CNNs to diagnose intracranial hemorrhage differed in several important ways. Whereas the above-described studies utilized relatively small datasets (<200 CT head studies), Arbabshirani et al. included over 46,000 CT head studies. To generate labels for this large number of studies, the authors expanded on other work investigating NLP applications to radiology reports [110, 111] and employed NLP to extrapolate a subset of human-annotated labels to generate machine-readable labels for the remainder of the radiology report dataset. The trained image classification model, which achieved an AUC of 0.846 for diagnosing intracranial hemorrhage, was then prospectively validated in a clinical workflow to flag new studies as either "routine" or "stat" in real time depending on the presence of intracranial hemorrhage. During this 3-month validation period, the network reclassified 94 of 347 CT head studies from "routine" to "stat." Of the 94 studies flagged, 60 were confirmed by the interpreting radiologist as positive for intracranial hemorrhage. An additional four flagged studies were later reevaluated by a blinded overreader and deemed likely to reflect hemorrhage; in other words, the trained network had found hemorrhage that was missed by the interpreting radiologist.

Seeking to diagnose a broader range of intracranial pathologies, Prevedello et al. [112] trained a pair of CNNs using several hundred labeled head CTs for the purpose of identifying a number of critical findings. A CNN for processing images using brain tissue windows was able to diagnose hemorrhage, mass effect, and hydrocephalus with an AUC of 0.90, while a separately trained CNN evaluating images using a narrower “stroke window” achieved an AUC of 0.81 for the diagnosis of an acute ischemic stroke.

Approaching this challenge of simultaneously surveilling for multiple critical findings, Titano et al. [113] utilized a larger dataset of over 37,000 head CTs, first employing NLP to derive machine-readable labels from the radiology reports. These labels were then used for weakly supervised training of a 3D CNN modeled on ResNet-50 architecture to differentiate head CTs containing one or more critical findings (including acute fracture, intracranial hemorrhage, stroke, mass effect, and hydrocephalus) from those with only noncritical findings, achieving a sensitivity matching that of radiologists (sensitivity 0.79, specificity 0.48, AUC 0.73 for the model). To validate the clinical utility of the trained network, the authors performed a prospective double-blinded randomized controlled trial comparing how quickly the model versus radiologists could evaluate a head CT for critical findings, demonstrating that the model performed this task 150 times faster than the radiologists (mean 1.2 s vs. 177 s). Pending further multicenter prospective validation, such a tool could be used in a clinical radiology workflow to automatically triage head CTs for review.

15.4 Conclusion

Having already demonstrated success at a diverse range of neuroradiology tasks, artificial intelligence is poised to move beyond the proof-of-concept stage and impact many facets of clinical practice. The continued advancement of AI for neuroradiology depends in part on overcoming hurdles both technical and logistical in nature. The need for large-scale training data can be addressed by the release of more public annotated datasets, through development of applications that facilitate the creation of labels from existing radiology reports and DICOM metadata, crowdsourcing initiatives, and through improving data augmentation methodologies. The high computational costs of applying deep learning to volumetric data may be overcome by advances in GPU hardware and new techniques that better leverage multicore GPU architectures. Several open-source platforms now exist that facilitate deep learning efforts, including Keras, Caffe, and Theano, and the arrival of turnkey AI development applications is likely imminent. Similarly, while deep neural network architectures currently vary widely in design, standards may arise for specific classes of neuroimaging tasks. Finally, once a deep learning application is developed it must undergo validation, which faces its own regulatory and practical hurdles. For example, the opacity of deep networks, which traditionally function as “black boxes,” can make auditing a challenge, although this may be partially addressed through technical means like generating saliency overlays (i.e., “heat maps”). Regulatory bodies are considering new programs that would allow a vendor to make minor modifications to its existing application without requiring a full resubmission for approval [114], potentially enabling AI tools to continue improving during the postmarket phase.

These advancements, coupled with the tremendous interest in AI applications to neuroradiology, ensure that the field’s pace of evolution will continue to hasten. Whether or not we will witness an AI application that is able to pass the neuroradiology equivalent of the Turing Test—that is, AI possessing diagnostic abilities truly comparable to those of a neuroradiologist—remains a point of considerable debate. It is clear, however, that AI will become an increasingly important part of clinical neuroradiology and will carry with it the accompanying benefits to both patients and physicians.

15.5 Take-Home Points

- Neuroimaging represents an intriguing target for AI applications due to the high morbidity and mortality associated with neurological diseases.
- Technical challenges remain due to the volumetric and multiparametric nature of neuroradiological imaging; however advances in GPU power and development of novel deep learning architectures may enable these challenges to be overcome.
- AI applications to neuroimaging have shown success at handling a range of tasks involving all stages from an imaging study's acquisition through its interpretation, including study protocoling; shortening image acquisition times of conventional, DTI, and ASL MRI; generating synthetic images using a different imaging modality; and lesion segmentation.
- Newer applications successfully identify and quantify specific disease processes including infarcts, tumors, and intracranial hemorrhage, and more robust approaches have shown success in surveilling for multiple acute neurological diseases.

References

1. Le Bihan D. Diffusion MRI: what water tells us about the brain. EMBO Mol Med. 2014;6(5):569–73.
2. Prasoon A, Petersen K, Igel C, Lauze F, Dam E, Nielsen M. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: Medical image computing and computer-assisted intervention – MICCAI 2013 [Internet]. Berlin: Springer; 2013. p. 246–53. (Lecture Notes in Computer Science). Available from: http://link.springer.com/chapter/10.1007/978-3-642-40763-5_31. Accessed 30 May 2018.
3. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multiscale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med Image Anal. 2017;36:61–78.
4. Acosta-Cabronero J, Williams GB, Pereira JMS, Pengas G, Nestor PJ. The impact of skull-stripping and radio-frequency bias correction on grey-matter segmentation for voxel-based morphometry. NeuroImage. 2008;39(4):1654–65.
5. Sadananthan SA, Zheng W, Chee MWL, Zagorodnov V. Skull stripping using graph cuts. NeuroImage. 2010;49(1):225–39.
6. Popescu V, Battaglini M, Hoogstrate WS, Verfaillie SCJ, Sluimer IC, van Schijndel RA, et al. Optimizing parameter choice for FSL-brain extraction tool (BET) on 3D T1 images in multiple sclerosis. NeuroImage. 2012;61(4):1484–94.
7. Smith SM. Fast robust automated brain extraction. Hum Brain Mapp. 2002;17(3):143–55.
8. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. NeuroImage. 2012;62(2):782–90.
9. Woolrich MW, Jbabdi S, Patenaude B, Chappell M, Makni S, Behrens T, et al. Bayesian analysis of neuroimaging data in FSL. NeuroImage. 2009;45(1, Suppl. 1):S173–86.
10. Jenkinsen M, Pechaud M, Smith SM. BET2: MR-based estimation of brain, skull and scalp surfaces. Eleventh annual meeting of the organization for human brain mapping, Toronto, Ontario; 2004. p. 716.
11. Shattuck DW, Sandor-Leahy SR, Schaper KA, Rottenberg DA, Leahy RM. Magnetic resonance image tissue classification using a partial volume model. NeuroImage. 2001;13(5):856–76.
12. Ségonne F, Dale AM, Busa E, Glessner M, Salat D, Hahn HK, et al. A hybrid approach to the skull stripping problem in MRI. NeuroImage. 2004;22(3):1060–75.
13. Iglesias JE, Liu CY, Thompson PM, Tu Z. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans Med Imaging. 2011;30(9):1617–34.
14. Eskildsen SF, Coupé P, Fonov V, Manjón JV, Leung KK, Guizard N, et al. BEaST: brain extraction based on nonlocal segmentation technique. NeuroImage. 2012;59(3):2362–73.
15. Shi F, Wang L, Dai Y, Gilmore JH, Lin W, Shen D. LABEL: pediatric brain extraction using learning-based meta-algorithm. NeuroImage. 2012;62(3):1975–86.

16. Kleesiek J, Urban G, Hubert A, Schwarz D, Maier-Hein K, Bendszus M, et al. Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. NeuroImage. 2016;129:460–9.
17. Duy NHM, Duy NM, Truong MTN, Bao PT, Binh NT. Accurate brain extraction using active shape model and convolutional neural networks. ArXiv180201268 Cs [Internet]. 2018. Available from: http://arxiv.org/abs/1802.01268. Accessed 21 May 2018.
18. Kahali S, Adhikari SK, Sing JK. On estimation of bias field in MRI images: polynomial vs Gaussian surface fitting method. J Chemom. 2016;30(10):602–20.
19. Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging. 1998;17(1):87–97.
20. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging. 2010;29(6):1310–20.
21. Nyul LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. IEEE Trans Med Imaging. 2000;19(2):143–50.
22. Sun X, Shi L, Luo Y, Yang W, Li H, Liang P, et al. Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions. Biomed Eng Online [Internet]. 2015;14. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4517549/. Accessed 21 May 2018.
23. Jenkinson M, Smith S. A global optimisation method for robust affine registration of brain images. Med Image Anal. 2001;5(2):143–56.
24. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage. 2002;17(2):825–41.
25. Andersson JLR, Jenkinson M, Smith S, Jenkinson M, Smith S, Andersson JLR, et al. Nonlinear registration aka spatial normalisation FMRIB technical report TR07JA2. 2007. Available from: https://www.scienceopen.com/document?vid=13f3b9a9-6e99-4ae7-bea2-c1bf0af8ca6e. Accessed 21 May 2018.
26. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage. 2011;54(3):2033–44.
27. Bernal J, Kushibar K, Asfaw DS, Valverde S, Oliver A, Martí R, et al. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. ArXiv171203747 Cs [Internet]. 2017. Available from: http://arxiv.org/abs/1712.03747. Accessed 30 April 2018.
28. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90.
29. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans Med Imaging. 2016;35(5):1240–51.
30. Akkus Z, Ali I, Sedlář J, et al. Predicting deletion of chromosomal arms 1p/19q in low-grade gliomas from MR images using machine intelligence. J Digit Imaging. 2017;30(4):469–76.
31. Schemmel A, Lee M, Hanley T, Pooler BD, Kennedy T, Field A, et al. Radiology workflow disruptors: a detailed analysis. J Am Coll Radiol. 2016;13(10):1210–4.
32. Balint BJ, Steenburg SD, Lin H, Shen C, Steele JL, Gunderman RB. Do telephone call interruptions have an impact on radiology resident diagnostic accuracy? Acad Radiol. 2014;21(12):1623–8.
33. Bhat A, Shih G, Zabih R. Automatic selection of radiological protocols using machine learning. In: Proceedings of the 2011 workshop on data mining for medicine and healthcare (DMMH '11) [Internet]. New York: ACM; 2011. p. 52–5. Available from: http://doi.acm.org/10.1145/2023582.2023591. Accessed 22 May 2018.
34. Brown AD, Marotta TR. Using machine learning for sequence-level automated MRI protocol selection in neuroradiology. J Am Med Inform Assoc. 2018;25(5):568–71.
35. Wang S, Su Z, Ying L, Peng X, Zhu S, Liang F, et al. Accelerating magnetic resonance imaging via deep learning. In: IEEE; 2016. p. 514–7. Available from: http://ieeexplore.ieee.org/document/7493320/. Accessed 22 May 2018.
36. Yang Y, Sun J, Li H, Xu Z. Deep ADMM-Net for compressive sensing MRI. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, editors. Advances in neural information processing systems 29 [Internet]. Curran Associates; 2016. p. 10–8. Available from: http://papers.nips.cc/paper/6406-deep-admm-net-for-compressive-sensing-mri.pdf
37. Zhu B, Liu JZ, Cauley SF, Rosen BR, Rosen MS. Image reconstruction by domain-transform manifold learning. Nature. 2018;555(7697):487–92.
38. Golkov V, Dosovitskiy A, Sperl JI, Menzel MI, Czisch M, Sämann P, et al. q-Space deep learning: twelve-fold shorter and model-free diffusion MRI scans. IEEE Trans Med Imaging. 2016;35(5):1344–51.
39. Kim KH, Choi SH, Park S-H. Improving arterial spin labeling by using deep learning. Radiology. 2017;287(2):658–66.
40. Law M, Wang R, Liu C-SJ, Shiroishi MS, Carmichael JD, Mack WJ, et al. Value of pituitary gland MRI at 7 T in Cushing's disease and relationship to inferior petrosal sinus

sampling: case report. J Neurosurg. 2018:1–5. https://doi.org/10.3171/2017.9.JNS171969.

41. Schaap K, Vries YC, Mason CK, Vocht F, de Portengen L, Kromhout H. Occupational exposure of healthcare and research staff to static magnetic stray fields from 1.5–7 Tesla MRI scanners is associated with reporting of transient symptoms. Occup Env Med. 2014;71(6):423–9.
42. Bahrami K, Shi F, Zong X, Shin HW, An H, Shen D. Reconstruction of 7T-like images from 3T MRI. IEEE Trans Med Imaging. 2016;35(9):2085–97.
43. Vemulapalli R, Nguyen HV, Zhou SK. Chapter 16 – deep networks and mutual information maximization for cross-modal medical image synthesis. In: Deep learning for medical image analysis [Internet]. Academic Press; 2017. p. 381–403. Available from: https://www.sciencedirect.com/science/article/pii/B9780128104088000225. Accessed 22 May 2018.
44. Liu F, Jang H, Kijowski R, Bradshaw T, McMillan AB. Deep learning MR imaging-based attenuation correction for PET/MR imaging. Radiology. 2017;286(2):676–84.
45. Han X. MR-based synthetic CT generation using a deep convolutional neural network method. Med Phys. 2017;44(4):1408–19.
46. Kalkers NF, Ameziane N, Bot JCJ, Minneboo A, Polman CH, Barkhof F. Longitudinal brain volume measurement in multiple sclerosis: rate of brain atrophy is independent of the disease subtype. Arch Neurol. 2002;59(10):1572–6.
47. Mollison D, Sellar R, Bastin M, Mollison D, Chandran S, Wardlaw J, et al. The clinico-radiological paradox of cognitive function and MRI burden of white matter lesions in people with multiple sclerosis: a systematic review and meta-analysis. PLoS One. 2017;12(5):e0177727.
48. Jack CR, Slomkowski M, Gracon S, Hoover TM, Felmlee JP, Stewart K, et al. MRI as a biomarker of disease progression in a therapeutic trial of milameline for AD. Neurology. 2003;60(2):253–60.
49. Simmons A, Westman E, Muehlboeck S, Mecocci P, Vellas B, Tsolaki M, et al. MRI measures of Alzheimer's disease and the AddNeuroMed study. Ann N Y Acad Sci. 2009;1180(1):47–55.
50. Bauer S, Wiest R, Nolte L-P, Reyes M. A survey of MRI-based medical image analysis for brain tumor studies. Phys Med Biol. 2013;58(13):R97.
51. Landman BA, Ribbens A, Lucas B, Davatzikos C, Avants B, Ledig C, et al. MICCAI 2012 workshop on multi-atlas labeling. In: Warfield SK, editor. CreateSpace independent publishing platform; 2012. 164 p.
52. Mendrik AM, Vincken KL, Kuijf HJ, Breeuwer M, Bouvy WH, de Bresser J, et al. MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans [Internet]. In: Computational intelligence and neuroscience. 2015. Available from: https://www.hindawi.com/journals/cin/2015/813696/. Accessed 23 May 2018.
53. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging. 2015;34(10):1993–2024.
54. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, et al. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci Data [Internet]. 2017;4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5685212/. Accessed 23 May 2018.
55. Kamnitsas K, Bai W, Ferrante E, McDonagh S, Sinclair M, Pawlowski N, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation. ArXiv171101468 Cs [Internet]. 2017. Available from: http://arxiv.org/abs/1711.01468. Accessed 23 May 2018.
56. Brosch T, Tang LYW, Yoo Y, Li DKB, Traboulsee A, Tam R. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to Multiple sclerosis lesion segmentation. IEEE Trans Med Imaging. 2016;35(5):1229–39.
57. Valverde S, Cabezas M, Roura E, González-Villà S, Pareto D, Vilanova JC, et al. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. NeuroImage. 2017;155:159–68.
58. Chen L, Carlton Jones AL, Mair G, Patel R, Gontsarova A, Ganesalingam J, et al. Rapid automated quantification of cerebral leukoaraiosis on CT images: a multicenter validation study. Radiology. 2018;288(2):573–81. https://doi.org/10.1148/radiol.2018171567.
59. Liew S-L, Anglin JM, Banks NW, Sondag M, Ito KL, Kim H, et al. A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. Sci Data. 2018;5:180011.
60. Styner M, Lee J, Chin B, Chin MS, Commowick O, Tran H-H, et al. 3D segmentation in the clinic: a grand challenge II: MS lesion segmentation. MIDAS J. 2008;2008, 638.
61. Carass A, Roy S, Jog A, Cuzzocreo JL, Magrath E, Gherman A, et al. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. NeuroImage. 2017;148:77–102.
62. Commowick O, Cervenansky F, Ameli R. MSSEG challenge proceedings: multiple sclerosis lesions segmentation challenge using a data management and processing infrastructure [Internet]. 2016. Available from: http://www.hal.inserm.fr/inserm-01397806/document. Accessed 23 May 2018.
63. Moeskops P, Viergever MA, Mendrik AM, de Vries LS, Benders MJNL, Išgum I. Automatic segmentation of MR brain images with a convolu-

tional neural network. IEEE Trans Med Imaging. 2016;35(5):1252–61.

64. Chen H, Dou Q, Yu L, Qin J, Heng P-A. VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images. NeuroImage. 2018;170:446–55.
65. Wachinger C, Reuter M, Klein T. DeepNAT: deep convolutional neural network for segmenting neuroanatomy. NeuroImage. 2018;170: 434–45.
66. Chang PD. Fully convolutional deep residual neural networks for brain tumor segmentation. In: Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries [Internet]. Cham: Springer; 2016. p. 108–8. (Lecture Notes in Computer Science). Available from: https://link.springer.com/chapter/10.1007/978-3-319-55524-9_11. Accessed 6 May 2018.
67. AlBadawy EA, Ashirbani S, Mazurowski Maciej A. Deep learning for segmentation of brain tumors: impact of cross-institutional training and testing. Med Phys. 2018;45(3):1150–8.
68. Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, et al. Heart disease and stroke statistics-2017 update: a report from the American Heart Association. Circulation, 2017. 135(10):e146–603.
69. Yang Q, Tong X, Schieb L, Vaughan A, Gillespie C, Wiltz JL, et al. Vital signs: recent trends in stroke death rates – United States, 2000–2015. MMWR Morb Mortal Wkly Rep. 2017;66(35):933–9.
70. Saver JL. Time is brain—quantified. Stroke. 2006;37(1):263–6.
71. Barber PA, Demchuk AM, Zhang J, Buchan AM. Validity and reliability of a quantitative computed tomography score in predicting outcome of hyperacute stroke before thrombolytic therapy. Lancet. 2000;355(9216):1670–4.
72. Nagel S, Sinha D, Day D, Reith W, Chapot R, Papanagiotou P, et al. e-ASPECTS software is non-inferior to neuroradiologists in applying the ASPECT score to computed tomography scans of acute ischemic stroke patients. Int J Stroke. 2017;12(6):615–22.
73. Chen L, Bentley P, Rueckert D. Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks. NeuroImage Clin. 2017;15:633–43.
74. Maier O, Schröder C, Forkert ND, Martinetz T, Handels H. Classifiers for ischemic stroke lesion segmentation: a comparison study. PLoS One. 2015;10(12):e0145118.
75. Albers GW, Marks MP, Kemp S, Christensen S, Tsai JP, Ortega-Gutierrez S, et al. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. N Engl J Med. 2018;378(8):708–18.
76. Nielsen A, Hansen MB, Tietze A, Mouridsen K. Prediction of tissue outcome and assessment of treatment effect in acute ischemic stroke using deep learning. Stroke. 2018;49(6):1394–401. https://doi.org/10.1161/STROKEAHA.117.019740.
77. Hope TMH, Seghier ML, Leff AP, Price CJ. Predicting outcome and recovery after stroke with lesions extracted from MRI images. NeuroImage Clin. 2013;2:424–33.
78. Hope TMH, Parker Jones Ō, Grogan A, Crinion J, Rae J, Ruffle L, et al. Comparing language outcomes in monolingual and bilingual stroke patients. Brain. 2015;138(4):1070–83.
79. Rondina JM, Filippone M, Girolami M, Ward NS. Decoding post-stroke motor function from structural brain imaging. NeuroImage Clin. 2016;12:372–80.
80. Poptani H, Kaartinen J, Gupta RK, Niemitz M, Hiltunen Y, Kauppinen RA. Diagnostic assessment of brain tumours and non-neoplastic brain disorders in vivo using proton nuclear magnetic resonance spectroscopy and artificial neural networks. J Cancer Res Clin Oncol. 1999;125(6): 343–9.
81. Shao Y, Lunetta RS. Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points. ISPRS J Photogramm Remote Sens. 2012;70:78–87.
82. Emblem KE, Zoellner FG, Tennoe B, Nedregaard B, Nome T, Due-Tonnessen P, et al. Predictive modeling in glioma grading from MR perfusion images using support vector machines. Magn Reson Med. 2008;60(4):945–52.
83. Alcaide-Leon P, Dufort P, Geraldo AF, Alshafai L, Maralani PJ, Spears J, et al. Differentiation of enhancing glioma and primary central nervous system lymphoma by texture-based machine learning. Am J Neuroradiol. 2017;38(6):1145–50.
84. Zacharaki EI, Wang S, Chawla S, Yoo DS, Wolf R, Melhem ER, et al. Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. Magn Reson Med. 2009;62(6):1609–18.
85. Zacharaki EI, Kanas VG, Davatzikos C. Investigating machine learning techniques for MRI-based classification of brain neoplasms. Int J Comput Assist Radiol Surg. 2011;6(6):821–8.
86. Emblem KE, Due-Tonnessen P, Hald JK, Bjornerud A, Pinho MC, Scheie D, et al. Machine learning in preoperative glioma MRI: survival associations by perfusion-based support vector machine outperforms traditional MRI. J Magn Reson Imaging. 2014;40(1):47–54.
87. Zhou M, Chaudhury B, Hall LO, Goldgof DB, Gillies RJ, Gatenby RA. Identifying spatial imaging biomarkers of glioblastoma multiforme for survival group prediction. J Magn Reson Imaging. 2017;46(1):115–23.
88. Macyszyn L, Akbari H, Pisapia JM, Da X, Attiah M, Pigrish V, et al. Imaging patterns predict pa-

tient survival and molecular subtype in glioblastoma via machine learning techniques. Neuro Oncol. 2016;18(3):417–25.

89. Kickingereder P, Bonekamp D, Nowosielski M, Kratz A, Sill M, Burth S, et al. Radiogenomics of glioblastoma: machine learning–based classification of molecular characteristics by using multiparametric and multiregional MR imaging features. Radiology. 2016;281(3):907–18.
90. Swinburne N, Schefflein J, Sakai Y, Oermann E, Titano J, Chen I, et al. Machine learning for semi-automated classification of glioblastoma, brain metastasis and CNS lymphoma using MR advanced imaging. Ann Transl Med. 2018; https://doi.org/10.21037/atm.2018.08.05.
91. Kim HS, Goh MJ, Kim N, Choi CG, Kim SJ, Kim JH. Which combination of MR imaging modalities is best for predicting recurrent glioblastoma? Study of diagnostic accuracy and reproducibility. Radiology. 2014;273(3):831–43.
92. Hu X, Wong KK, Young GS, Guo L, Wong ST. Support vector machine multiparametric MRI identification of pseudoprogression from tumor recurrence in patients with resected glioblastoma. J Magn Reson Imaging. 2011;33(2):296–305.
93. Artzi M, Liberman G, Nadav G, Blumenthal DT, Bokstein F, Aizenstein O, et al. Differentiation between treatment-related changes and progressive disease in patients with high grade brain tumors using support vector machine classification based on DCE MRI. J Neurooncol. 2016;127(3):515–24.
94. Tiwari P, Prasanna P, Wolansky L, Pinho M, Cohen M, Nayate AP, et al. Computer-extracted texture features to distinguish cerebral radionecrosis from recurrent brain tumors on multiparametric MRI: a feasibility study. Am J Neuroradiol. 2016;37(12):2231–6.
95. Hong S-J, Kim H, Schrader D, Bernasconi N, Bernhardt BC, Bernasconi A. Automated detection of cortical dysplasia type II in MRI-negative epilepsy. Neurology. 2014;83(1):48–55.
96. Ahmed B, Brodley CE, Blackmon KE, Kuzniecky R, Barash G, Carlson C, et al. Cortical feature analysis and machine learning improves detection of "MRI-negative" focal cortical dysplasia. Epilepsy Behav. 2015;48:21–8.
97. Rudie JD, Colby JB, Salamon N. Machine learning classification of mesial temporal sclerosis in epilepsy patients. Epilepsy Res. 2015;117:63–9.
98. Chockley K, Emanuel E. The end of radiology? Three threats to the future practice of radiology. J Am Coll Radiol. 2016;13(12, Part A):1415–20.
99. Jha S, Topol EJ. Adapting to artificial intelligence: radiologists and pathologists as information specialists. JAMA. 2016;316(22):2353–4.
100. Holodny AI. "Am I about to lose my job?!": a comment on "computer-extracted texture features to distinguish cerebral radiation necrosis from recurrent brain tumors on multiparametric MRI: a feasibility study". Am J Neuroradiol. 2016;37(12):2237–8.
101. Davenport TH, Keith J, Dreyer DO. AI will change radiology, but it won't replace radiologists [Internet]. Harvard Business Review. 2018. Available from: https://hbr.org/2018/03/ai-will-change-radiology-but-it-wont-replace-radiologists. Accessed 25 May 2018.
102. Al-Ayyoub M, Alawad D, Al-Darabsah K, Aljarrah I. Automatic detection and classification of brain hemorrhages. WSEAS Trans Comput. 2013;12:395–405.
103. Scherer M, Cordes J, Younsi A, Sahin Y-A, Götz M, Möhlenbruch M, et al. Development and validation of an automatic segmentation algorithm for quantification of intracerebral hemorrhage. Stroke. 2016;47(11):2776–82.
104. Phan A-C, Vo V-Q, Phan T-C. Automatic detection and classification of brain hemorrhages. In: Intelligent information and database systems [Internet]. Cham: Springer; 2018. p. 417–27. (Lecture Notes in Computer Science). Available from: https://link.springer.com/chapter/10.1007/978-3-319-75420-8_40. Accessed 6 May 2018.
105. Desai V, Flanders AE, Lakhani P. Application of deep learning in neuroradiology: automated detection of basal ganglia hemorrhage using 2D-convolutional neural networks. ArXiv171003823 Cs [Internet]. 2017. Available from: http://arxiv.org/abs/1710.03823. Accessed 26 April 2018.
106. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. ArXiv14094842 Cs [Internet]. 2014. Available from: http://arxiv.org/abs/1409.4842. Accessed 25 May 2018.
107. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet large scale visual recognition challenge. Int J Comput Vis. 2015;115(3):211–52.
108. Phong TD, Duong HN, Nguyen HT, Trong NT, Nguyen VH, Van Hoa T, et al. Brain hemorrhage diagnosis by using deep learning. In: Proceedings of the 2017 international conference on machine learning and soft computing (ICMLSC '17) [Internet]. New York: ACM; 2017. p. 34–9. Available from: http://doi.acm.org/10.1145/3036290.3036326. Accessed 25 May 2018.
109. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, Suever JD, Geise BD, Patel AA, et al. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. Npj Digit Med. 2018;1(1):9.
110. Chokshi F, Shin B, Lee T, Lemmon A, Necessary S, Choi J. Natural language processing for classification of acute, communicable findings on unstructured head CT reports: comparison of neural network and non-neural machine learning techniques. bioRxiv. 2017; https://doi.org/10.1101/173310.

111. Zech J, Pain M, Titano J, Badgeley M, Schefflein J, Su A, et al. Natural language-based machine learning models for the annotation of clinical radiology reports. Radiology. 2018;287(2):570–80.

112. Prevedello LM, Erdal BS, Ryu JL, Little KJ, Demirer M, Qian S, et al. Automated critical test findings identification and online notification system using artificial intelligence in imaging. Radiology. 2017;285(3):923–31.

113. Titano J, Badgeley M, Schefflein J, Pain M, Su A, Cai M, et al. Automated deep neural network surveillance of cranial images for acute neurologic events. Nat Med. 2018;24(9):1337–41.

114. Speeches by FDA officials – transforming FDA's approach to digital health [Internet]. Available from: https://www.fda.gov/NewsEvents/Speeches/ucm605697.htm. Accessed 30 May 2018.

16 The Role of AI in Clinical Trials

Irene Mayorga-Ruiz, Ana Jiménez-Pastor, Belén Fos-Guarinos, Rafael López-González, Fabio García-Castro, and Ángel Alberich-Bayarri

16.1 Introduction

Medical imaging has become a key diagnostic element in clinical practice, providing insights on several diseases that would not be detected otherwise. Beyond being a cornerstone in the current framework of clinical practice, medical imaging is progressively achieving a crucial role in the field of clinical trials.

A clinical trial is an experimental evaluation of a product, substance, medication, and diagnostic or therapeutic technique that, in its application to human beings, aims to assess its effectiveness and safety. In each of the four phases which any product must pass in order to reach the market, different aspects will be evaluated, beginning with toxicity and safety and finishing with the effectiveness of the product under evaluation:

- Phase I: Evaluation of treatment safety and the determination of all the possible side effects in a small cohort (group of subjects/patients). In a cohort of a phase, around 15–25 people are involved. People involved in clinical trial are those patients that, for example, in case of cancer treatment in the conventional treatment has not been effective and for which there are no alternative therapies. Patients with tumors for which there is no standard treatment with proven efficacy are also included.
- Phase II: The evaluation of the treatment effectiveness, always taking safety into account. Cohort size increases. There are around 40–60 patients involved. The patients involved have a very specific tumor or disease.
- Phase III: Last step before a drug can be released on the market. In this phase, drug effectiveness is evaluated, and any possible side effects have to be monitored. Both safety and effectiveness of the new drug will be compared against similar drugs available in the market to assess the added value of the new drug. Treatment is given to larger cohorts, usually and due to this need of big patient recruitment (>100 patients). The phase III clinical trials are carried out in several sites.
- Phase IV: After the approval of the new drug by the regulatory agencies, several studies are carried out with the aim of providing extra information about drug-related risks, benefits, and best use. Phase IV clinical trial cohorts

I. Mayorga-Ruiz (✉) · A. Jiménez-Pastor ·
B. Fos-Guarinos · R. López-González · F. García-Castro ·
Á. Alberich-Bayarri
Quantitative Imaging Biomarkers in Medicine, Valencia, Spain
e-mail: irenemayorga@quibim.com;
anajimenez@quibim.com; belenfos@quibim.com;
rafaellopez@quibim.com; fabiogarcia@quibim.com;
angel@quibim.com

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_16

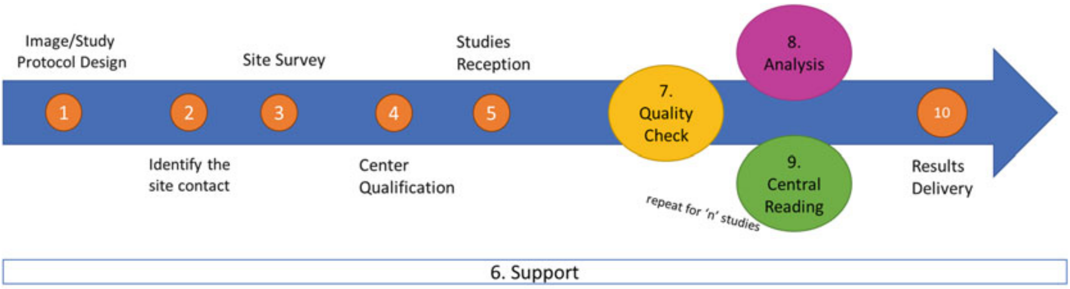


Fig. 16.1 Medical image standardization pipeline for clinical trials

are similar to phase I, II, and III cohorts if new medical indications of the drugs are being evaluated.

Regarding imaging, regulatory agencies recommend the centralized management of the explorations [1]. Therefore, in clinical trials, image acquisitions are usually uploaded to electronic platforms, which store the image studies from all the sites involved in the clinical trial. Even more, agencies also recommend the centralization of the radiological image reading, in order to avoid reading inconsistencies among sites due to criterion differences among radiologists. All these services, including the centralized storage and reading of the images, are usually provided by a core laboratory. A complete workflow of the tasks to be performed by a core laboratory specialized in medical imaging can be appreciated in Fig. 16.1.

The introduction of medical imaging in clinical trials has allowed the evaluation of different treatment lines in a more objective and less invasive way for the subject, reducing the iatrogenic phenomena derived from any medical intervention. With the increasing involvement of medical imaging in the world of clinical trials, the need of radiological and nuclear medicine centralized reading has also increased. This scenario presents a perfect niche for the application of artificial intelligence (AI) techniques and imaging biomarker quantification algorithms. Imaging biomarkers and AI techniques can be used as a means of improving the workflow and supporting diagnostic decisions of the radiological and nuclear medicine specialists.

The use of medical imaging, biomarker quantification, and the application of AI to any of the four phases allow the objectivation of drug safety and effectiveness evaluation in a shorter period of time. This represents a paradigm shift in clinical trials, with the creation of the concept of clinical trial in real time. Real-time evaluation of treatment effectiveness and safety allows the extraction of conclusions before the expected end of the trial or even increasing the cohort size because of the statistical power required in a glimpse of partial satisfactory results under the current conditions. This flexibility allows to perform the statistical assessment of the drug effectiveness almost at any time during the study and reduces the time-to-market of drugs.

In this chapter, the different applications of AI in medical imaging that enhance clinical trials workflows are presented, proposing a methodology for imaging standardization in clinical trials and setting the initial steps toward the future implementation of in silico clinical trials, that is, the computational evaluation of drug toxicity, effectiveness, and efficacy through simulation of the interactions of the drug with the human body multi-scale models.

16.2 Standardization of Medical Imaging in Clinical Trials

The standardization of medical imaging in clinical trials is crucial to avoid biases and errors that may invalidate the conclusions extracted in a study. Guidelines proposed by the Food and Drug Administration (FDA) and the European

Medicines Agency (EMA) aim to set the procedures to follow in the design and management of medical imaging in clinical trials [1–3]. These guidelines specify the need of centralized storage, management, and reading of medical images in multisite clinical trials to ensure the homogenization of images.

Standardization is especially important in multisite clinical trials. In this kind of clinical trials, different medical imaging equipment is involved, as each site might have, for instance, magnetic resonance (MR) or computed tomography (CT) scanners from different vendors. This equipment heterogeneity can introduce variations in the acquisition and reconstruction of imaging studies and, therefore, uncertainty in the imaging biomarker quantification.

The stepwise methodology for the standardization of medical imaging in clinical trials, shown in Fig. 16.1, covers from the design of the image acquisition protocol to the quantification of medical images using AI techniques.

16.2.1 Before the Start of Clinical Trial

16.2.1.1 Image Acquisition Protocol Design

For the standardization of medical images, the first point that has to be taken into account is the correct design of the image acquisition protocols. This will allow the use of imaging biomarkers and artificial intelligence techniques in clinical trials.

For the correct definition of the image acquisition protocol, it is important to define, first and foremost, which is the image modality that best suits the requirements of the clinical trial. It is also mandatory to properly choose the series/sequences which better depict the disease or the underlying biological process to be evaluated. The correct design of the acquisition protocol is the first step to ensure the reliability, repeatability, and quality of radiological readings and imaging biomarker quantification in clinical trials. The parameter definition is of utmost importance and must be designed following this rule: the protocol should minimize image acquisition time and radiation dose (in case of ionizing radiation modalities) while maximizing image quality in terms of spatial resolution, contrast, and also temporal resolution when required (i.e., dynamic examinations).

The different image acquisition technology configurations within the machines are named using different commercial acronyms, and parameters are sometimes expressed in different units (i.e., reception bandwidth in MR provided in “pixels,” Philips; “Hz/pixel,” Siemens/Canon; “kHz,” GE); therefore, the core lab has to design the image acquisition protocols by taking into account the configurations in all the scanners involved in the study. Even if the same acquisition parameters are specified, sometimes the different acquisition algorithms will introduce differences between vendors. For example, although in a CT examination the dose administered to the patient can be reduced significantly while maintaining image quality by using dose modulation options provided by the manufacturers, the dose modulation algorithms are different depending on the manufacturer, and this introduces variability in the image quality and in the extracted data.

16.2.1.2 Site Validation

The second step to guarantee the quality of the imaging studies is the validation of the sites involved in the clinical trial**.** The site validation procedure is composed by three validation steps that ensure, from a theoretical and technical point of view, the capabilities of the sites to acquire and transmit images above the set quality threshold.

1. Site survey

   The site survey is a specific document in clinical trials that aims to collect information about the technical capabilities of the sites participating in the study.

   These technical capabilities include the acquisition modalities available on each site and the equipment characteristics. Also, it is important to assess the DICOM file transmission and exportation capabilities from the PACS; as in any clinical trial, the images are transferred to the central platform of the imaging core laboratory.

The site survey is a document specific to each clinical trial which will vary according to the clinical trial characteristics. The site survey document will vary depending on the image acquisition modality (MRI, CT, X-ray, SPECT, and PET) and the series that are going to be acquired.

Examples of document structure for the site survey in a clinical trial using dynamic contrast-enhanced (DCE) MR are the following:

– MR machine

  Some of the information required related to the MR machines in the site survey are the manufacturer and the model of the scanner. Also, the software version of the machine is required. One important point related to MRI acquisition are the coils used during the image acquisition. Related to coils, the information required is if the coil is multichannel and the number of channels.

– DICOM capabilities

  Related to the DICOM capabilities, the most important information is the capability of the site to export images in DICOM format. This is crucial as DICOM format is the medical imaging standard. Another point that has to be asked in the site survey is the Internet connection of the site. This is important for multisite clinical trials that centralized the images in an electronic platform which usually has the anonymization module integrated. Related to that, for those sites with bad Internet connection, it is important to be sure about its anonymization capabilities as far as they will send the images instead of using the electronic platform through standard courier.

– Contrast

  Finally, and regarding the contrast media used in the DCE acquisition, to ensure the homogenization of the images in multisite clinical trials, it is important to require information about contrast manufacturer and trade. This is important because image contrast is influenced by contrast molarity. Also, the pump information is needed in order to ensure that site's capabilities are homogeneous among sites. In case the clinical trial modality is a CT with contrast, this information will be also important. Therefore, contrasts such as iodine concentration, volume, and injection speed have to be asked.

2. Cross-calibration

   This is a novel step, traditionally not present in the clinical trial process involving medical imaging but will progressively become more and more important, since most of the data that will be analyzed for the evaluation of treatment response in the next years will be related to imaging biomarkers. The cross-calibration of the imaging biomarkers measured using different equipment in a clinical trial will therefore be a must.

   After collecting the different site surveys sent to the sites involved and once it has been assessed which of the sites comply with the requirements of the clinical trial, the next step to assure the quality of the images is the technical validation of the sites through the evaluation of the possible biases that might be introduced by the equipment.

   The cross-calibration of the equipment of the different sites involves the detection of the biases introduced by each piece of equipment in order to calibrate the biomarker results accordingly. Cross-calibration should always be performed using an imaging phantom specific to the image modality included in the trial.

   An imaging phantom is a device specifically designed to evaluate, analyze, and tune the performance of several imaging acquisition modalities. Phantoms are specific to an imaging modality, e.g., the evaluation of the imaging quality of a CT scanner will be performed with a phantom designed for CT scanners. The need for the cross-calibration of equipment is being introduced in guidelines for image standardization in clinical trials by some radiological societies, such as the American College of Radiology [2, 3].

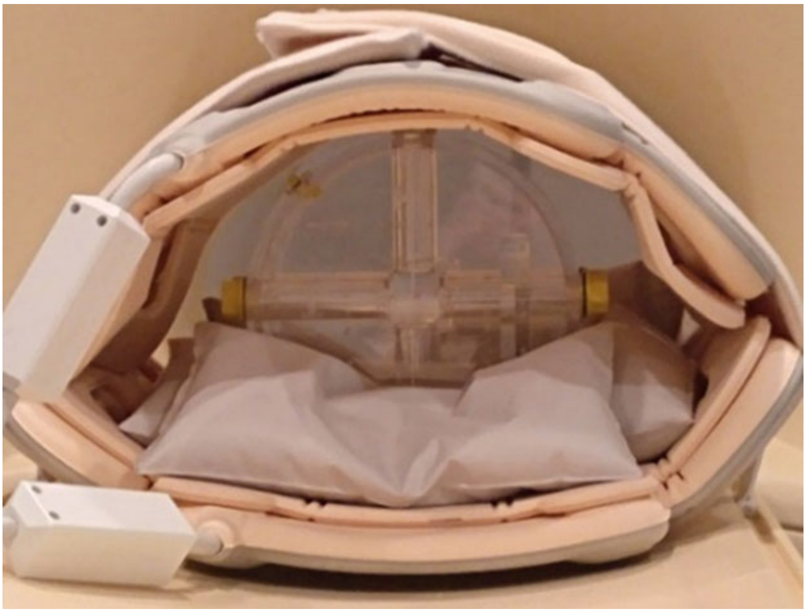


Fig. 16.2 MRI imaging phantom for T1 and T2 calibration (Leeds Test objects, Leeds, UK)

Figure 16.2 shows an example of an imaging phantom intended for T1 and T2 mapping calibration in MR in Fig. 16.3; the CT imaging phantom for Hounsfield unit (HU) measurement performance evaluation is shown.

The sequences (MR) or series used for the acquisition of the cross-calibration examinations have to be the same as the ones that were defined for the imaging study acquisition of the clinical trial. Using this methodology, the inherent biases of the equipment can be determined and therefore taken into account before applying AI algorithms for quantification.

The cross-calibration procedure will guarantee that the imaging biomarker data is reliable and comparable among sites, as it will help reduce or remove equipment biases. For the cross-calibration, the correction factors to be applied to the results of every machine in order to normalize them considering a reference pattern will be calculated. An example of cross-calibration report of a dual-energy X-ray absorptiometry (DXA) machine is shown in Fig. 16.4 where the correction factors have been provided considering the reference values of a DXA phantom.

3. *Dummy run* exploration

   A dummy run study is the acquisition of a subject following the guidelines established in the clinical trial and with the specific image acquisition protocol. The purpose of this examination is to evaluate that the center has understood the patient preparation procedures and has properly introduced all the parameters in the scanner configuration. With this last step, the performance of the image acquisition in each site is evaluated in order to correct the acquisition protocols, if necessary. The core laboratory will receive the dummy run and verify image quality and parameters in the DICOM header, to verify they are within the ranges specified in the protocol.

16.2.1.3 During the Clinical Trial

Quality Assurance of Medical Images

Finally, an additional key aspect in the process of medical image standardization in clinical trials is image quality. Because of that, a quality assurance check has to be done for each of the studies that are included in the trial. There are two goals

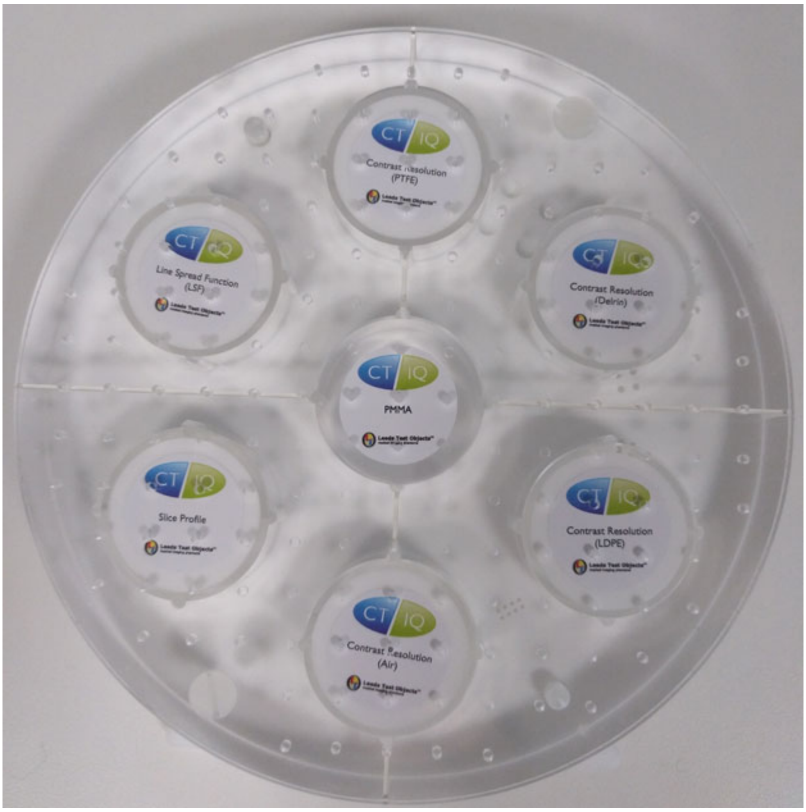

CT IQ
Contrast Resolution (PTFE)
Line Spread Function (LSF)
Contrast Resolution (Delrin)
PMMA
Slice Profile
Contrast Resolution (LDPE)
Contrast Resolution (Air)
Leeds Test Objects

Fig. 16.3 CT imaging phantom for the study of Hounsfield unit performance (Leeds Test objects, Leeds, UK)

of this quality check step:

- First, to guarantee the quality of the images involved in the trial by reviewing the image compliance with the acquisition protocol and by detecting the existence of image artefacts that can affect image quantification. These items ensure the accuracy and reliability of the quantified imaging biomarkers.
- Second, to reduce the number of excluded subjects, because the problems detected in an examination can be rapidly fixed, reducing the possibility of another subject being excluded for the same reason.

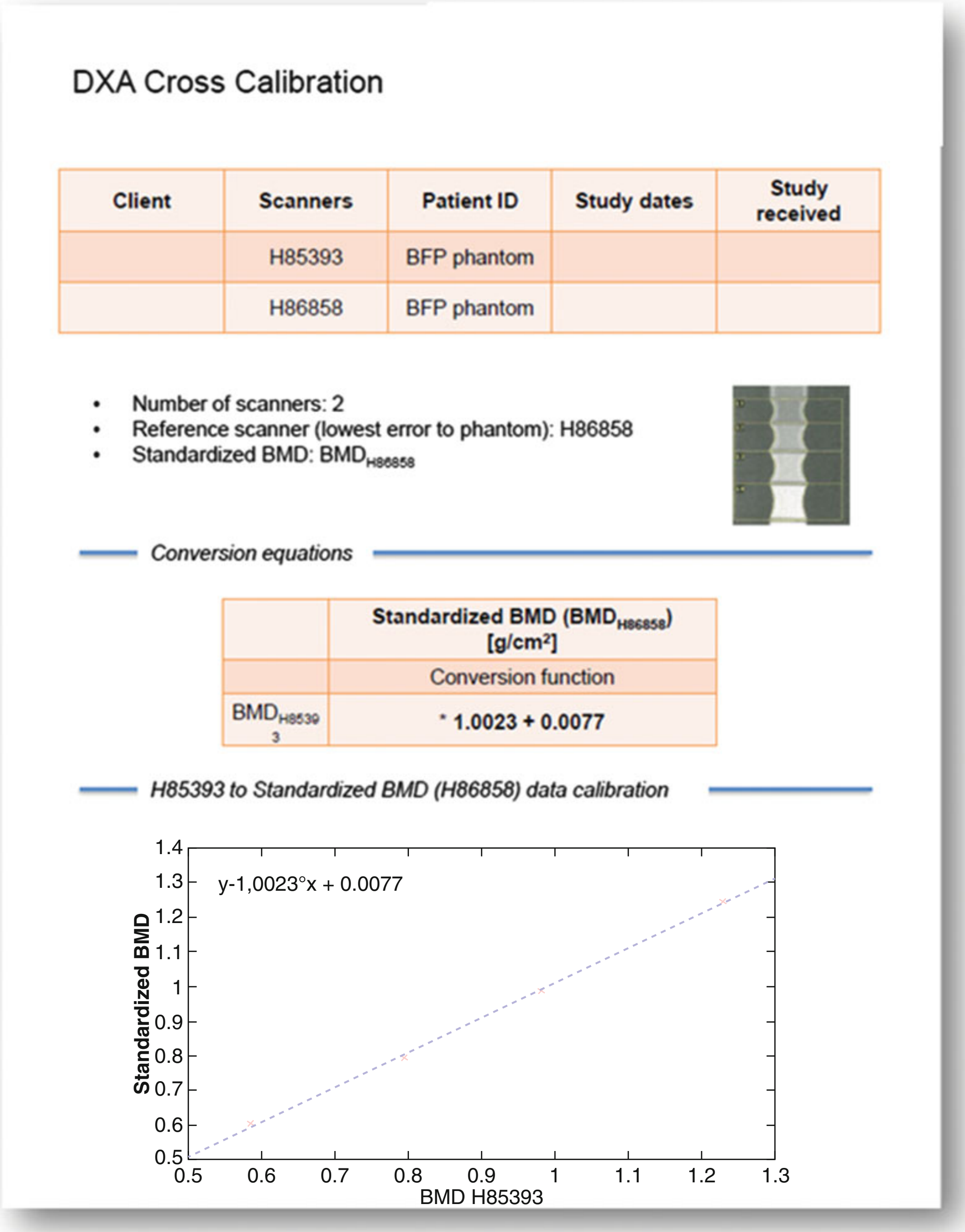
For example, some of the aspects that are checked in a quality assurance of a low-dose thorax CT are the following:

– Protocol compliance: Low-dose thorax CT should be acquired using 120 KVp as tube voltage and 50 mAs as current. Also, slice thickness should be below 2 mm having a pixel size smaller than 1 mm. All those acquisition parameters ensure the accuracy of the images.

– On the other hand, artefacts are checked. Artefacts in thorax CTs usually are due to metallic objects or patient movement. If one of those artefacts is detected, images have to be excluded from the trial.

16.3 Artificial Intelligence in Clinical Trials

As of today, AI techniques are not being widely used in clinical trials, as they are not yet considered by regulatory agencies such as the Food and Drug Administration (FDA) or the European Medicines Agency (EMA) [4, 5]. However, AI is being included in clinical trials to help assess exploratory objectives and endpoints. With these advanced tools, researchers are collecting data



DXA Cross Calibration

Client	Scanners	Patient ID	Study dates	Study received
	H85393	BFP phantom		
	H86858	BFP phantom		

- Number of scanners: 2
- Reference scanner (lowest error to phantom): H86858
- Standardized BMD: BMD_{H86858}

Conversion equations

	Standardized BMD (BMD_{H86858}) [g/cm²]
	Conversion function
BMD_{H8539} 3	*** 1.0023 + 0.0077**

H85393 to Standardized BMD (H86858) data calibration

Fig. 16.4 Example of cross-calibration report of a dual-energy X-Ray absorptiometry (DXA) machine

and results that support the need of including these techniques in regulatory procedures in order to introduce them in clinical trials.

AI can be applied for the following aspects:

- Patient stratification and inclusion: can help to better select those patients that will benefit from therapy
- Automated assessment of quality assurance
- Automating the extraction of quantitative imaging biomarkers
- Shortening image reading times
- Creating multivariate models from the extracted data to increase statistical power of the results

AI in clinical trials can be used in any of the steps that should be covered in a clinical trial. Thanks to AI classification algorithms, patient stratification and inclusion can be speeded. AI classification tools can be trained to detect given patient data the compliance of it to the inclusion criterions. This could be helpful in the patient's recruitment. Also, and as quality assurance parameters are well defined, AI tools could check automatically all the studies reducing time and the need of human supervision.

Regarding medical imaging, AI can be used for the automatic extraction of imaging biomarkers. But also, it can be used to help radiologist in reading the studies. AI can detect suspicious regions due to image characteristics that are suspicious of cancerous tissue. This could help radiologist to focus first on that regions while they are reading the study. Also, it will help the radiologist to be sure that they do not forget any region to evaluate. AI tools which can also be used to create multivariate models using data collected from the clinical trial can model or predict the evolution of a disease helping the pathology management.

AI brings many benefits to medical image processing, allowing to automatize tasks that were performed manually in the past, like structure, tissue, or organ segmentation. Thanks to this automatization, AI avoids human intervention, reducing the subjectivity and variability inherent to any human-dependent process. The improvement of objectivity is a key point in clinical trials, as it allows the homogenization of, for example, image segmentation. Another essential benefit introduced by AI is the improvement of the radiologist's workflow, specifically shortening reading times. This can be achieved creating a CAD system to automatically classify images. The CAD will perform a screening to differentiate between normal and abnormal images. Those imaging studies that were classified as abnormal will be given priority at the time of the radiological reading, hence improving the time efficiency of the radiologists.

Furthermore, AI allows the automatization of imaging biomarker quantification. This means that images can be quantified automatically, with minimum human intervention, reducing time and economic impact as radiologist specialist do not have to do repetitive and time-consuming tasks such as organ segmentation and they can focus on the expert reading of the images. Also, as imaging biomarkers can detect slight changes to the naked human eye in shorter times, the concept of clinical trials in real time appears. Real-time clinical trials are the result of introducing imaging biomarkers in clinical research. Imaging biomarkers are very sensitive to small functional or morphological changes that may occur in the body over a very short period of time after treatment. These changes cannot be detected in the traditional radiological qualitative reading, as they are not discernible by the human eye; this happens in the evaluation of diffusion and perfusion images. Because of the improved sensitivity in a shorter period of time, the evolution of the different branches of treatment can be studied in real time. This helps to reduce time while facilitating decision-making in clinical trials.

In medical imaging clinical scenarios, the use of AI techniques is mostly based on deep learning algorithms, as they are more sensitive and can reach higher performances than other subsets of AI and machine learning. One of the drawbacks of deep learning is that large labeled datasets are needed to achieve optimal results. In a clinical environment, this is a limitation, since it's very difficult to obtain properly annotated data. Due to the lack of large labeled datasets of medical images, a very common approach to improve the efficiency of neural networks is to use transfer learning [6]. Transfer learning allows to use a pretrained network and retrain it using fine-tuning with a medical imaging dataset. A very extended approach is to use deep networks trained using ImageNet which is made by millions of real-world images. Despite these images having different visual properties than medical images, the parameters learned by the networks trained with this dataset can be used as a starting point to train further from there with medical image data. It has been proved that this approach improves the performance of the network than starting from random parameters.

Data augmentation techniques are commonly used to increase the number of available training data. Data augmentation artificially increases the size of the dataset, hence reducing the possibility of overfitting. Applying minor transformations to the original images allows to create a richer dataset. Some of the usual transformations applied are the addition of Gaussian noise, rotation, flip, and translation of the images [7, 8].

Bearing these considerations in mind and while regulatory aspects are not fully defined, there are many applications in which AI can be used in clinical trials to help improve the time efficiency and homogenization of radiological readings and biomarker quantification. Two applications of AI in clinical trials are introduced and described in this chapter.

16.3.1 Classification Algorithms

Classification algorithms are used to predict classes or labels given an input data. This label prediction is performed using a trained model, which is able to split the images into classes with a class probability, i.e., the probability of the image of belonging to a certain class.

The most common function to measure the performance of classification networks in deep learning is cross-entropy [9]. High values of cross-entropy imply a worse performance in classification. Therefore, the objective during training is to learn the network parameters which minimize cross-entropy between the predicted labels and the ground truth labels.

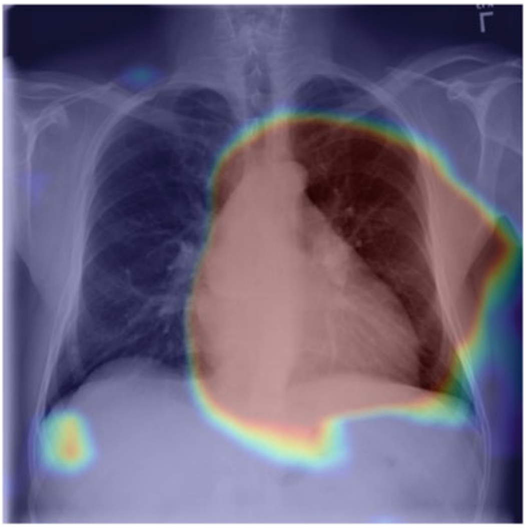


Fig. 16.5 Probability map given by the chest X-ray classifier

An example of a classification algorithm is a chest X-ray classifier. This AI tool can be applied in the screening step of a clinical trial. It could help radiologists focus on abnormal studies with a certain pathology that can be included in the clinical trial. The chest X-ray classifier methodology allows the classification of chest X-rays in two classes, healthy and pathological, while given a class probability for each possible pathology. So, given an input image, the chest X-ray classification model predicts the probability of that image to be normal vs abnormal. Figure 16.5 shows a probability map given by the classifier (an abnormal chest X-ray with cardiomegaly).

16.3.1.1 Segmentation Algorithms

Segmentation is a particular type of classification. Each pixel of an image is classified as a part of a structure or as a part of the background.

In medical imaging segmentation applications, the need of large datasets hindered the development of tools that could be used in clinical trials. The emergence of a specific neural network architecture called UNET has partially mitigated this problem. The main advantage of UNET is its high performance,

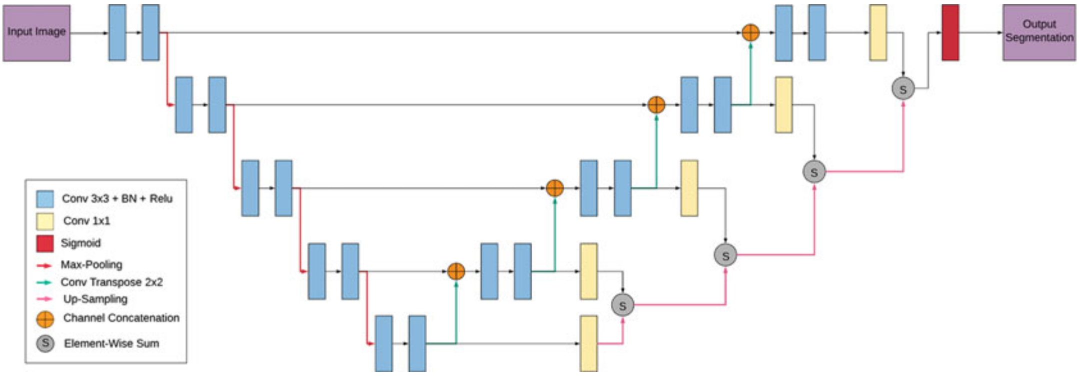


Fig. 16.6 UNET architecture

even with reduced datasets with a limited number of samples. UNET consist of deep convolutional neural networks (CNN) composed by a compression stage in which multi-resolution image features are extracted and a second stage of expansion where the compressed characteristics are decompressed to obtain a mask image with the same size of the input image (Fig. 16.6).

In segmentation problems, the function used to evaluate the performance of the segmentation is called the Dice score. The Dice score analyzes the degree of agreement between two binary masks. One of the masks will be the one segmented by the specialist and the other the one segmented by the AI application. The higher the Dice score (between 0 and 1), the higher the similarity between the masks. Therefore, the goal during training is to obtain the network parameters that maximize the DICE score between the predicted segmentation masks and ground truth segmentation masks.

Automated organ segmentation can be applied in any region. An example of organ segmentation that can be introduced in clinical trials is prostate segmentation (Fig. 16.7). This segmentation will help the radiologist to reduce time in the evaluation of prostate size and will allow to assess functional changes, as, for example, in cellularity or neovascularization, with the aid of imaging biomarkers.

Another application that could significantly shorten the time needed for manual organ segmentation is the automatic liver segmentation. Manual segmentation of the liver is a very time-consuming task, not compatible with routine clinical practice or clinical trials. However, the automatic segmentation of the liver takes mere seconds, allowing to quantify, for example, diffuse liver diseases with fat and iron fraction biomarkers (Fig. 16.8) [10]. The automatic liver segmentation removes interobserver variability, hence minimizing possible errors due to human interaction and ensuring the accuracy and reliability of the imaging biomarker results. Iron and fat fraction biomarkers are of the utmost importance in the evaluation of the diffuse liver diseases, as it reduces the number of invasive biopsies and, therefore, clinical interventions.

16.4 Digital Twin and In Silico Clinical Trials

Another application in which AI could be used in clinical trials is the creation of digital twins for the execution of in silico clinical trials [11]. A digital twin is a digital model that mimics a real physiological condition, process, or system and usually created using big data techniques. The digital twin models must be created under a multi-scale approach, from molecule, cells, tissues, organs, systems, and human scales of simulation. They will be used to evaluate new drug molecule interaction with cells containing specific DNA characteristics that will have consequences in the tissue and with all other scales.

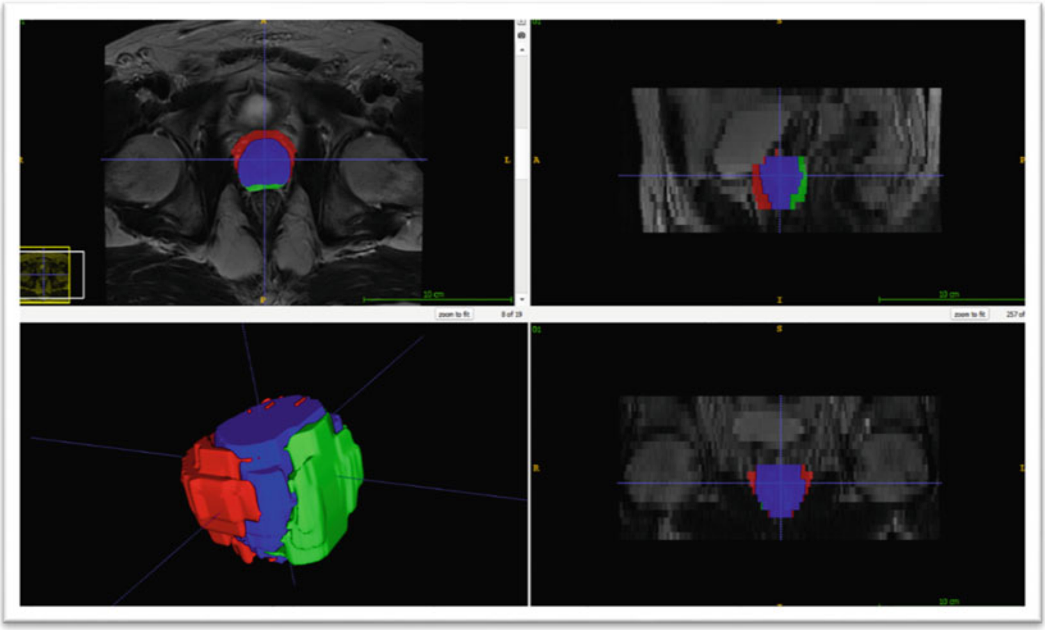


Fig. 16.7 Automatic prostate segmentation

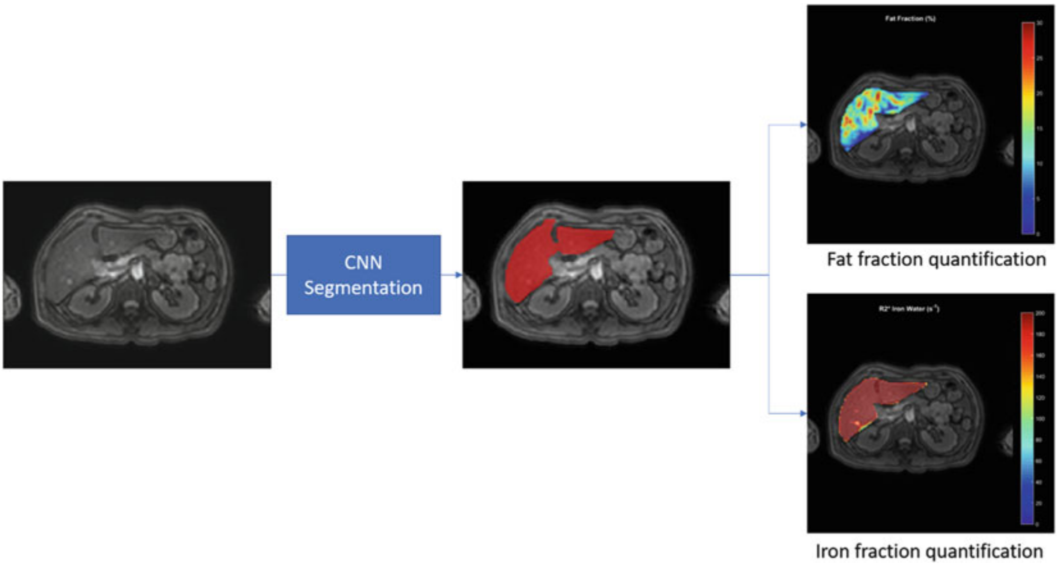


Fig. 16.8 Methodology for automatic liver segmentation and fat and iron quantification

Using these in silico clinical trials which means using digital twins instead of animal models or patients allows, for example, the evaluation of drug toxicity without the need to perform clinical trials with animals or humans at first stages reducing risks [12].

The main limitation of digital twins are the high computational requirements needed for the development of an accurate digital twin. Other limitations of these disruptive concepts of digital twins and in silico clinical trials are the need of developing a multi-scale model that accurately replicates the physiological conditions that need to be evaluated. Therefore, to create a good model, the size of the datasets used for training the system is crucial. As with AI techniques, this presents an important limitation for the introduction and regulation of digital twins and in silico clinical trials, as large labeled datasets are very rare in the medical environment.

16.5 Conclusion

In this chapter the importance of image standardization in clinical trials has been introduced, proposing a stepwise methodology to guarantee image quality along the different steps of the clinical trial in terms of imaging. Also, this chapter shows the advantages of using imaging biomarkers and AI tools to manage patient recruitment and stratification, image quality assurance, radiological reading, and disease management. These quantification techniques allow the reduction of the radiological reading times and ensure the accuracy of the image quantification results.

The evolution in the next years of the computational techniques and the AI tools will promote the creation of digital twins that will allow the use of the in silico clinical trials, reducing this way the number of patients involved, homogenizing the results of the trials, and extracting conclusions in shorter times reducing time-to-market of the products evaluated.

16.6 Summary

Medical image standardization in clinical trials is a crucial point to ensure reliability of the results extracted in a clinical trial. This standardization is important either in the conventional radiological reading and the imaging biomarker extraction. In this chapter, a standardized methodology for medical image validation is proposed. This procedure takes care from the sites capabilities to the quality of the scanners performing a cross-calibration of the equipment involved in the clinical trial. On the other hand, once the MR equipment and the sites are validated, the introduction of AI tools, imaging biomarkers and digital twins in clinical trials could reduce time and economic impact appearing the concept of real-time in clinical trials as treatment branches response can be monitored continuously.

References

1. FDA Guidance for Industry. Clinical trials imaging endpoints process standard. 2018. https://www.fda.gov/downloads/drugs/guidances/ucm268555.pdf. Accessed 24 May 2018.
2. EMA ICH Guidelines. http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000035.jsp&mid=WC0b01ac0580027645. Accessed 24 May 2018.
3. NCI. NCI-CQIE qualification materials. https://www.acrin.org/corelabs/ncicqiequalificationprogram/sitequalificationmaterials.aspx. Accessed 12 May 2018.
4. Food and Drug Administration. FDA official web. https://www.fda.fda.gov. Accessed 10 May 2018.
5. European Medicine Agency. EMA official web. https://www.ema.europa.eu/ema. Accessed 10 May 2018.
6. Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging. 35(5):1285–98.
7. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. 2017.
8. DeVries T, Taylor GW. Dataset augmentation in feature space, 2017. arXiv:1702.05538v1.

9. Cheatsheet ML. Official web. http://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html. Accessed 17 Jul 18.
10. França M, Alberich-Bayarri Á, Martí-Bonmatí L, et al. Accurate simultaneous quantification of liver steatosis and iron overload in diffuse liver diseases with MRI. Abdom Radiol. 2017;42: 1434–43.
11. Bruynseels K, Santoni de Sio F, van den Hoven J. Digital twins in health care: ethical implications of an emerging engineering paradigm. Front Genet. 2018;9:31.
12. Viceconti M, Henney A, Morley-Fletcher E, editors. In silico clinical trials: how computer simulation will transform the biomedical industry. Brussels: Avicenna Consortium; 2016.

Part VI

Quality, Regulatory and Ethical Issues

Quality and Curation of Medical Images and Data

17

Peter M. A. van Ooijen

Recent years have shown an explosive growth in the use of artificial intelligence (AI) and deep learning (DL) not in the least for medical applications. These new technological developments have started a whole new discussion on how we can use the vast amount of available data in health care for processing by these computerized systems. However, especially the applications in health care demand a high level of (patient) data privacy and security. Furthermore, increasingly the requirement of getting appropriate consent from the patient, client, or participant is enforced [1] leading to additional challenges when collecting (retrospective) data. Another concern is that—although an abundance of data are acquired in health care—much of the health-related data are unstructured and not standardized. The actual ownership of medical data is also part of this discussion where different ownership rules can be involved with original, de-identified, anonymized, and processed data. Questions that arise from this are, for example, what data are still personal data for an individual patient or participant in a clinical trial and who actually owns the data that is produced by self-learning computer systems.

Once these issues and questions are solved and data can be collected and used, in many cases the big data are collected with a specific goal in mind in which the focus is on data quantity instead of data quality. This can hamper proper implementation and even lead to incorrect processing of the data or incorrect conclusions [1, 2]. In the era of machine learning and deep learning, the old adage of computer science that defines "garbage in, garbage out" gained renewed meaning and importance and the quality assessment and curation of the (imaging) data for AI and DL is said to take up to 80% of the data scientists' time [3, 4].

This chapter discusses the issue of data quality by looking at the process of curation of medical images and other related data and the different aspects that are involved in this when moving forward in the era of AI.

17.1 Introduction

When trying to answer questions about curation of medical images and data in the era of AI, one first has to answer the questions what the definition of artificial intelligence is. Different sources provide different answers to this question. Three often heard and read definitions are:

P. M. van Ooijen (✉)
University of Groningen, University Medical Center
Groningen, Groningen, The Netherlands
e-mail: p.m.a.van.ooijen@umcg.nl

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_17

1. Artificial intelligence is a computerized system that exhibits behavior that is commonly thought of as requiring intelligence.
2. Artificial intelligence is the science of making machines do things that would require intelligence if done by man.
3. AI is the science and engineering of making intelligent machines, especially intelligent computer programs.

In short, in machine or deep learning the algorithmic rules are no longer put into the system by a human observer, but the machine uses input data and known outcomes as training data to develop the algorithm. Therefore, data quality is a very important issue since the development of the algorithm is directly linked to the (quality of the) data collection used. Keep in mind that the results provided by such systems are always preliminary since every new bit of data entered into the learning system potentially alters the algorithm. Therefore, over time data also need to be of a constant high quality in order to avoid degradation of the algorithm because of newly arriving data and knowledge. This requires not only data collection quality but also a process of curation of collected data to increase the value and usability.

The University of Illinois' Graduate School of Library and Information Science defines data curation as "the active and ongoing management of data through its life cycle of interest and usefulness to scholarship, science, and education. Data curation activities *enable data discovery and retrieval, maintain its quality*, *add value*, and provide for *reuse over time*. This new field includes *authentication, archiving, management, preservation, retrieval*, and *representation*." [5, 6]. This data curation process is deemed a requirement to achieve an imaging biobank or data repository that is findable and reusable [7].

Current estimations suggest a doubling of the total amount of data in the world every 2–3 years [2]. Simultaneously, the percentage of the data collected digital instead of analogue increased dramatically in the past two decades. Although no fixed numbers over a long period of time are published, we can assume that similar increases in data have occurred in the past decades concerning medical imaging. In the nineties of the twentieth century, the digitalization of imaging commenced with the introduction of standardized data structure and communication with DICOM (Digital Imaging and Communication in Medicine) and the development of picture archiving and communication systems (PACS). These allowed a more convenient and standardized collection of the imaging data and also could guarantee the long-term storage and accessibility of the imaging data, provided a proper storage medium and migration strategy is employed [8]. The data increase itself was triggered by the evergrowing requirement for high-quality imaging data and mainly pushed forward by the developments in computer tomography (CT) and magnetic resonance imaging (MRI).

The increase in digital data collection also lowered the threshold to acquire data and thus allowed higher sampling frequency with more comprehensive data, thus further increasing the amount of data produced. These different factors have led to the collection of multi-TB PACS archives over the years with a variety of information per patient from different modalities, sequences, protocols, etc. Also, post-processed data obtained during the analysis and review from a variety of tools and workstations can be included in the patient data in the PACS as well as reports and other meta-information. When conducting retrospective data collection from such a PACS environment, the challenge is to include the relevant selection from the dataset acquired and generated that can be used for analysis and will lead to the required insight. What data to collect and at which frequency is still a human decision and thus prone to error, variation, and personal or institutional preferences. Because of this, the risk of collecting largely useless data collections is present. Often it is those types of collections of questionable quality that have to be used in artificial intelligence and deep learning.

Different machine learning and deep learning systems are developed both supervised and nonsupervised and new networks are being published frequently. Selection of the proper environment or network is therefore also part of the challenge of deep learning, and this selection should be adapted to the properties of the data collection used to train and test the network. Regardless of the system selected, the availability of an appropriate training dataset is vital [2]. The above demonstrates that in DL the quality of the dataset used is vital in every step of the development.

17.2 Data Discovery and Retrieval

As stressed before, data selected to be used as training input for artificial intelligence environments have to comply with high quality standards. The data need to be correct, have proper and validated labels, be accurate, be still "up to current standards," etc. However, even if the data are of high quality, it also needs to be of sufficient size since applying AI to too small datasets will not render significant findings because of lack of power. Therefore, the right data collection(s) must be found and if needed combined to obtain a sufficiently large amount of unbiased data including all possible variations [4, 9–11].

The discovery and retrieval of (imaging) data in health care has a dimension on its own in that it is almost always personal health data from an individual. This hampers the discovery and retrieval of (imaging) data because multiple factors have to be considered when collecting health-related retrospective data from the electronic medical record (EMR) or picture archiving and communication system (PACS) or when acquiring prospective data through clinical trials or population studies. In many instances, the tools to mine these clinical systems in a structured, meaningful, and easy fashion are lacking but required for obtaining the datasets useful and adequate to perform AI [3].

And then again, when the correct cohorts are identified and the required approvals are obtained, the variability in the data collection can be enormous. First of all, medical imaging equipment is far from standardized and imaging data from different hospitals using equipment from different vendors or the same vendor but different equipment generation or protocol used can be incomparable in their image presentation and diagnostic quality, not only because of the fast development of new equipment but also because of the (subtle) differences in the technical implementation and scan sequences used by the different vendors.

Furthermore, these sequences for specific clinical questions are also not standardized and will provide different images based on the local preferences of a certain department or even a specific radiologist. Also, the variety in the naming of the protocols used by different vendors (especially in MRI) is decreasing the quality of the data. Therefore, the development of guidelines to data acquisition and standardization of protocols is a requirement to allow the construction of large and above all useful data collections [9]. Additionally, the imaging data that are usually collected nowadays are based on the already processed data in the shape of DICOM images while the raw data from which these human interpretable images originate (e.g., the k-space data of MR and sinograms of CT) are not stored while these could be a valuable source, with possibly less variation, for computerized analysis [3, 12].

Another major question to consider is the ownership and control of the data. The legal perspective concerning this question is covered in another chapter of this book, but there is also a more practical question to consider. Where does the data reside? In health care, we have observed a slow movement from hospital-centric data model to a more patient-centric data model. This also means integration of new information in this patient-centric model through for example the Internet of Things (IoT) and wearables. Furthermore, open science and open data are increasingly advocated by governments and funding agencies resulting in large collections of data mostly available in the cloud establishing sandbox environments to be used by anyone to train and validate their software [11].

The risk of putting data into the cloud is that we are in a sense losing control of this data, and thus discovery and retrieval of relevant data is severely hampered by the fact that we upload all our health-related information into a variety of dispersed non-connected and non-standardized cloud solutions [13].

The question arising from all this is, even if enough high-quality data are collected, are we able to find (or discover) the right data. And if we find certain data, are we able and allowed to actually retrieve the information contained in the system and combine different sources unambiguously into one single dataset. If we are able to gather the information contained in those different databases, it might bring us the capability of data merging and such obtain a new linked dataset with much richer information. However, this could also have implications on the usability of the data since combining multiple datasets could infringe the privacy of the individual that could not be recognized in the separate datasets but is identifiable by the combined data through data linkage.

The legal obligation to protect the privacy of the patient or participant is one of the crucial things to take care of when collecting data in health care [10, 11]. Current methodologies for anonymization and de-identification are often suboptimal [14]. Furthermore, the anonymization or de-identification has to be performed such that the scientific research value of the data is retained in the de-identified dataset while still removing all personal health information [15]. Therefore, new algorithms should be developed to conceal identities effectively both protecting the individual privacy and still maintaining the full value of the same data for analysis. These three aspects of de-identification, privacy, and data value can work against each other with opposing requirements and struggle with variability in data content and lack of standardization, thus hampering the automation of this process. Current repositories of research data such as the TCIA [7, 16] still have a workflow in place where curators visually check and when needed correct every DICOM file (image and header) entered into their database to ensure data privacy and correct handling of the data.

One specific challenge here is the fact that DICOM headers may contain proprietary information that is not part of the standard DICOM but could include information on the acquisition or nature of the imaging data enclosed that is vital for adequate advanced (post)processing of the data. However, these private tags may also include references to personal health information (PHI) or other information that could infringe the privacy of the subject [12]. The same holds for the comment fields that are available in the DICOM header, content of these fields is free text, and their use is often depending on local conventions. This content may thus vary per hospital or even per modality within a hospital. Therefore, these fields could also contain PHI manually entered by a technician or radiologist.

Besides the header information included with the DICOM file, the actual image contents may also pose the risk of disclosing privacy sensitive information. For example, in secondary captures, topograms, and ultrasound examinations where in each of these exams sensitive information can be burned into the image, removing this information is possible but difficult to automate since the location at which the sensitive information is stored in the images may vary and can be difficult to detect automatically. Furthermore, so-called DICOM containers can also be constructed where the DICOM header is present but instead of an image another file type is included into the file such as a PDF file. These files could even be full patient reports with all PHI included. Another special kind of DICOM file that needs to be handled with care is the DICOM Structured Report (SR). The SR file typically holds the report of the radiologist describing the image review and conclusion and thus could also reveal sensitive information depending on local policies or the reporting method of the individual radiologist.

A final challenge that needs to be identified is the fact that facial features can easily be obtained from MR and CT datasets of the head. By performing surface or volume rendering reconstruction of those datasets, the face of the subject involved becomes visible. Studies have shown that facial recognition technology is able to combine these reconstructions with pictures from, for example, social media profiles to reveal the identity of the imaged subject [17, 18]. Especially since name tagging of pictures in modern-day social media directly links the person's name to the facial features.

The importance of careful curation of the data because of privacy risks has been reported on by different studies where the ability to breach the personal health information privacy was demonstrated on de-identified dataset. One example by Sweeney [19] shows that in 35 cases of an anonymized dataset obtained from a US hospital re-identification of the studies involved was possible by cross-linking to publicly available newspaper stories about hospital visits in the area.

17.3 Data Quality

The main reason for performing data curation is to increase the data quality of the data collection. However, an important consideration to start with is the question if the data quality is sufficient and useful for their application to artificial intelligence in the first place. It can be argued that when using big data an occasional bad sample or outlier will have little effect on the algorithm because of the large number of correct samples. However, the data richness also implies that the machine learning environment could use faulty inputs to determine the algorithm causing it to work on the training and testing dataset, but not in general use. One well-known example of AI and DL using suboptimal input datasets is a situation where the network is trained on a large multicenter database and with a test set performs adequate. However, at more careful inspection it is evident that the network is not trained to identify the pathology in the images but to recognize the features of a specific imaging device or hospital of origin because of unbalance with respect to the incidence of pathology in the dataset.

When looking at deep learning and machine learning as a system where the data together with the model are used to eventually come to a prediction, it is evident that the success of this system is not only depending on the quality of the model, but also on the quality of the data [20]. If the data, the model or both are of insufficient quality, the prediction will not be reliable.

The quality of the data used is thus essential for the validity of the outcome. A paper by Chalkidou et al. [21] showed that the current practice with data science and artificial intelligence leads to false discoveries because of fundamental flaws in the way the studies are performed. The issues that occur with those studies are a small sample size (12–72 cases, mean 44 cases) of often heterogenous cohorts, selection bias, and missing validation dataset (only 3/15 examined studies had a validation dataset).

There are multiple challenges defined that could negatively affect the quality of a dataset. These challenges are poor data collection practice, missing or incomplete values, non-standardized inconvenient storage solutions, intellectual property, security, and privacy [20, 22]. Assessing the quality of a dataset can therefore be challenging. To increase the use of data quality measures of datasets, multiple suggestions have been made to introduce some kind of data quality or maturity model. By assessing the dataset against such a model, the quality can be determined more objectively and possible use of the dataset is more evident.

One such a model for data quality was proposed by Lawrence [20]. He proposed to introduce a three-band model with subdivision into different levels per band. In this model, C4 would be the worst dataset and A1 the best. Band C would look at accessibility of the data. This could vary from C4 where the data might exist, but existence isn't even verified to C1 where data are collected in a standardized and known format and ready to be used without any constraints on the use. The next band, band B, would be about faithfulness and representation of the data. In this band, questions should be answered such as: Is the data that we got also what we expected? How are missing or incorrect values handled? What kind of encoding is used for the different data fields? How was the data collected? Is there bias in the dataset? Etc. In this band, the top quality would be B1 where we have a dataset that is C1 and where the limitations of the data are known to the user. Band A puts the data into the context. Here the ultimate question has to be answered if the dataset is appropriate to get to the correct prediction. It could be that in this phase expert annotation of the existing data or collection of additional data is required. Here level A1 would be curated data that are adjusted properly to allow getting the answer to the (clinical) question.

Based on the model by Lawrence, Harvey later introduced a version describing four data quality levels A–D more targeted to the medical domain [22]. The levels run from D where data are inaccessible, with unknown format and un-anonymized (current EMRs and PACSs in hospitals). In level C, anonymization is performed and ethical clearance obtained, but still the data are unstructured and show noise and gaps (EMR/PACS-based research collections). Level B introduces true representation with structured and visualizable data (structured and curated research collections). Finally, level A is a dataset containing contextual annotated and task ready data. According to Harvey, only A is AI usable data.

Although these kinds of models could be useful to categorize datasets for the purpose of machine learning, widespread application has not been established yet.

17.4 Adding Value

A report by EMC in 2014 [23] showed that in 2013 of the data collected in the global digital environment only 22% could be useful for analysis if—and only if—it would be properly tagged or characterized. However, they concluded that the tagging is mostly lacking in the collected data and that only 5% was valuable target-rich data. At that time they projected that in 2020 possible useful data would be increased to about 37% of all data collected with a doubling in target-rich data to about 10%.

In the case of medical imaging data, the proper annotation or tagging of the imaging data is also of vital importance [12] (level A of the model of Harvey described in the previous section). In order to train or validate AI and ML systems, a proper annotation is needed to define the ground truth that is used to learn and check results. However, no standardized syntax or method is available to collect the ground truth, and furthermore, the actual ground truth is difficult to obtain in most cases.

The two main standardized annotation methods are the Annotation and Image Markup (AIM) standard and the DICOM Presentation State (PS). AIM is developed within the National Cancer Informatics Program of the National Cancer Institute [10]. With AIM information is annotated and these annotations can be stored in a DICOM-compliant manner for later analysis. Although AIM is frequently reported to be used in research, it is not a widely accepted and used standard yet, and although DICOM PS is part of the globally accepted DICOM standard, it is still little used by software developers to report on annotations. Furthermore, clinically obtained annotations can in most cases not be used directly when performing AI because the annotations could contain personal health information which should not be present in research data. Therefore, annotations have to be redone when using the data for AI training and validation. Segmentation of the imaging data is even worse; no current widely accepted standard exists to store and communicate segmentation results between different tools from different vendors/sources.

The ground truth currently frequently used for training AI will result from a radiological report, a pathology examination report, or surgical reports. In this case, the value of the data on the "ground truth" relies both on the expertise of the observer describing the result, the accuracy of the description, and on the quality of the measurement methods. However, the accuracy of the results described by a physician is compromised by the fact that many reports are still free text without standardized lexicon or terminology resulting in multi-interpretable ambiguous reports with an abundance of synonyms. Furthermore, these different reports can even provide different measurements or conclusions and distinguishing which of these is the actual ground truth is a challenge that can often not be tackled. Natural language processing (NLP) could be a solution in situations where structured reporting and coding is not being used (as unfortunately still is the case in most hospitals).

17.5 Reuse Over Time

Part of the value of a dataset is the ability to use that dataset repeatedly over time. With the advent of bg data approaches, the data discovery and retrieval tend to shift from a targeted approach where specific data are collected to an approach where as much data as possible are gathered without a clear goal in mind because of possible future applications or novel insights that can be obtained [2]. When collecting data in this manner, assumptions have to be made on what data to collect and keep for future reference and use. Therefore, assessing the quality of this data collection is very cumbersome since the application of the data is still unknown. Furthermore, reuse also introduces other challenges and questions concerning the legal aspects of data privacy and intended use [2].

To allow reuse over time, the data should comply with the FAIR principle and be Findable, Accessible, Interoperable, and Reusable [24]. The FAIR guiding principles, that can be found in a table published by Wilkinson et al. at *(*https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/*)*, define the principles that should be met to obtain a FAIR data collection.

To achieve this, a proper IT infrastructure is required to store these data [10] including an accurate description and indexing of the data. Such systems are often described in terms of and referred to as imaging biobanks. Imaging biobanks are defined as IT systems holding relevant data and allowing interoperability between them in a federated set-up [25]. Currently, multiple (research) institutes, scientific organizations, and funding agencies are advocating the opening up of imaging data for reuse over time and designing and building environments to allow this. Examples are the Cancer Imaging Archive (CIA), the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), and the Osteoarthritis Initiative (OAI). Those archives or imaging biobanks contain collections of anonymized and curated (imaging) data that can be used for scientific purposes.

As an example, the CIA is a repository for cancer imaging and related information from the US National Cancer Institute [7]. With a content of over 30 million radiology images from over 37,500 subjects, it holds a wealth of information on cancer imaging. Data descriptions are used to categorize and organize the database into collections by tumor type. All data are manually curated and anonymized.

Although these initiatives exist, the need for better ways to construct FAIR data repositories is still prominently discussed, frequently also stressing the specific requirements to such datasets when they are to be used for machine and deep learning purposes [12].

17.6 Some Tools of the Trade

As in any application domain dealing with data, a vast number of tools are available to support in the different steps of the data curation process of medical data. There are tools for collection and anonymization of the data, for enrichment of the data, and for cleaning and curating the data. Without the illusion of being complete, Table 17.1 shows some examples of open source and freeware tools available for the different steps. When selecting tools to help you to obtain valid datasets, it is important that you select tools that are as simple as possible, and it might also help to restrict to using a defined set of tools within your research group or institution.

17.7 Conclusions

Only recently has data curation made the calendar of medical imaging research. Therefore, the understanding and role of data curation in the medical imaging domain is still limited. Often new research projects do not take into account the cost and manpower required to perform data curation either when collecting the data from the start (data curation “by design”) or when data are collected from existing sources and, if needed, combined. However, in order to obtain datasets that can be used for future purposes, obtaining high-quality data is obligatory and data curation should be a requirement.

Table 17.1 List of examples of freely available tools for data handling and curation

Tool	Purpose	Where to find
CTP	Data collection/anonymization	https://www.rsna.org/ctp.aspx
TextAnonHelper	Text anonymization	https://bitbucket.org/ukda/ukds.tools.textanonhelper/wiki/Home
DeFacer	Anonymization by removal of facial features	https://surfer.nmr.mgh.harvard.edu/fswiki/mri_deface
POSDA	Archival and Curation of DICOM datasets	https://github.com/UAMS-DBMI/PosdaTools [26]
OpenRefine	Data cleaning tool	http://openrefine.org/
Colectica for Excel	Excel extension for documentation	https://www.colectica.com/software/colecticaforexcel/
Open Clinica	Clinical Data Management tool	https://www.openclinica.com/
RedCap	Clinical Data Management tool	https://www.project-redcap.org/
XNAT	Platform to support imaging-based research	https://www.xnat.org/

References

1. Rosenstein BS, et al. How will big data improve clinical and basic research in radiation therapy? Int J Radiat Oncol. 2015;95:895–904.
2. Mayer-Schonberger V, Ingelsson E. Big data and medicine: a big deal? J Intern Med. 2017.
3. Ridley EL. How to develop deep-learning algorithms for radiology. AuntMinnie.com. 2017. https://www.auntminnie.com/index.aspx?sec=sup&sub=aic&pag=dis&ItemID=118078. Accessed 6 June 2018.
4. Redman TC. If your data is bad, your machine learning tools are useless. Harv Bus Rev. 2018. https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless. Accessed 6 June 2018.
5. U of Illinois. 2018. https://www.clir.org/initiatives-partnerships/data-curation/. Accessed 9 May 2018.
6. Freitas A, Curry E. Big data curation. In: Cavanillas JM, et al., editors. New horizons for a data-driven economy. Cham: Springer International Publishing; 2016.
7. Prior F, Smith K, Sharma A, Kirby J, Tarbox L, Clark K, Bennett W, Nolan T, Freymann J. Data descriptor: the public cancer radiology imaging collections of the Cancer Imaging Archive. Sci Data. 2017;4:170124.
8. van Ooijen PMA, Viddeleer AR, Meijer F, Oudkerk M. Accessibility of data backup on CD-R after 8 to 11 years. J Digit Imaging. 2010;23(1):95–9.
9. Aerts HJWL. Data science in radiology: a path forward. Clin Cancer Res. 2018;24(3):532–4.
10. Kansagra AP, Yu J-PJ, Chatterjee AR, Lenchik L, Chow DS, Prater AB, Yeh J, Doshi AM, Hawkins M, Heilbrun ME, Smith SE, Oselkin M, Gupta P, Ali S. Big data and the future of radiology informatics. Acad Radiol. 2016;23:30–42.
11. Tang A, Tam R, Cadrin-Chenevert A, Guest W, Chong J, Barfett J, Chepelev L, Cairns R, Michell R, Cicero MD, Gaudreau Poudrette M, Jaremko JL, Reinhold C, Gallix B, Gray B, Geis R. Canadian Association of Radiologists white paper on artificial intelligence in radiology. Can Assoc Radiol J. 2018;69:120–35.
12. Kohli M, Summers R, Geis R. Medical image data and datasets in the era of machine learning—whitepaper from the 2016 C-MIMI meeting dataset session. J Digit Imaging. 2017;30:392–9.
13. Lupton D. Who owns your personal health and medical data? This Sociological Life BLOG. 2015.
14. Aryanto KYE, Oudkerk M, van Ooijen PMA. Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. Eur Radiol. 2015;25(12):3685–95. https://doi.org/10.1007/s00330-015-3794-0.
15. Moore SM, et al. De-identification of medical images with retention of scientific research value. Radiographics. 2015;35:727–35.
16. Clark K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26:1045–57.
17. Prior FW, Brunsden B, Hildebolt C, et al. Facial recognition from volume rendered magnetic resonance imaging data. IEEE Trans Inf Technol Biomed. 2009;13(1):5–9.
18. Mazura JC, Juluru K, Chen JJ, Morgan TA, John M, Siegel EL. Facial recognition software success rate for the identification of 3D surface reconstructed facial images: implications for patient privacy and security. J Digit Imaging. 2012;25(3):347–51.
19. Sweeney L. Only you, your doctor, and many others may know. Technology Science. 2015. http://techscience.org/a/2015092903. Accessed 6 June 2018.
20. Lawrence ND. Data readiness levels. 2017. arXiv:1705.02245v1 [cs.DB].
21. Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. PLoS One. 2015;10:e0124165.

22. Harvey H. Is medical imaging data ready for Artificial Intelligence? AuntMinnieEurope. 2017. https://www.auntminnieeurope.com/index.aspx?sec=sup&sub=pac&pag=dis&ItemID=615032. Accessed 6 June 2018.
23. EMC. The digital universe of opportunities: rich data and the increasing value of the internet of things. Executive summary data growth, business opportunities, and the IT imperatives. EMC. 2014. https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm. Accessed 9 June 2018.
24. Wilkinson MD, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.
25. ESR. ESR position paper on imaging biobanks. Insights Imaging. 2015;6(4):403–10.
26. Bennett W, Metthews J, Bosch W. SU-GG-T-262: open-source tool for assessing variability in DICOM data. Med Phys. 2010;37:3245.

Does Future Society Need Legal Personhood for Robots and AI?

18

Robert van den Hoven van Genderen

18.1 A Paradigm Shift

Over history the human race has proved to adapt to environmental changes and if possible adapt the environment. Technology is to be maximized in application to serve humanity. Will technology be adapted to a further human evolution or will the humans evolve to the development of technology. Or will there be a continuous integration of both resulting in a cyborgic society, in the sense of human and AI integrated beings?

This paradigm shift will start with the evolution of human controlled robots and AI appliances toward an ever more autonomous system in a variety of applications, autonomous vehicles, and other forms of transport, social, financial, and economic services, industrial and production processes, and health and medical industry.

Rather than an expert human surgeon, one could already be operated by the Smart Tissue Autonomous Robot (STAR). In a recent set of experiments, STAR's inventors showed that it makes more precise cuts than expert surgeons and damages less of the surrounding tissue. The researchers presented their results at the recent robotics conference IROS 2017.[1]

Methods such as EEG and fMRI are in use for noninvasive and indirect forms of brain–computer interfaces by acquiring biological signals from outside the human body as well. These brain–computer interfaces are already applied as an aid for control and communication employed by paralyzed people by translating neural signals into command signals that can control devices. One example of this method is to record signals from the motor cortex, which can then efficiently be utilized to exert control over devices like a computer cursor. With advancing technology, these devices get faster and more and more approach the function of normal movements. Current research even takes it one step further by planning to use computers to send feedback information to the brain. Warwick used the same procedure to create

"Personhood" can be read as "legal personality".
This chapter has been based on former articles, insights and presentations by the author. The terms "robot" and "AI entity" are used interchangeably.

R. van den Hoven van Genderen (✉)
Center for Law, Internet and Intellectual Property Law at Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Switchlegal Lawyers, Amsterdam, The Netherlands
e-mail: rob.vandenhovenvangenderen@switchlegal.nl

[1] https://spectrum.ieee.org/the-human-os/biomedical/devices/in-fleshcutting-task-autonomous-robot-surgeon-beats-human-surgeons

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_18

a so-called biological robot.[2] While signals from the robot's sensors stimulate the neural tissue and output signals from the neural tissue drive the motors of the robot. Warwick also believes that by now it would be possible to grow a biological human brain, meaning a nervous system, and let it develop inside the robot. This raises the question of the potential consciousness of such a brain and the implications of this. Furthermore, humans could wish to donate their own neurons, but with today's possibilities, realizing a copy of part of the human brain in its functioning is not an option yet.

There is a common understanding that the development of robotics and artificial intelligence will have an immense effect on society and the behavior of humans without knowing what the disruptive results will look like. Moore's law states that new technology is being produced at an exponential rate but will we be capable to accept these changes within the same rate?

> As neuroscientist Moran Cerf debated why trying to master the brain is a bit of a catch-22: "If the human brain were so simple that we could understand it, we would be so simple that we couldn't."[3]

We are doing our best to understand our own brain but will take artificial intelligence to do the mapping of the brain and still will take a long time to fully understand both.[4] As with all new technology there are two theses that determine our view: fear and overestimation of human control. This fear i.e. resulted in the open letter of several scientists to the European Commission against the idea to give electronic legal personhood to AI entities.[5] Also the late Steve Hawking, Elon Musk, and others warned several times of the risk of artificial intelligence and call for controlled development of AI.[6] This development should be governed by a set of ethical and legal and operational principles.[7] There is a general fear that the unique human race will disappear after reaching "singularity."[8] Although Sofia, the lifelike robot from Hanson Robotics, received citizenship from Saudi Arabia in the autumn of 2017 as a PR stunt, there is the wrong assumption that the embedding of AI and robotics in our legal system will lead to giving human rights to robots. This is based on the idea that a legal model should be based on the existing legal position of the natural person.[9] Also, the warnings for the development of autonomous systems creating "killer robots" by several scholars and leaders of industry create a feeling of distress and fear. There is a danger that stopping all further innovation of AI is hampering further development of mankind. Should we seek to stop this technological evolution? Or, is it acceptable to integrate AI fully in our society and also in our legal system?

New technological developments have a risk: there is a fear of the negative consequences. With the development of robotics and AI, a new paradigm shift will be there. AI and robots will be used not just as tools, but more and more as a replacement for people. To date, it is uncommon for robots or AI to completely remove people from the equation, but as businesses become more reliant on the use of robots and AI, the human becomes a smaller part of the economic model: loss of jobs, loss of control, and ultimately fear for the future of mankind. We also have the ten-

[2] Warwick K., et al. (2004). "Thought Communication and Control: A First Step Using Radiotelegraphy." *IEE Proceedings on Communications* 151(3):185–189.

[3] https://waitbutwhy.com/2017/04/neuralink.html

[4] i.e.: https://www.nature.com/news/worldwide-brain-mapping-project-sparks-excitement-and-concern-1.20658; and, https://www.humanbrainproject.eu/en/

[5] https://bit.ly/2xfMToe

[6] https://futureoflife.org/ai-principles/

[7] ibidem.

[8] The acceleration of technological progress has been the central feature of this century. We are on the edge of changes comparable to the rise of human life on Earth. The precise cause of this change is the imminent creation by technology of entities with greater [intellectual capacity] than human intelligence. See Vinge [1].

[9] a. A legal status for a robot can't derive from the Natural Person model, (. . .) since the robot would then hold human rights, such as the right to dignity, the right to its integrity, the right to remuneration or the right to citizenship, thus directly confronting the Human rights. This would be in contradiction with the Charter of Fundamental Rights of the European Union and the Convention for the Protection of Human Rights and Fundamental Freedoms, https://bit.ly/2xfMToe

dency to accentuate the negative aspects of any new technology, certainly when we do not fully understand the technology and its consequences. As Sir Arthur Clarke stated in his novel *Profiles of the Future: An Inquiry into the Limits of the Possible* his so-called third law: any sufficiently advanced technology is indistinguishable from magic.[10] And magic is incomprehensible and, therefore, dangerous, certainly as we all are the sorcerer's apprentice. As Steven Hawking, Elon Musk, and others have warned us:

> AI is the "biggest risk we face as a civilization" and "AI is a rare case where we need to be proactive in regulation instead of reactive because if we are reactive in AI regulation it's too late, AI is a fundamental risk to the existence of civilization . . . as a whole."[11]

Is the development of AI so special that the legal system has to be adapted? The origin of the law, although possibly influenced by technological developments, is to regulate society by a normative structure. Until now, law has been developed by humans, for humans, and—initially—to govern the relations between natural persons and, later on, artificial legal persons. But many things have changed during the historical development of the law, in the long journey from the Roman legal system to our modern legal system. New technologies will change society and will reflect on the change of this legal framework. As Lauren Burkhart citing Clark A. Miller and Ira Bennett on "reflexive governance" observes we better be prepared to have an open mind for changes in technology by "identifying not only what gadgets might arise but also how gadgets intersect in society, with one another and with people, how people identify with, make use of, oppose, reject, apply, transform, or ignore [technologies]."[12]

To what level society must adapt to technological innovations has to be based on the needs of that society, be it economic or social needs. If a sentient entity, in the sense of possible autonomous intelligent agency in robotics and other AI systems, now, or in the near future, could be expected to act with legal effect, that is to say perform tasks with legal consequences, it could be desirable to adapt the legal framework accordingly. This decision, however, should be based on the assumption that an AI robotized society will benefit from—to a certain degree—the legal personality of robots. Legal scholars are generally hesitant to adapt the law on the basis of technological changes. But "if the facts too long deviate from the legal status and the right is unsustainable, the law must ultimately yield to the actual situation."[13] Already society has undergone changes as a result of this development. Semi-autonomous cars are now a point of legal, moral, and social discussion because the central subject in traffic laws is the driver and their control over the vehicle is a requirement for safety on the road. This gives rise to a question that is not new, nor solely legal: a question that was already described by Geldart in a discipline overruling way:

> The question is at bottom not one on which law and legal conceptions have the only or the final voice: it is one which law shares with other sciences: political science, ethics, psychology, and metaphysics.[14]

It is of the utmost importance to also consider ethical values and fundamental rights issues in the possible decision to give a certain legal status to robots. Neil Richards and Jonathan King's statement in their paper on Big Data ethics could well be applied to robotics:

> We are building a new digital society, and the values we build or fail to build into our new digital structures will define us. Critically, if we fail to balance the human values that we care about, like privacy, confidentiality, transparency, identity and free choice with the compelling uses of Big Data, our Big Data Society risks abandoning these values for the sake of innovation and expediency.[15]

[10] Clarke [2], p. 14.

[11] Titcomb (2017) AI is the biggest risk we face as a civilization, Elon Musk says. Available at: http://www.telegraph.co.uk/technology/2017/07/17/ai-biggest-risk-face-civilisation-elon-musk-says/. Accessed 11 October 2017.

[12] Lauren Burkhart citing Miller and Bennett [3]; Burkhart [4].

[13] Tjong Tjin Tai [5], p. 248.

[14] Geldart [6], p. 94.

[15] Richards and King [7], p. 394.

This is certainly the case if we neglect these values if also other technological developments going beyond digitalization are creeping up to us. There is an increase in research on brain–computer interfaces and biotechnology. The integration of artificial enhancements of human mind and body will create other social as well as ethical and legal questions by creating new schisms in society. But these ethical questions are also tendentious. Giving robots this kind of legal personhood would raise an ethical question because the human dignity is at stake. Giving robots legal personhood diminishes the self-esteem of humans. Humanity tends to place itself above all other living beings on top of the food and brain chain. Accepting ultra-intelligent beings as part of our society would possibly destroy this dignity.

Still a positive attitude toward AI developments is the way forward. The house of Lords of the UK announced this notion in its reaction to the House of Lords Select Committee report on Artificial Intelligence: "AI in the UK: ready, willing and able?",[16] as follows:

> The Government recognises the importance of artificial intelligence to the UK's economy, its businesses, public services, its workers and consumers. As an enabling and exponential group of technologies, AI can drive efficiencies, boost productivity and accelerate innovation. It is key to realising the ambitions of the Industrial Strategy and ensuring the UK is at the forefront of existing and future industries.[17]

Maybe this conviction will save the UK from the negative influence of the Brexit.

18.2 Legal Position

The legal structure of our society is the result of cultural, social, economic, and ethical convictions laid down in norms as an artificial layer to conduct all activities of actors within this arena we call society. Within our legal system natural persons and legal persons have, for a long time, been the key players. Large and small businesses, private organizations, and government organizations are entitled to perform tasks with legal effect and can be held responsible for the things they do, having legal personhood. The perception about legal personhood evolves within culture and time. In the Middle Ages, for instance, animals could also be held responsible for their acts.[18] Technological development develops in the direction of artificially intelligent programs possibly embodied in all kinds of physical instruments and a variety of robotic entities in more or less anthropomorphic shapes that can perform a variety of tasks. Coupled with the exponentially expanded Internet, decision making by these AI entities with legal consequences is approaching. The consideration whether an autonomously functioning artificial intelligent entity or robot must have a certain legal subjectivity or not will be dependent upon social and economic necessities and, not least of all, the cultural social and legal acceptance by other actors in the societal arena. The acceptance will be also dependent on the choice for a utilitarian or more ethical, right-oriented framework. In other words, can a future society function without any form of legal personality for autonomous, artificially intelligent entities or is it a "conditio sine qua non?"

It is important to consider what kind of reasoning will be applied to the determination of the legal status of robotics. This status could be built on an augmented layer of required legal elements based on the continuous development of autonomy and intelligence of the robot. Or one could analyze the characteristics of the current players with legal personality and select which elements will be desirable to give robots that degree of legal personality that is considered useful in society.

Cautious proposals are already being made to comply with the future and to find legal solutions. However, the actual legal implications of an AI integrated society are set aside. Although the European Parliament accepted a motion on the civil law aspects of the development of AI generated

[16] https://www.parliament.uk/documents/lords-committees/Artificial-Intelligence/AI-Written-Evidence-Volume.pdf

[17] https://www.parliament.uk/documents/lords-committees/Artificial-Intelligence/AI-Government-Response.pdf

[18] Berriat Saint-Prix [8].

robotics, in creating electronic legal personhood, it is at a rather high level of abstraction:

> 59 f) creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently.[19]

In this motion, though, the essence is recognized: legal interaction with other parties. The orientation on electronic (legal) persons though is limiting the possibility of application of other future technologies as biotechnological constructs. Even in this case, legal orientation is not seeing beyond the nearest technological manifestations.

18.3 AI and Robots as Actor

For an analysis of the legal positioning of robots and AI, we cannot escape defining or describing these phenomena. Of course, there are several definitions developed by scientists and lawyers. For the sake of clarity, this chapter will not delve into all of these conceptions. There are a range of robots varying from the simple one task-oriented industrial robot to the autonomous car and the anthropomorphic robot companion. Bertolini defined a robot in a broad sense, encompassing this wide variety of robotics and AI entities as follows:

> [A] machine, which (i) may be either provided of a physical body, allowing it to interact with the external world, or rather have an intangible nature—such as a software or program,—(ii) which in its functioning is alternatively directly controlled or simply supervised by a human being, or may even act autonomously in order to (iii) perform tasks, which present different degrees of complexity (repetitive or not) and may entail the adoption of not predetermined choices among possible alternatives, yet aimed at attaining a result or provide information for further judgment, as so determined by its user, creator or programmer, (iv) including but not limited to the modification of the external environment, and which in so doing may (v) interact and cooperate with humans in various forms and degrees.[20]

Legal scholars differ in observation of the rules that would apply to AI and robots and the need for a separate legal qualification of AI in society. In determining the need for the legal personhood of AI entities, it should be taken into account that these systems will clearly vary in function. There will be obvious differences in the degree of autonomy resulting in a variety of legal requirements dependent on a social need to have robots perform tasks as more or less autonomous acts.

For the possible legal analysis and classification of robots, it is required to look at (1) the embodiment or nature of the robot, (2) the degree of autonomy, (3) the function of the robot, (4) the environment, and (5) the nature of the interaction between human and robot.[21] As always we will create rules on the basis of this new technological development as we for example did after the invention of the airplane, in creating air traffic rules. The difference is that there is an increasing autonomy in the action of the AI entity that does not fit in the existing legal framework. And of course, there is a pre-judicial question if humankind will accept the more utilitarian vision on seeing legal personality as just a means to serve society or the more ethical rights-oriented deontological orientation.

[19] Whereas it is of vital importance for the legislature to consider all legal implications. All the more now that humankind stands on the threshold of an era in which ever more sophisticated robots, bots, androids, and other manifestations of AI seem poised to unleash a new industrial revolution that is likely to leave no stratum of society untouched; report Delvaux with recommendations to the Commission on Civil Law Rules on Robotics [9]; European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))<A8-0005/2017>.

[20] Bertolini [10], p. 219. Compare also definition by "robotpark": "A robot is a mechanical or virtual artificial agent (called "Bot"), usually an electro-mechanical machine that is guided by a computer program or electronic circuitry. Robots can be autonomous, semi-autonomous or remotely controlled and range from humanoids such as ASIMO and TOPIO, to nanorobots, "swarm" robots and industrial robots. A robot may convey a sense of intelligence or thought of its own."

[21] Bertolini [10].

On the basis of these considerations, we can formulate the following questions:

1. Is there a need for a framework for AI and robot law in the sense of a law relating to, or as a result of, the use of robot technology in society? And, if so, what are the preconditions for establishing such a law in our legal system?
2. Does the robot need a certain degree of legal personhood that does not yet exist in positive law and is it necessary to regulate that degree of legal personhood? And, if so;
3. Is there a "gradation" of legal embodiment that connects with existing forms of legal personality or is a sui generis construction desirable taking into account the variability of AI systems and robotics?

18.4 Legal Subjectivity

Legal subjectivity, legal personality, or legal personhood is a condition that is attributed to a certain entity to perform within the legal structure of society. Before looking further into the question of what legal personhood would mean for an autonomous robot, one has to consider the existing quality of legal personality, or in other words, what does it mean to be considered a legal person. The technical legal meaning of being a legal person is in a simplified version: *"a subject of legal rights and duties."*[22]

This does not necessarily refer to humans as "natural persons." The idea of legal personhood involves the status of an entity as a person before the law, leading to recognition of certain rights and obligations under the law. Consequently, a legal person has the duty to obey the law, while enjoying the benefit of protections to rights and privileges accorded to a legal person.

In most legal systems, legal personhood can be understood as being capable of having legal rights and duties and legal capacity within a legal system, to act with legal effect such as to enter into contracts, to be liable, to be a subject of legal remedies. A legal (artificial) person is considered equal to a natural person, as far as property law is concerned, unless the law explicitly states the contrary.[23]

The legal construct of personhood in the law, however, operates as a bundle of fundamental assumptions involving the biological understanding of human beings, the understanding of an entity as a rational agent, and the existence of consciousness when it concerns natural persons.[24]

Still epiphenomenalists tend to believe that consciousness is a type of illusion that exists without any causal influence, even though there is no explanation as to how such an illusion could form in a deterministic world. Others have argued that consciousness is an impossible concept for us to understand or explain, simply because we humans can't conceive it fully.[25]

The overlap of the assumptions of the concept of consciousness and the relative priority accorded to each assumption is continuously evolving to accommodate new issues arising time, place, and culture. For instance, human slaves in the Roman Empire, as well as in later centuries, were not considered human beings for a long time, nor did they have human rights. They had the possibility of peculium though, to have and hold a certain amount of property as their own private property that their masters allowed them to spend or use as their own. Still, they were considered to be property: legal objects, as the subject of that could be bought or sold. So, we see that in that time legal objects and legal subjects could coincide. In the USA, on the other hand, slaves could be punished for criminal acts so as to exclude the criminal liability of their masters.[26] This is comparable with the treatment of animals in criminal law as was common in Europe in

[22] Solum [11], pp. 1238–1239.

[23] Dutch Civil Code (Burgerlijk Wetboek, BW), Book 2, Article 1 and 2.

[24] Ohlin [12], p. 210.

[25] https://plato.stanford.edu/entries/epiphenomenalism/

[26] American law was inconsistent in its constitution of the personality of slaves. While they were denied many of the rights of "persons" or "citizens," they were still held responsible for their crimes which meant that they were persons to the extent that they were criminally accountable. The variable status of American slaves is discussed in Fagundes [13]; Naffine [14], p. 346.

the Middle Ages as will be described below. Among humans, so-called natural persons, there has always been a difference in the contents of the legal capacity of legal personality. Also in "modern times" there existed, and still exists, legal discrimination among natural persons. Until recently, for example, women in all Western societies were not considered to have comparable legal capacities as their male counterparts. Until 1957, married women in the Netherlands could not perform legal acts without the consent of their husbands.

Changes continue to take place regarding the legal status of minors and their capacity to perform activities with legal capacity. Rights based on age or gender to drive cars, vote, buy weapons, or marry vary per culture, time, and place.

In addition, society has allowed the creation of artificial business entities such as the corporation, firm, or foundation, based on the necessity that these entities have to have the power and legal status to perform economic acts with legal consequences and have to have legal credibility. The legal person is also referred to as the metaphoric expression for a dogmatic fiction.[27] But it is a fiction that proved to be very useful. But it will be a continuous search how to apply its usefulness.

In our present society, we have discussions and even legal actions to consider personhood for animals. There have also been recent actions granting personhood to inanimate objects such as the Whanganui River in New Zealand and several rivers in India, suggesting that the scope of the legal construct of personhood may be expanding if the need arises.[28]

Whether an entity should be considered a legal person depends on the following question: should this entity be made subject of a specific set of legal rights and duties? The answer depends upon the cultural, economic, and political circumstances. There is considerable confusion about this central legal question, as well as deep intellectual divisions.[29] Legal personhood can be considered for humans, animals, or inanimate objects if you think of law from an essentialist perspective, as an artificial pragmatic construct, meant to service society. Of course, this also applies to legal objects and all norms translated in laws by humans. Or one could choose the concept of comparatism in the sense of Cartesian dualism. This would entail separating the concepts of legal personhood and legal objects on the basis of their characteristics as consciousness, matter, will, etc. To compare these concepts, one could take the common characteristics to find the most applicable legal status for different manifestations of robots or AI-driven systems. But, in addition, a complete dualistic principle of the concept of legal personhood is possible based on the utilitarian functional requirement of legal capacity of the entity concerned as was the case with artificial personhood.

Legal personhood is a flexible and changeable aspect of the legal system. As stated by international lawyer, Ian Brownlie, it is well recognized that the subjects of law in any legal system are not exactly identical in their nature and rights or in the extent of their rights and nature, depending on the needs of the community.[30] And certainly in international law, the recognition of the responsibility as a legal subject varies and often is used to protect the "legal subject" state to push the other legal subject in front of them:

> There is no international criminal law which applies to states as accused, but there is an increasing body of rules, administered in part by international tribunals, which subjects the conduct of individuals (potentially including state officials) to international criminal law. These developments, particularly in the field of human rights, have added another category of personality (albeit heavily qualified) to those within the international legal system, namely individuals and sometimes corporations created by national law.[31]

Specifically, in international law it is recognized that the scope of legal personality is measured by the need of society under different circumstances.[32]

[27] Maximilian Koessler, The Person in Imagination or Persona Ficta of the Corporation, 9 La. L. Rev. (1949).

[28] Hutchinson [15].

[29] Naffine [14], p. 346.

[30] Brownlie [16], p. 58.

[31] Crawford [17], p. 17.

[32] "All that can be said is that an entity of a type recognized by customary law as capable of possessing rights and

To see if and how AI-driven entities as autonomous robots need legal personhood, it is helpful to compare them with the bearers of legal rights and obligations that currently exist in our society, namely, natural and artificial legal persons. For a more essentialist utilitarian vision, though, it is necessary to look at the bare necessity that is essential to the function of the autonomous robot in a more metaphysical way. This is the artificial legal layer—a legal fiction—that can be applied or taken away in the sense of a construct of the character of *Alice's Adventures in Wonderland*, the "Cheshire Cat," the non-existing entity that can be there if one needs it and vanishes when superfluous.[33]

18.5 Humans as (Natural) Legal Persons

Does the human mind control the rationality of decision making and therefore can be trusted to make "computational" decisions as is the point of view of cognitive sciences? The sentient and conscious characteristics of human beings that are often declared essential to the legal conception of natural persons separate the natural person from the legal persons. To identify which aspects of legal personality might apply to AI entities as autonomous functioning robots, an explanation of the relevant characteristics of natural and unnatural legal persons can be helpful. Legally, the individual as a natural person is the bearer of rights and obligations due to the fact that it concerns a living person and not a fictional entity.

However, there is some agreement on what is characteristic of the individual: each individual differs from the other in the physical sense, but in a legal sense, each man of flesh and blood is the bearer of rights and obligations.

From a historical perspective, we can look at the concept of person and personhood as defined by Thomas Hobbes in his famous work Leviathan. According to Hobbes, a person is:

> He whose words or actions are considered, either as his own, or as representing the words or actions of another man, or of any other thing to whom they are attributed, whether truly or by fiction.
>
> When they are considered as his own, then is he called a natural person: and when they are considered as representing the words and actions of another, then is he a feigned or artificial person.[34]

Hobbes explains the origin of the word coming from the Latin "persona" and the Greek "prosperon," a mask used in theaters.[35] Still the Romans reserve this "persona" phenomenon to living (natural) humans, including women and slaves.

Hobbes separates the phenomenon of legal personality for nonhuman actors from artificial legal persons; if the person does not speak for himself, but their action or representation is attributed, one can speak of artificial personality. Hobbes' concept does not necessarily imply that this must be a human. Of course, he did not account for autonomous robots but he might well have considered this if he had been confronted with autonomous and maybe sentient robots. As referred to by Pagallo, the idea that a legal subject can be an "artificial person" should be traced back to the notion of "persona ficta et rapraesentata" developed by the experts of Canon Law since the thirteenth century. And Thomas Hobbes' Leviathan has thus a precedent in the work of Bartolus de Saxoferrato (1313–1357).[36]

duties and of bringing and being subjected to international claims is a legal person. If the latter condition is not satisfied, the entity concerned may have legal personality of a very restricted kind, dependent on the agreement or acquiescence of recognized legal persons and opposable on the international plane only to those agreeing or acquiescent." Crawford [17], p. 117.

[33] Naffine [14].

[34] Hobbes [18].

[35] The word "person" is Latin, instead whereof the Greeks have "prosopon," which signifies the face, as "persona" in Latin signifies the disguise, or outward appearance of a man, counterfeited on the stage; and sometimes more particularly that part of it which disguiseth the face, as a mask or vizard: and from the stage hath been translated to any representer of speech and action, as well in tribunals as theatres. Text Hobbes [18].

[36] In his commentary on *Digestum Novum* (48, 19; ed. 1996), Bartolus reckons that an artificial person is not really a person and, still, this fiction stands in the name of the truth, so that we, the jurists, establish it: "*universitas proprie non est persona; tamen hoc est fictum pro vero, sicut ponimus nos iuristae*." This idea triumphs with legal

Another feature of the natural person is found in the spiritual aspect of the natural person. In religious scriptures, one often finds references to the presence of the soul. Aristotle declared that all living entities had a form of soul: plants, animals, and humans. The difference was that plant's soul was a vegetative soul, only directed[37] on reproduction and growth; animals had a sensitive soul, directed on mobility and sensation; and humans have a rational soul capable of thinking, planning, and reflection.

Homer spoke of soul only in the case of human beings, in sixth- and fifth-century usage soul is attributed to every kind of living thing. What is in place, then, at this time is the notion that soul is what distinguishes that which is alive from that which is not. The adjective "ensouled" [*empsuchos*] as the standard word meaning "alive" was applied not just to human beings, but to other living things as well.[38]

Artificial legal persons and objects are considered not to have a soul. According to the catechism of the Catholic Church, "soul" means the spiritual principle in man. The soul is the subject of human consciousness and freedom.

The freedom of decision is the ethical and legal background of the responsibility we have as natural beings. Individuals are sovereign in their decisions and therefore legally responsible for their actions. So the concept of soul is gradually evolving to moral consciousness based on free will, spiritual sovereignty. Neuroscience nowadays though considers consciousness as a narrative that incorporates our senses by neural actions, how we perceive the world, and everything we do. But even within that definition, neuroscientists still research why we are conscious and how best to define it in terms of neural activity.[39]

Jean Bodin claimed that sovereignty, not specifically relating to consciousness, must reside in a single individual. This sovereignty can be transferred to other "legal entities, i.e., the state, a company, or any other organizational unit" that is recognized by law in the specific legal system. These legal entities must be considered "legal entities" with the power to make decisions with legal effects.[40]

The important question is whether independent, technical, and electronic instruments, combinations of hardware and software or algorithms, can be considered as bearers of rights; whether these might be vested with the power to act as legal entities and thus can perform legal acts, or whenever they are mandated to produce such acts. Their actions could also lead to liability that is not directly traceable to any other responsible body as is the case with employees, children, and animals. Or will there always be a natural individual behind the acting entity as the ultimate bearer of the rights and legal responsibilities?

An individual will always be a legal entity with legal personality but a legal entity; an artificial entity will not have the same rights as a natural person. The legal entity, being a natural person, the subject of rights and duties, can act with legal implications. There is no question whether one of these natural persons is fictitious or natural. The natural person is the human of flesh and blood. But inherent to this person is that he is able to function socially and, if legally competent, able to perform acts with legal consequences.

Natural persons can vote for other individuals in elections and be elected to represent other individuals. They may join a political party or a church. They will be the subject of human rights, the right to life, privacy, freedom of expression, right to education, and freedom of religion. Individuals may be put in prison if convicted for

positivism and formalism in the mid-nineteenth century. In the *System of Modern Roman Law* (1840–1849) ed. (1979), Friedrich August von Savigny claims that "only human fellows properly have rights and duties of their own, even though it is in the power of the law to grant such rights of personhood to anything, e.g., business corporations, governments, ships in maritime law, and so forth." The same line of thought is stated in Pagallo [19], p. 156.

[37] Aristotle, de Anima.

[38] https://plato.stanford.edu/entries/ancient-soul/

[39] https://digitalcommons.law.lsu.edu/cgi/viewcontent.cgi?article=1615&context=lalrev

[40] Bodin [20].

a felony. Individuals can marry another person or enter into a civil partnership. They may have children by natural birth from other individuals and will have an automatic natural and legal relationship. Yet also this might change overtime. Due to biotechnology, individuals can also be fully or partially naturally inseminated, derived from insemination with sperm or ova from a third party. It is even possible that children are the result of a DNA merging from three different individuals.[41]

For the time being, at least, this has no legal consequences. But maybe in future the boundary between natural and non-natural persons will not be that clear anymore. This could also have legal consequences that will also relate to a schism in the manifestation in being subjected to fundamental rights.

The law and legal opinions may not give an answer or have a final say on these questions. This is the terrain that legal science shares with other sciences: political science, medicine, ethics, psychology and metaphysics.[42]

Meanwhile, the question remains what features are relevant to determine what a real or natural person is? Biotechnology and AI are converging. Artificial limbs and organs are already integrated in the human body. Also, several experimental couplings of the brain to the Internet of Things has already occurred.[43] Bioengineering is developing at an incredible pace.

We already discriminate on the basis of free will and intelligence. If an individual is not mentally able to independently perform legal acts, he is placed under curatorship, and if the individual is not legally defined as an adult (in the Netherlands and other countries, 18 years), natural persons are not able to perform acts with legal consequences. This is not an absolute rule. Minors and adults under guardianship can buy a sandwich, ice cream, or even a bicycle, but will not be able to buy a car or a house. Their parents or trustees have a duty to support them and to represent them. There is a transition period between full responsibility for the actions of children and adulthood, which usually begins between 14 and 16 years old, and in China even from 10 years onward. The parent or guardian is not liable if he is not at fault for a harmful act by the child. But even within this system there are cultural and national differences. The age of full legal capacity is well established in the Netherlands and the USA at 18 years of age. However, an “adult” person in the USA is not allowed to buy alcoholic beverages, but is allowed to drive a car at the age of 16 or can purchase a fire arm, as referred to above. In many countries in Africa and Asia, for instance, there is no minimum age set for marriage. India recently had the maturity and judgment limit lowered to 16 years of age for the perpetrators of a crime. Thus, the law is far from consistent, not even nationally and certainly not in an international context. Legal standards are not equal for natural persons. There is also a tendency to look at the quality of the psychological capacity of natural persons. An example is the (not accepted) proposal to forbid women to have children when the parents are apparently not able to raise their children adequately, for example if they already have children expelled from home to external care.[44]

Furthermore, reference may be made to the historical context in the perspective of the standards relating to the legal capacity of natural persons. Time and culture varies with the legal status of natural persons. The abolition of slavery and, therefore, the abolition of the (partial) status as legal object only took place in 1794 in France, to be renewed by Napoleon in 1802, finally abolished in 1841 in 1838 in the UK after the abolition Act of 1833. The Netherlands and the USA finally accepted the abolition of slavery in

[41] Hamzelou (2016) Exclusive: World’s first baby born with new “3 parent” technique. Available at: https://www.newscientist.com/article/2107219-exclusive-worlds-first-baby-born-with-new-3-parent-technique/. Accessed 11 October 2017.

[42] Geldart [6], p. 94; Dewey [21], p. 655.

[43] Brainternet works by converting electroencephalogram (EEG) signals (brain waves) in an open source brain live stream. Minors [22] Can you read my mind? Available at: https://www.wits.ac.za/news/latest-news/research-news/2017/2017-09/can-you-read-my-mind. Accessed 11 October 2017.

[44] Proposal Ira Winds, Livable Rotterdam alderman.

1863. Nevertheless, there are people still living and working under “slave-like” circumstances.[45]

Women only got their democratic voting rights across the Western world at the beginning of the twentieth century. Until the abolition of the law on incapacity on 14 June 1956, married women in the Netherlands were legally incapacitated.[46] Belgium maintained this rule until April 1958. However, until 1971, the Dutch Civil Code stipulated that the man was the “head of the family” and that the woman owed him obedience. To increase the complexity about legal positions, we may also refer to the fact that a distinction is made in the rights of individuals as such. Same-sex marriages are still not allowed in a majority of countries. In conclusion, it can be established that the actual content of the legal status of individuals is not homogeneous. The legal status of natural persons is not manifest and is dependent on time as well as social-cultural circumstances. This point of view can also be applied to the legal characterization of the robot.

18.5.1 Human-Like Behavior as Determination for Legal Personhood

Some legal scholars argue that legal personhood should be limited to human beings or at least to serve the legal system that is construed by and used for the benefit of human beings. Their fear is that *extending the class of legal persons can come at the expense of the interests of those already within it.*

Free Will

Another consideration that is used to obtain the qualification of a natural person is the existence of free will, a basic element of real autonomy. This free will concept though in a legal sense has to be bound by the norms of ethics and morality. The idea of qualifying an autonomous thinking and self-decisive robot as an individual based on the autonomy and free will is a fairly extensive one. Free will, as indicated by Descartes, is based on the fact that we, as human beings, have the experience by which free will steers our behavior. Aristotle had the conviction that this free will also exists within animals.[47] And is not our “free will” determined by circumstances, history, and genes? And are we conscious of this free will? Is that consciousness? According to Shaun Nichols in an article in the *Scientific American*, it is just a series bioelectric signals, not more, referring to neurons firing in certain brain areas, no more and no less.[48]

Intelligence

For autonomous thinking there is also the need for intelligence. This aspect is also often used to determine the humanlike behavior, needed to determine the determination of a human and therefore a natural person.

The problem is that the concept of intelligence is not very extensively defined due to the different concepts of intelligence, i.e., rational intelligence and social intelligence. Howard Gardner theorized that there are multiple intelligences comprised of nine components: naturalist, existential, musical, logical–mathematical, bodily–kinesthetic, linguistic, spatial, interpersonal, and intrapersonal intelligence.[49]

David Wechsler formulated in 1955 a well-known general definition of intelligence: “The aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment.”[50]

Without going into the different theories that exist about the many forms of intelligence, I would limit this reference to the intelligence

[45] Aziz and Hussain (2014) Qatar’s Showcase of Shame. Available at: https://www.nytimes.com/2014/01/06/opinion/qatars-showcase-of-shame.html?_r=0. Accessed 12 October 2017; The Global Slavery Index [23] https://www.globalslaveryindex.org/findings/. Accessed 12 October 2017.

[46] On June 14, 1956, the House settled the bill by Minister JC Furnace, so that married women were legally competent as from January 1, 1957.

[47] Descartes [24].

[48] Shaun Nichols, Is free will an illusion? https://www.scientificamerican.com/article/is-free-will-an-illusion/

[49] Gardner [25].

[50] Wechsler [26].

needed to participate as an individual in society. To this end, it is necessary that there is understanding of the consequences of acts performed in this social traffic (with legal effect). When AI entities will be capable, now or in the near future, to meet the Turing test, the qualification for intelligence on a "human" level will not mean that AI entities will be comparable with humans on all levels.[51]

Intelligence and Animals, No Condition for "Human Rights"

Compliance with this test, something that other animal primates certainly cannot meet, gives the impression one has to do with a human being. Yet there are regular attempts to give these other primates a form of legal personality. The chimpanzee, an entity that is regarded as reasonably intelligent, was subject in the appeal court in New York on an appeal to personal liberty (Habeas Corpus).[52] It must be confirmed in a legal rule that a person cannot be kept in confinement unless it is decided by a court of law. This right, though, is reserved for natural persons. The status as a natural person was not accepted. The court stated that chimpanzees, although cognitively complex, are not entitled to the same legal status as human beings: "We conclude that a chimpanzee is not a 'person' entitled to the rights and protections afforded by the writ of habeas corpus."[53]

Only people can have rights, the court states, because only people can be held legally accountable for their actions. "In our view, it is this incapability to bear any legal responsibilities and societal duties that renders it inappropriate to confer upon the chimpanzees legal rights" ... that have been afforded to human beings.[54] On the other hand, the court also states that: "the classification of a being or entity as a "person" is made solely for the purpose of facilitating determinations about the attachment of legal rights and duties";

The Nonhuman Rights Project, the appellant in this case, did not agree with the ultimate conclusion of the court and stated:

> The Court ignores the fact that the common law is supposed to change in light of new scientific discoveries, changing experiences, and changing ideas of what is right or wrong; it is time for the common law to recognize that these facts are sufficient for the establishment of personhood for the purpose of a writ of . . . [55]

Although Descartes was able to claim that animals are mere machines due to their lack of cognitive abilities, the discussion above has indicated that this vision is slightly impaired. Animals are not "things"; therefore, provisions with respect to issues on animals apply and should be in compliance with the laws, regulations and rules of unwritten law, reasonable restrictions, obligations and principles of law, and public order and decency.[56] Although animals still have no rights, they will be treated on the basis of their role in society, yet with certain rights based on the obligations of natural persons in society. Abuse or neglect of animals will not be accepted and rules as such are also included in the Criminal Codes;, and certain rights for animals, in the Netherlands since 2011, are included in "the law

[51] The Turing Test published by Alan Turing [27] was designed to providence a satisfactory operational definition of intelligence. Turing defined intelligent behavior as the ability to achieve human-level performance tasks, sufficient to fool an interrogator.

[52] State of New York, Supreme Court, Appellate Division Third Judicial Department. Decided and Entered: December 4, 2014 (518336). Available at: http://decisions.courts.state.ny.us/ad3/Decisions/2014/518336.pdf. Accessed 20 October 2017.

[53] Ibidem, p. 6.

[54] Ibidem, p. 5: Amadio v Levin, 509 Pa 199, 225, 501 A2d 1085, 1098 [1985, Zappala, J., concurring] [noting that "'[p]ersonhood' as a legal concept arises not from the humanity of the subject but from the ascription of rights and duties to the subject"]).

[55] The Nonhuman Rights Project (NhRP) further stated: chimps and other select species—bonobos, gorillas, orangutans, dolphins, orcas, and elephants—are not only conscious, but also possess a sense of self, and, to some degree, a theory of mind. They have intricate, fluid social relationships, which are influenced by strategy and the ability to plan ahead, as well as a sense of fairness and an empathetic drive to console and help one another. In many ways (though certainly not all), they are like young children. The NhRP contends, based on this, that chimpanzees are capable of bearing some duties and responsibilities.

[56] Dutch Civil Code, Book 3, Article 2a.

on animals."[57] This animal has no legal personality but there is a societal tendency to have more rights applicable for animals and not just to the legal and beneficial owner of an animal. The owner and others have been given more responsibilities with regard to the animal in the context of acting carefully and friendly. Animal are not considered to be objects.

It is imaginable that some categories of social robots as pets, companion robots ,and sex robots would be considered in comparable way.[58]

Yet there are also voices to provide animals with some form of legal personality. Animals can take various decisions under the influence of different information. Is this proof that they have a comparable free will that also prove a cognitive base for their decisions on the information obtained? As long as we cannot decide on this element of free will of animals and even discuss the free will of humans a decision on free will as an element necessary for legal subjectivity will be not ultimately convincing.

18.5.2 Non-natural (Artificial) Legal Persons

For modern economic structures we cannot imagine a society without the existence of non-natural legal persons, states, companies, organizations, and other institutions, take part in the social, economic, and legal structure of our global economy. This vision was not new.

In ancient Egyptian society, the legal structure of a foundation was used to maintain temples. In Roman civilization, there were several legal entities such as the "universitates personarum," which was similar to a corporation or government college with their own identity and independent legal personality.

It seems that this concept of artificial legal personality for institutions disappeared for several centuries until it rose as a phoenix from the ashes of the Early Middle Ages again within the Roman Catholic Church. According to Maximilian Koessler, the imaginative personality of a corporation or juristic person was reinvented and appeared for the first time in the writings of an Italian jurist, Sinibaldus Fliscus (de Flisco or Fiesco), better known as Pope Innocentius IV (Pope between 1243 and 1254).[59]

A well-known Dutch international organization with legal personality, the first multinational corporation, was the Dutch East India Company (VOC), founded in 1602. This is another clear example of adapting the legal reality to the social and economic needs of the times.

A legal person, as to property, is in an equal position as an individual and natural person, unless otherwise provided by law. A legal person is, in a similar way to the individual, a legal entity to participate in socially relevant legal relationships. A legal person can go to court if its interests are affected, or can be sued in court if it acted unlawfully in the view of another legal or natural person.

As John Dewey indicated already in 1926 in the Yale Law Journal: "The Corporation is a right-and-duty-bearing entity."[60]

As stated, corporations are not equal to humans, but they do have a legal personality to act in a legal sense. Although it is a legal fiction, granted to organizations and other entities they can only act in a legal manner that is in the best interest and the purpose of this legal entity. Thus, the fiction is a kind of augmented reality as a legal layer to social reality and not imaginary, at least not within a society that is based on a legal reality. There is a global spectrum of legal persons in civil law. In the USA, this means even, to some extent, the application of the Bill of Rights guarantees to corporations. Carl Mayer describes this situation in the USA on the basis

[57] Article 350 paragraph 2 of the Dutch Penal Code (Wetboek van Strafrecht) and Law of May 19, 2011, on an Integrated Framework for Regulations on Captive Animals and Related Topics (Animals Act).

[58] Darling [28].

[59] Maximilian Koessler, the person in imagination or persona ficta of the corporation p. 437, Louisiana law review, volume 9 number 4 May 1949 (https://digitalcommons.law.lsu.edu/cgi/viewcontent.cgi?article=1615&context=lalrev).

[60] Dewey [21], p. 26.

of the development of equal treatment under the 14th Amendment. Companies are considered persons for the purpose of the 14th Amendment, i.e., companies should have the right to equal protection and due process.[61] Of course, these conceptions are not equally applied across the globe. As stated before, legal as well as social conceptions differ throughout countries, cultures, and political structures.

Can we derive useful comparisons from these characteristics to define a legal framework for the artificial intelligent entity?

18.6 Autonomous Artificial Intelligent Entities

Can we always understand the working of the algorithm of the autonomous robot if it is self-learning? Can we understand the brain of humans if they are self-learning? If we can create rule for autonomous entities as natural persons why should it not be possible to strive for a place in our legal framework for artificial autonomous entities? The simplicity of the idea was already described by Descartes seeing the "autonomous fountains robots" in the French royal gardens. Now conceivable as robots:

> For we can easily understand a machine's being constituted so that it can utter words, and even emit some responses to action on it of a corporeal kind, which brings about a change in its organs; for example, if it is touched in a particular part it may ask what we wish to say to it; if in another part it may exclaim that it is being hurt, and so on, but it never happens that it arranges its speech in various ways, in order to reply appropriately to everything that may be said in its presence, as even the lowest type of man can do.[62]

Although Descartes clearly had an open mind to future developments he could have never imagined algorithms that will be capable to respond in a rational way. Still we expect AI to mimic the human behavior and the robot is assumed to be an autonomous function artificial intelligent self-learning entity. AI is described as a system applied to an advanced computer technology, which is aimed at imitation of intelligent human behavior,[63] partly to understand (human) intelligence and also to create intelligent creatures that can operate autonomously in complex, changing situations.[64] Will such a system need legal personhood? This will depend on the dimension where it will function, in society, in culture, and its intended purpose. For instance, as applied to robots with multipurpose tasks that require intelligence and social behavior, a certain legal competence is thinkable. As the possible cooperation between those autonomous robots and natural persons will be very probable, a legal mutual commitment based on trust is a perquisite.

This line of thinking is also observed in the earlier referred to motion of the European Parliament in consideration 50:

> Notes that development of robotics technology will require more understanding for the common ground needed around joint human–robot activity, which should be based on two core interdependence relationships as predictability and directability; points out that these two interdependence relationships are crucial for determining what information need to be shared between humans and robots and how a common basis between humans and robots can be achieved in order to enable smooth human–robot joint action . . .

Of course, this moment is still shrouded in the nebulae of the future, but it is probably nearer than we think given the pace of technological developments in this context.

18.6.1 AI in Robotic Entities

There is a fast development of increasing use of AI in numerous processes, as for instance in assistance and guidance: a human–robot interaction in Japan where robots function as help and guidance for travelers with minimal human con-

[61] Mayer [29].

[62] Descartes, *Discourse on Method and Meditations on First Philosophy*, New Haven & London: Yale University Press. (1996), p. 3435.

[63] Shoyama [30], p. 129.

[64] Russell and Norvig [31], pp. 1 and 18; also referring to the following definition of AI: The act of creating machines that perform functions that require intelligence when performed by people. Kurzweil [32].

trol.[65] Another example of a semi-autonomous functioning system is IBM's Summit and Watson as it carries out numerous tasks at the moment in the field of DNA research, teaching, and seed breeding, to name a few.[66] Nevertheless, this system still receives its initial instructions from an individual. Even under these limiting circumstances, one could consider that there are certain legal effects that result from its own functioning. This could provide for certain attributed legal personhood, be it that there have to be limits to the extent of legal consequences, as will be explained later.

In today's society such systems or robots are still (at least partly) controlled by natural persons. However, there is an undeniable trend toward the use of self-thinking and self-acting systems. Also, natural persons are controlled in their professional activities in comparable ways by other natural persons or (artificial) legal persons. AI applications will be in the field of all kinds of industries, such as hosting, social and physical support, care robot in physical and social sense, the sex robot, industrial robots, medical robots, surveillance robots, military robots, drones, etc. In the medical sector, molecular nano-robots are deployed of chemical or organic origin.[67]

The fear of the unknown creeps up on us when AI becomes uncontrollable in the sense that we cannot understand the processes that move the AI system or entity because the self-learning and teaching element is beyond our human comprehension. This is the so-called super-intelligence and is the result of the singularity based on Moore's law and paradigm shift. Moore observed the fact that the capacity of microprocessors doubled every 2 years. Vinge and Kurzweil broadened this concept to other technological developments, including a shift to other forms of technology if the former development would hamper the further progress, for instance from micro-processing to nano-processors. This increase would also manifest itself in the development of intelligence by artificial means, resulting in super-intelligent entities of a bio-digital character or, of course, a manifestation not yet known to mankind.

Nick Bostrom has defined super-intelligent systems as: "Any intellect that radically outperforms the best human minds in every field, including scientific creativity, general wisdom and social skills."[68]

It is alluring to elaborate further into the apocalyptic scenarios predicted by Vinge and Bostrom and others but I will restrict myself to the legally relevant perspective. The robot is not yet super-intelligent but can be considered as a dynamically evolving concept that started as a machine, fueled with AI, and is constantly evolving into a complex autonomous functioning robot and—maybe in a later stage—super-intelligent or semi-humanoid system.[69] The nature of this entity—electronic or organic-chemical—is less relevant for its legal characterization. The state of intelligent autonomy and its function in society will be more relevant in determining its legal status.

One could refer, in this respect, to the development of the "intelligent" car. This is already happening and therefore an understandable example. The modern automobile is quickly developing an increasing autonomous mode of operation. We already drive with all kinds of warning systems, automatic breaks, distance keeping, etc. According to the road traffic law, the driver is the responsible party. But how to justify this when the driver is gradually losing control over the car and, instead, depends on numerous

[65] https://bit.ly/2tzJs6M

[66] See http://www.ibm.com/watson/ and https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/

[67] Examples are the molecular machines as designed by prof. Ben Feringa, Nobel laureate in 2016.

[68] Bostrom [33].

[69] Already in the 1960s this development was predicted: let an ultra-intelligent machine be defined as a machine that can far surpass all the intellectual activities of any person, however clever. Since the design of machines is one of these intellectual activities, an ultra-intelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of humans would be left far behind. Thus, the first ultra-intelligent machine is the last invention that humanity need ever make, provided that the machine is docile enough to tell us how to keep it under control. Good [34], cited by Vinge [1].

providers of information? These providers are the manufacturer, the infrastructure, road managers, other motorists, the producer of the software, the meteorological department, the designer of the algorithm at the heart of the learning vehicle, and third-party data providers that control or affect navigation and engine control. Therefore, the suggestion in the earlier mentioned European Parliament motion to adapt the outdated Vienna Convention on Road Traffic of 1968 is not undue.

> 10. Expects the Commission to ensure that the Member States adapt the existing legislation, such as the Vienna Convention on Road Traffic of 8 November 1968, in a uniform manner in order to make driverless driving possible, and calls on the Commission, the Member States and the industry to implement the objectives of the Amsterdam Declaration as soon as possible;[70]

But what if a direct link between the brain activity of the "driver" and the software control is made? Not so futuristic, there are already cars that respond to drivers who threaten to fall asleep where certain movements betray a delay in reflexes. Going one step further, those links are analyzed in an external autonomous system that will control the traffic flow. This (is?) not the plot of a science fiction novel. Elon Musk is also moving into neuro-tech; he launched Neuralink, a company that is researching methods to upload and download thoughts. Ultimately, Neuralink aims to change the way in which we interact with devices by linking our brains to the machines we interact with most often: cars, mobile devices, and even smart items in the smart home of the future. This also is happening in the academic research at the University of Witwatersrand, SA, as referred to earlier: the "Brainternet" project streams brainwaves onto the Internet. Essentially, it turns the brain into an Internet of Things (IoT) node on the World Wide Web. "The concept of IoT refers to connecting any device with an on and off switch to the Internet."[71]

Another example is an AI application that is used in the selection of candidates for jobs. Beyond the algorithmic selection of candidates based on their email or letter, by so-called applicant tracking systems (ATS), AI can evolve into AI robots that can be used during a conversation to watch an individual's posture, eye movements, sweating, tuning stability, and other mental and physical reactions. This analytical achievement will be developed to even a greater extent in the "care industry," where autonomously functioning robots will apply client custom-made solutions to the needy without the necessity of guidance from outside.

To determine the legal classification of the AI entity as a simple tool, a legal object that is used as an instrument, or as an autonomous artificial intelligent entity that will operate independently and could be classified for legal activities, we have to determine its role and status.[72] Whether robots should be compared to legal persons or legal objects is to be answered for a great deal on the basis of function and autonomy. This decides whether they are assessed similarly as thing, as minor, as non-subordinate, as movable property and animals,[73] or as independent legal entities.

A complicating factor is that it is not so easy to tie to a breakdown of legal persons and legal objects. The artificial legal person that is a company can be an object too, it can be sold, and it can be divided, but also it can be held responsible for its actions. To determine a sensible solution

[70] Reference to the Declaration of Amsterdam of the Council, of 14–15 April 2016, on cooperation in the field of connected and automated driving ("Amsterdam Declaration").

[71] Minors [22] Can you read my mind? Available at: https://www.wits.ac.za/news/latest-news/research-news/2017/2017-09/can-you-read-my-mind. Accessed 11 October 2017.

[72] The Principles of European Tort Law ("PETL") refers to liability for "auxiliaries" (6: 102)—an apt term for both robots, although in PETL it is meant particularly for people. Article 3: 201 of the Draft Common Frame of Reference (DCFR) of the Principles, Definitions and Model Rules of European Private Law refers to workers or "similarly engaged" others, in which the phrase "similarly engages" others may contain cases of accidental damage; see: Giliker [35], pp. 38 et seq. Then the robot will have to be seen as "another," where the employer is liable under the condition that he still has "the least abstract possibility of directing and supervising its conduct through binding instructions"; Von Bar and Clive [36], pp. 34–55.

[73] Schaerer et al. [37], pp. 72–77.

for a new legal personhood structure, if needed, we have to develop an original analysis.

18.7 The Problem of Human–Robot Integration

Christof Koch and Giulio Tononi wonder if we could sell our soul on eBay, considering the soul to be our brain contents, character, and consciousness.[74] They think that would be possible by uploading your thoughts, memories, and personality to a computer or robot. They are convinced that this depends only on mathematics and logic and on the imperfectly known laws of physics, chemistry, and biology; it does not arise from some magical or otherworldly quality. Therefore, there's no reason why consciousness can't be reproduced in a machine—in theory, anyway.

Also Ray Kurzweil supposes this could be possible:

> Suppose we scan someone's brain and reinstate the resulting "mind file" into a suitable computing medium. Will the entity that emerges from such an operation be conscious?
>
> This being would appear to others to have very much the same personality, history and memory.[75]

According to utilitarianism, the ethical theory that states that the best action is the one that maximizes utility in the sense of the best solution for the well-being of the(human) society, the outcome of an action is the most important factor in deciding whether the action is good or bad.[76] We have enhanced our standard of health continuously by biomedical and technological adjustments and appliances. So why should we have a problem with the next step of improving the pace of evolution?

There are some disputable elements though. In the transhumanist movement, people are using technological devices and body adaptations to "upgrade" the human body to an above-natural level. However, such inventions generally are more expensive compared to conventional technology due to its "cutting-edge" nature. This could enlarge an already existing gap between higher educated socially well-to-do people and non-highly-educated persons in socially deprived parts of the (global) society.

Human–Robot Integration

Muscular or neural implants or other prosthetics are now even able to provide movement by brain activity and sensation of touch. These prosthetics are coming closer and closer to a replicate of an actual human arm/leg than ever before. It is only a matter of time before these artificial limbs even surpass the abilities of a natural arm or leg. An artificial limb could be made to be significantly stronger compared to a regular human limb, and it's easily replaceable.

DARPA[77] is doing research on the creation of implantable neural chips, which are able to cognitively improve soldiers of the US army. Because DARPA's goal is directly to improve the cognitive abilities to an above-natural level, this can already been seen as a current transhumanist invention. Its sole purpose is to upgrade our biology by use of technology. Scientific research institutions are also focusing on creating neural implants to treat patients suffering from brain trauma. As the scientific community is investing in this line of research as well, significant improvements in brainpower can be expected due to these neural implants in the near future.[78]

Persons like Randal Koene have dedicated his research in his nonprofit foundation Carboncopies.org to create "substrate independent mind" (SIM), which is independent of the substrate on which this network is created. In this way, a human mind could be transferred to a

[74] https://spectrum.ieee.org/biomedical/imaging/can-machines-be-conscious

[75] http://www.kurzweilai.net/pdf/RayKurzweilReader.pdf, p. 91.

[76] Based on Theories by Jeremy Bentham, An Introduction to the Principles of Moral and Legislation, 1789, London, and John Stuart Mill, Utilitarianism, 1861, London.

[77] Defense Advanced Research Projects Agency (DARPA) is the wing of the U.S. Department of Defense which is responsible for developing emerging technologies for military use.

[78] Also see: https://futurism.com/brain-based-circuitry-just-made-artificial-intelligence-faster/

computer or robot body. In theory, this could make humans (at least their minds) immortal. Question is if this entity could still be considered as a human with the same rights as a natural person.

Of course, there will be moral ethical and legal questions arising from these developments. Should it be required to enact regulations about requirements of the circumstances that allow these adjustments and even replacements of (parts of) the human body and mind?

Would it be prohibited to remove any healthy tissue from healthy patients that is meant to be replaced by an above-natural artificial body part.

Would it be allowed for healthy humans to "upgrade" their body parts to an above-natural level? If forbidden, this could lead to shady businesses as regretfully still is the case considering abortion and even plastic surgery in many countries. Certain is that we need a new legal framework to create an acceptable way to deal with these moral and ethical issues.

Criminal Law for Robots?

If our society will be inhabited by a large quantity of autonomous intelligent robots and cyborgs, how should we deal with those elements that will act with illegal intentions and that will develop criminal minds. The self-learning algorithms could choose to develop in a less law-abiding way we had hoped. Looking with simplifying glasses at the criminal law, one can regard it is as an instrument given to the state by its subjects or obtained in a less democratic way by a state authority, with the purpose to secure law and order and security in the society. The content is directed on the offender of these specific societal rules of behavior and social values and norms and consists of punishment of this behavior with the intention to punish, correct, or re-socialize the offender. This system is developed to keep human behavior between the lines of society but, of course, is dependent on the time, in the sense of the era, culture, and political system.

An exemplary issue for a lot of people contemplating the legal difference between the legal position of robots and humans is the question how to punish a robot if "it" commits a crime. Also, scholarly colleagues often ask me in what way we should punish robots if they would commit a crime, as this is a pitfall to give up the legal positioning of robots. Of course, the question is easier to state than the answer.

It will be dependent on whether we accept the robot as a legally and morally accountable entity, a sentient legal subject, or just as an object. The question is whether this legal description will suffice for a clear separation. Loyal to the comparison with other mammals and, in particular, with human beings as well as artificial legal persons, we have to start the comparison with these "structures."

Comparison with existing legal persons only suffices if we want to connect to the ideas of the positive legal system of criminal law where there is a strong conviction that the deed has been committed by natural persons or, at least, under the responsibility of natural persons as in the case of artificial legal persons. After all, companies can commit crimes. These crimes are mostly of a financial character, such as fraud, money laundering, or tax crimes, but also environmental crimes involving pollution by chemical and oil industries, or false reporting as in "Diesel gate" in the automobile industry and even discrimination of clients or in the personnel area. Mostly, the punishments are fines, sometimes extremely high if it is considered to be a crime against competition rules, for example. Very seldom the crimes are considered murders but maltreatment, or even culpable death, which are not uncommon in the case of chemical and medical industries. Also, states can commit crimes as polluters, financial villains, or war criminals. Generally, such crimes will be paid out of the financial reserves of the company, and, in rare cases, the responsible board members, or in the case of war, responsible state commanders, will be put on trial.

It is not unimaginable to submit other entities than human beings under the realm of criminal law. As alluded to above, in the Middle Ages, several criminal proceedings were held against animals in the same way as they were held against humans. In 1266, in Fontenay aux Roses, a pig was convicted and brought to death in the city square because the beast had bitten and killed a child. The judge ordered the executioner first to cut off a paw to be followed by a beheading. Before the execution, the pig was dressed in the clothes of a human being. There were other cases against horses, cows, and bulls that wounded or killed humans or other beasts. In such cases, the animal was punished without holding their masters accountable. The animals also had the rights of legal support by an assigned counselor.[79] To make the comparison even clearer I cannot leave out the description by the counselor (barrister) and president of parliament, Mr. Chaseneux of a court case in 1488 against rats where he was the counselor in defense of the rats. Because his clients were divided over several cities and cities the subpoena had to be brought out to the rats by proclaiming them after mess in every city so it would be possible to make it noticed by the defendants. After this had been done the counselor pleaded that his clients could not be expected to be present at court because all this has brought so much attention to the cats of the cities and village that his clients would endanger their lives if they would be summoned to court.[80] The last case against an animal in the Netherlands was against a bull in the town of Zwolle in 1664 after he impaled his own master. His counselor could not do much to save his client; he was stoned and buried alive in conformity with what was written in the Bible in Exodus 21:28.[81] Most interesting for the parallel with the robots is that the owner was not held legally responsible.

A different perspective was proposed by authors to compare and qualify robots with different types or breeds of dogs. The animal has a natural analogy to robots, where the "type" of animal, in this case a friendly or aggressive type of dog, is relevant. Some breeds of dogs are prohibited to breed as for instance the pit-bull fighting dog. Personally, I think that the danger often lies with the master of the dog.[82]

The analogy between humans and robots, though, has to be based on the fact that also is relevant for humans: is there an intention to commit the crime and is the relevancy in the potential of the act to create a negative result or damage to the victim? Sentiency in the sense of consciousness is relevant. Will it be an act committed by the responsible actor or is the entity use as an instrument by a (human) third party? And how will the punishment be executed? Remove the human brain from the robot for isolation in a fridge for 6 years? Disintegration of the robot? Penalty for the human that initiated the act? It will be sensible to study the different scenarios.

18.8 An Alternative Personhood

Personhood in a legal sense is not carved in stone; there is elasticity of the concept due to the elasticity of societal needs, dependent on what is deemed acceptable within certain social, cultural, political, and geographical parameters. If animals are accepted to have a certain status in that society and culture, they can have a legal status going beyond that of a mere object. If a company has legal personhood because it is socially and economically desirable, why should it not be acceptable and even desirable to give a robot a certain legal status and to have a new kind of personhood. This practical view or utilitarian view of legal personhood is made by the first abstraction by Naffine. To see if this analysis will be of help to determine what legal position could be applicable to AI entities, one may consider this model. Naffine gives three possible models for legal personhood:

1. The (lucid) Cheshire Cat
2. Any reasonable human creature
3. The responsible subject

[79] Erven D onder de Linden en zoon [38], pp. 201–203.

[80] Berriat Saint-Prix [8].

[81] "If a bull gores a man or woman to death, the bull is to be stoned to death, and its meat must not be eaten. But the owner of the bull will not be held responsible."

[82] Kelly, Schaerer & Gomez, Liability in Robotics: An International Perspective on Robots as animals, paper Nevada University (https://bit.ly/2tkSkxU).

18.8.1 Abstraction of the Legal Position of the Robot by a Narrative

18.8.1.1 The Cheshire Cat

The concept of the legal person that Naffine mentions the Cheshire Cat is regarding the most lucent aspect of personhood.[83] According to this definition, to have personhood means nothing more than the formal capacity to be a carrier of legal rights and duties.[84] There is no moral or ethical dimension to this definition.

The person exists only as an abstract capacity to function in law, a capacity which is endowed by law because it is convenient for law to have such a creation.[85]

Anyone or anything can be considered a person in the eyes of the law, because the only reason that legal personhood exists in the first place is because of the practical advantages of such an attribution. This definition of legal personhood is the most comprehensive definition of personhood of the three.

This model does not have any moral, ethical, historical, or empirical content.[86] Following this definition, there is no reason why animals or other legally functioning entities should not be considered persons. As long as they are able to carry solely one right or legal duty, there is no reason to not grant them personhood, even if a human is necessary to enforce that right.[87] This should not be a problem since the same enforcement from a legal competence is required for minors and other legal incapacitated persons. The interesting part is that there should be no requirement for the scope and contents of the legal subjectivity.

This theory also denies the necessity of differentiating between natural persons and artificial persons, or other entities. In either case, the concept of personhood is an abstract concept; neither the natural person nor the artificial person is more real than the other. Both of their legal personalities are based on the fact that they retain a particular bundle of rights and duties. This is the essence of the Alice in Wonderland character of the vanishing figure: take away the rights and the duties of the person and its legal personality vanishes like the Cheshire Cat.[88] Supporters of this theory thus envisage the concept of legal personality as an empty slot that fits anyone or anything.[89]

Of course, this concept leaves open the question if other legal and natural persons are willing to perform legal actions with this "new Cheshire Cat." This point can be illustrated by the development of robots when this development reaches a point where robots and people look very much alike and almost cannot be distinguished. The concept "uncanny valley," introduced by Masahiro Mori, is used to indicate the point when feelings of eeriness and aversion to humanoid robots arise.[90] This is when human–robots appear almost, but not exactly, like real human beings. The question arises if humans want to create sentient robots that resemble human beings so much, also considering the legal status of robots, giving them rights that reflect their human-like status.

18.8.1.2 The Reasonable Human

The second concept Naffine proposes is that a (legal) person is any reasonable human creature.[91] A required rights position by birth with one requirement though is that one has to be reasonable. Simply put: to qualify as a legal person, one has to be human. This perspective is the most dominant and comes closest to the common language usage of the word person, at least from an Anglo-Saxon perspective. It is common legal knowledge that someone, in this context meaning a human person, becomes a

[83] Naffine [14], p. 350.

[84] Naffine [14], p. 350.

[85] Naffine [14], p. 351.

[86] Naffine [14], p. 351.

[87] Naffine [14], p. 351.

[88] Naffine [14], p. 353.

[89] Naffine [14], p. 356.

[90] Mori [39] The Uncanny Valley: The Original Essay by Masahiro Mori. Available at: https://spectrum.ieee.org/automaton/robotics/humanoids/the-uncanny-valley. Accessed 15 October 2017.

[91] Naffine [14], p. 357.

legal person at the very moment of being born or conceived, depending on the legal order and it certainly ends at death. Furthermore, there is the possibility to limit the scope of personhood if the rationality or psychological stability is not present but personhood, as such, still exists.

There are two ways in which this common legal knowledge is interpreted. Firstly, this reasoning could refer to a human being who has been born alive and has not yet died, and is thus considered a human, therefore a person. Secondly, it could refer to the rights and duties of a person that starts to exist as soon as someone is born as a human being and which ceases to exist as soon as this same person dies.[92]

Either way, personhood is linked with both biological and metaphysical notions of humanity. Taking this definition, personhood is not a purely legal matter anymore, but concerns instead the question of what it means to be human.[93] This is also the main criticism of this theory from the perspective of the Cheshire Cat definition. Supporters of the concept of the person as a rational human are, according to supporters of the Cheshire Cat concept, misguided because of their reliance on extra-legal biological or moral considerations.[94] The terms “human being” and “person” are being used synonymously and interchangeably by supporters of this second theory.[95]

The definition of the legal person as a human being has the advantage of simplicity. For someone to be considered a person, one does not require any quality except for that of being a human. Therefore, this theory includes all humans, regardless of their mental or physical state, thus being compatible with the human rights movement. In the meantime, this definition excludes—in line with the common legal view—other nonhuman animals from personhood. Corporations as artificial legal persons are able to carry personhood under this definition because they are reducible to the relations between the persons who manage them, own them, work for them, and act in mandate.[96] This definition of personhood, however, is not compatible with the demands of the qualification of differences based on the legal requirements by society. It should, however, be considered in giving legal status to AI entities in the same way the artificial legal person is considered as a vehicle for inter-human legal relations and therefore is served with legal capacity.

18.8.1.3 The Responsible Actor

A rights-based conception is to be found in the third concept of legal personality observed by Naffine, “the rational, responsible actor: a high-threshold definition since not all humans possess the qualities to be considered persons under this definition. It is going one step beyond the reasonable” human.[97] This definition insists on a certain level of mental capacity and therefore excludes young children, mentally incompetent humans, and animals.[98] This theory recognizes the human form of personhood, but does not see this as the critical characteristic that sets a human apart as a person. For the perception of a responsible actor it is, rather, the rationality, the mental attributes, and the ability to comprehend a certain situation determine this situation.[99] Although seeming to set this definition of the person as the ideal legal actor, it also encounters the danger of elitism. Moreover, the idea is not very original. Most legal orders already have a system of legal incapability in a private and criminal law sense. Naffine states that under this definition, the person can actually be meaningfully subjected to legal punishment for criminal acts.[100] Criminal law has to treat the person as a responsible actor with a free will because otherwise one cannot take responsibility for one’s actions. If a person is not capable of making rational decisions, then what is the point of punishing this person? This reasoning already is applied in many legal systems as “being not accountable

[92] Naffine [14].

[93] Naffine [14].

[94] Naffine [14].

[95] Naffine [14], p. 358.

[96] Solum [11], p. 1239.

[97] Naffine [14], p. 362.

[98] Naffine [14], p. 364.

[99] Naffine [14], p. 364.

[100] Naffine [14], p. 364.

for one's actions due to psychological stress or other mental or physical factors." One of the main goals of punishment in criminal law is the prevention of a person committing the same criminal offence again. If a person is not capable of making rational decisions in the first place, then they cannot be expected to learn from their punishment. Nonetheless, in the case of criminal law, this definition of a legal person is simplifying reality; in many ways the law shows awareness of the weaknesses and dependence of human individuals and in many ways the law does not require persons to be as rational and responsible as this Naffine definition requires a human to be.[101] This definition has a utilitarian aspect based on the measure of attributable sentiency. A person can only be responsible for the acts willfully and rationally committed. But could it not be applicable to nonhuman actors too?

18.8.1.4 Concluding on Legal Position

The choice within the legal system could be made to allocate legal personhood to anything according to the Cheshire Cat theory, regardless of the nature of the entity that it is allocated to.[102] Inanimate entities have been the subject of legal rights at various times in the past. As mentioned above, temples in Rome and church buildings in the Middle Ages have been regarded as persons in the past.[103] So have ships, an Indian family doll,[104] and Indian and New Zealand rivers.[105] And certainly a parallel can be drawn with business corporations and with government entities.[106]

As we zoom in on the example of corporate personhood, we can see a lot of parallels with the proposed electronic and AI entity personhood. Similar to a corporation, the aims of an AI entity robot may lie in economic profit for the producer or owner of a robot, or in the social welfare of a society. For example, a robot working for an automobile manufacturer may improve production and thus profit for the manufacturer, while a robot caring for an elderly person will be carrying out a civic service. The reason why personhood has been invoked for corporations and robots seems to correspond as well; they reduce the responsibility and liability of the owners in case of damage inflicted by the corporation or the robot. Corporate personhood has seen the liability of its shareholders limited to a certain extent by corporate legislation. Electronic persons or other organic entities could fall under similar legal qualifications. Taking this definition as our base, there should be no problem in granting personhood to AI considering their specific task or function.

Concerning Naffine's concept of legal personality being connected to the human, granting personhood to AI would be a problem. If personhood can only be granted to humans purely based on the fact that they are humans, then it would not be possible for AI to obtain legal personhood. Then how is it possible that corporations are granted personhood? But the legal connection to the natural person could be the trait d'union. The property of a corporation is eventually the property of its shareholders.[107] Damage done to a corporation would directly injure natural persons.[108] As such, corporations are reducible to the relations between the persons who manage them, own them, work for them, and so forth.[109] So, the fact that corporations have legal personalities does not necessarily mean that AI entities should be granted legal personality or the same legal capacity. The question lingers though if existing legal persons could represent legal persons (and/or natural persons) in the same way natural persons function in representation or in the use of mandates. Could the attribution of rights be compared with those attributed to natural persons although they would not have the same status as natural persons?

Rejection of the human being personhood concept, granting personhood to AI, is based

[101] Naffine [14], p. 365.

[102] Naffine [14], p. 351.

[103] Solum [11], p. 1239.

[104] Solum [11], p. 1239.

[105] Safi [40].

[106] Solum [11], p. 1239.

[107] Solum [11], p. 1239.

[108] Solum [11], p. 1239.

[109] Solum [11], p. 1239.

on the conception that acceptance would undermine the meaning of being a person because it reduces the exclusive belonging of personhood to humans. This exclusivity has been represented by religious texts such as the Bible: man is separate from nature and is created in God's own image. This hierarchy sets humans above "things," be it animals, property, or the environment.[110] This argument against granting personhood to AI seems to only be problematic if one uses the terms human being and person synonymously and interchangeably. Electronic or robots personhood does not have the intention to interfere with the exclusivity of humans' place in the world. According to the common legal view, a natural person (being a human) is different from a juridical person. A legal person does not have to be made up out of blood, flesh, and DNA, but exists to ease economic traffic and proceedings in a court of law.

Another argument against granting personhood to robots which aligns with this second definition of personhood is that, because of the special place that humankind has granted itself, it is not in the interests of humankind to grant robots personhood.[111] This argument shows similarities with slave owners stating that slaves should not have constitutional rights simply based on the fact that it is not in the interest of slave owners to grant them such rights and also deny them a comparable human status.[112]

Overall, robots do fit in with this second definition of the legal person with at least some difficulty and bending of the concept. Even though most arguments against the granting of personhood to AI entities can be put in to a practical perspective, in which such legal personality may be pragmatic and desirable, robots lack the ultimate aspect which needs to count as a person in the view of the supporters of this theory: humanity in its widest and nonlegal sense.

Returning to the concept of a person as the responsible actor,[113] the human form is not the critical characteristic that makes a legal person; the rational, mental attributes and ability to comprehend a situation will suffice to be defined as a person. These characteristics will make a person able to have full legal responsibility and to handle in a single capacity in its own right. In the current technological situation, robots are not (yet) able to perform as a legal person under this definition; it cannot act as the fully responsible and capable person that this theory prescribes it to be; robots are still too dependent on humans as they are not fully autonomous and sentient yet. But this can change rapidly.

However, we do not know how the future will unfold. Imagine a future in which humanoid AI walks around the globe with great mental capacity, able to comprehend its own situation and have responsibilities[114]; would this sort of robot qualify as a legal person within this definition?

The definition of the responsible, rational actor presumes the presence of a consciousness. Is this prerequisite for personhood something that robots could actually obtain?[115] We do not have a clear notion of what consciousness actually is and so there is little to be said about questions that go beyond our basic intuitions.[116] It could be that we cannot only get consciousness out of neurons but also out of artificial neurons as is the intention of the European RAMP project.[117] It might as well be that we cannot get consciousness out of anything except neurons and that we will never be fully able to reproduce it.[118] If robots would be able obtain a consciousness, and then according to this definition, there should be no problem granting personhood to robots. How would the consciousness of this AI be established? Since

[110] Lovejoy [41].

[111] Solum [11], p. 1260.

[112] Solum [11], p. 1261.

[113] Naffine [14], p. 362.

[114] See, e.g., the robot Sophia, of Hanson robotics, and (compare "Ava": Bush, E. (Producer), & Garland, E. (Director). (2014). *Ex machina* [Motion Picture]. United States).

[115] Solum [11], p. 1269.

[116] Solum [11], p. 1264.

[117] This project aims to build a biohybrid architecture, where natural and artificial neurons are linked and work together to replace damaged parts of the brain (https://ec.europa.eu/digital-single-market/en/news/artificial-neurons-replace-and-assist-damaged-parts-human-brain).

[118] Solum [11], p. 1265.

we do not have direct access to another person's mind, one can only assume consciousness based on behavior and self-reporting.[119] It might be that the artificial intelligent entity claiming personhood would do this on the basis of having a consciousness but would merely be faking its consciousness.[120]

An objection against granting legal personhood could be that robots lack any sort of feelings.[121] But even that could be developed in future AI, by humans or by AI itself. In the context of the legal person as the responsible, rational actor, this characteristic could actually be beneficial for the granting of personhood to AI. Supporters of this theory state that man should be a rational animal and requires that he should exercise a reasonable control over their passions.[122] As stated before, the criminal law system takes this actor as the ideal legal person.[123] A form of intelligence completely lacking feelings does not have to control its feelings because it does not have them in the first place.

Taking into account that a robot is not at a level yet in which it could function as a responsible, rational actor, robots cannot be granted personhood under this definition. Granting personhood under this concept in the future depends completely on how successfully AI will develop sentiency in robots. If AI performs in robots as a humanlike consciousness and could therefore act as the responsible rational actor this definition requires it to be, then this personhood could encompass AI.

18.9 The Artificial Intelligent Entity or Robot as Legal Actor

Do we need to compare the role and personality aspects of robots and other AI systems with existing legal personhood or at least with elements of existing personhood? In other words, is having legal personality desirable for robots and society?

The consideration that such an autonomously functioning artificially intelligent robot should have a secure legal subjectivity is dependent on the actual social necessity in a certain legal and social order. In other words, will a future society still function without any form of legal personality for autonomous artificially intelligent entities? Or will it have a need to place the entity within the framework of legal personhood?

The deployment of autonomous robots in the near future could be comparable to the efforts of individuals representing institutions and organizations and to the efforts of individuals working as mandated legal representatives. As an example, I refer to a social service that uses a care robot deployment in support of the needy. The robot is capable of managing the household, ordering products and services, conducting physical support, and analyzing medical problems and then even performing medical procedures.

The legal consequences of this development are great. A society that depends on autonomous systems and robots cannot do without a legal framework integrating this development. It is quite conceivable that there is a need, in this future society, for a degree of legal responsibility and legal personality of robots so that the legal consequences of such acts can have a place in the legal framework. A distinction needs to be made between fully autonomous functioning entities and those entities that operate on the basis of previous entries by legal persons. Although the "Cheshire Cat" structure seems to be too simple, not taking into account all social requirements that would be necessary to perform acceptable roles and to be recognized by other legal persons, we can specify the role and function and legal effect of the AI entity.

Furthermore, the development of self-learning algorithms should be embedded legally before proceeding to the question whether legal personality provision to robots is at order. In addition, demonstrating a defect in the software requires a profound technological knowledge of the functioning. It is unlikely that most claimants have easy access to this type of knowledge.

[119] Solum [11], p. 1266.

[120] Solum [11], p. 1266.

[121] Solum [11], p. 1269.

[122] Naffine [14], p. 364.

[123] Naffine [14], p. 364.

18.9.1 Sui Generis Construct, Legal Subject or Legal Object Specialis?

Definitions of legal subjects and legal objects are not so watertight specified as it may seem. Therefore, the flexibility in states in the range of different legal objects and legal subjects could offer an interesting analogy with classification of AI entities. Although the definition of a legal subject does yet not completely coincide with the characteristics of an AI entity, it shows an increasing number of interfaces. Because of the variation in types of AI entities, from vacuum cleaner to sex robot, it is impossible to provide a uniform legal regime for robots. But the same goes for legal persons such as limited companies, foundations, etc. These entities are classified by purpose and function and also have different rights and obligations. For individuals, there is a similar specification with regard to act. Children and adults under guardianship as such have a legal status under the supervision of another natural or legal person. Individuals will function under supervision or independently, and their activity affects their interpretation of legal personality and the performance of their acts. Government officials, secret service officials, and the military but also medical physicians and journalists have a different legal status from other individuals concerning their function and use of rights in society.

As a classification of the specific robots would be desirable, it will depend on the degree of legal subjectivity that is needed. The legal subjectivity and derived legal capacity need not be equal to the legal personality such as we know it in positive law. The possible extension of legal capacity could be based partly on the concept of existing legal personhood, leading to a new "sui generis" construction, based on elements of legal autonomy for the purpose of the functioning of the robot in society. In this context, a comparison with the "peculium-like" requirements as restricted liability could be of help.

This reasoning applies when it is possible to figure out who the user or owner of the system is, and when there is general acceptance about the responsibility for the system. In the future, this will become an increasing problem as systems function more autonomously and interact with similar systems. Car manufacturers of smart cars until now have still accepted a strict risk liability. This means that the producer accepts responsibility for errors or incomplete functioning of the system and of automatic control systems. But this system may easily come to an end because of the technical and financial burden.[124] Strict liability of the "user" could also be a solution when AI is fully deployed.

Is the boundary between legal subject and legal object always clear? Legal objects can be goods, services, rights, or objects that are the carrier subjects of rights and obligations. Objects can never be bearers of rights and obligations similar to a legal entity. The legal property concerns, in particular, business, products, and services, but is also applicable to more artificial legal person concepts like an organization or company. The lastly mentioned legal persons may perform as a legal object but are themselves legal entities. This special construction is also described as a set of active and passive proprietary elements. The sui generis construction for AI can take this in consideration. Robots could be legally considered either as objects or subjects depending on the legal activities of other legal actors. One could interact with AI entities with legal effect but the owner also could sell them or pawn them.

18.9.2 Liability and Legal Subjectivity

The liability of a legal person shall also apply to the director or directors, being natural persons at any time during the life span of the liability of the legal persons if they had the responsibility or

[124] 2018 US overview state legislation: http://www.ncsl.org/research/transportation/autonomous-vehicles-self-driving-vehicles-enacted-legislation.aspx. Also: EU Common Approach on the liability rules and insurance related to the Connected and Autonomous Vehicle EP study (http://www.europarl.europa.eu/RegData/etudes/STUD/2018/615635/EPRS_STU(2018)615635_EN.pdf).

were authorized to act for the legal person. This seems to apply to AI and robots as well. Robots can be classified simply as legal objects, but they can also occupy a special position. In several publications, the comparison has been made with slaves. As also referred to by Ugo Pagallo, Norbert Wiener, who compared robots with slaves: "the automatic machine, whatever we may think of any feelings it may have or may not have, is the precise equivalent of slave labour." Also referring to Leon Wein in *The Responsibility of Intelligent Artifacts* (1992), in the sense that: "As employees who replaced slaves are themselves replaced by mechanical 'slaves,' the 'employer' of a computerized system may once again be held liable for injury caused by his property in the same way that she would have if the damage had been caused by a human slave."[125] What is more, Voulon stated that the intelligent agent, such as a software robot, was compared with a slave, deployed to carry out a particular task.[126] We can easily draw parallels with existing machines that perform the needed legal actions to fulfill legal statements and transactions:

> Such a machine would need to have two abilities. First, it must be able to render correct outputs from given factual inputs. Second, its output needs to be reified some way in the real world. The vending machine is the archetypical example of a self-executing smart contract. Vending machines have been defined as 'self-contained automatic machines that dispense goods or provide services when coins are inserted'.[127]

In other words, the vending machine completes one side of a contractual relation. A funny example in this respect is the case of the British bookseller, Richard Carlile, in the year 1822, who invented a book-dispensing machine so as to avoid prosecution under the country's libel and sedition laws. He had been jailed previously and wanted to avoid any future liability, so the idea was to make it impossible for the Crown to prove that any individual bookseller actually sold the blasphemous material. He argued that it was purely a contract between the buyer and the machine with the publisher having no formal involvement. Here is Carlile's description of the machine as it appeared in *The Republican*:

> Perhaps it will amuse you to be informed that in the new Temple of Reason my publications are sold by Clockwork!! In the shop is the dial on which is written every publication for sale: the purchaser enters and turns the hand of the dial to the publication he wants, when, on depositing his money, the publication drops down before him.[128]

The Crown, however, was not amused. Use of the device was ineffective and both Carlile and his employee were convicted of selling blasphemous literature through the device.[129] Our society is full of these kinds of devices. The provider is usually very simple to identify: the city for parking meters, the selling company for soft drinks on the street, or hotels. But cigarette dispensers are somewhat more difficult. Is the other party the shop owner or the cigarette company? Although we do not know for sure we do not mind and just proceed with the transaction. In this respect, it is all about trust and credibility.

Pagallo, citing Chopra and White, also explained that, from the point of view of legal trust and credibility, for the acceptance of legal actions with legal effect, it must be clear on what mandate and on what legal attribution the agent is functioning.[130] For a vending machine, this is clear. For natural persons and AI entities it is not always clear. For natural persons representing legal persons, we have to look up in official registers what their legal status in attribution of legal capacity encompasses. If we make the comparison with the position of the Roman slave, it must also be taken into account that the relation between the slave and their master and the relation between the slave and society as a whole were more than instrumental. The slaves

[125] Pagallo [19], p. 3 (referring to Wiener [42]).

[126] Voulon [43].

[127] Raskin [44], p. 10 (citing Segrave [45]).

[128] Ibidem, p. 10/11 (referring to Carlile [46]).

[129] Ibidem.

[130] Chopra and White [47], p. 130, correctly remark, "to apply the respondent superior doctrine to a particular situation would require the artificial agent in question to be one that has been understood by virtue of its responsibilities and its interactions with third parties as acting as a legal agent for its principal." Pagallo [19], p. 132.

could perform a legal representative position and independent legal transactions and could appear as a witness in court. Moreover, the slave could be declared a "free man" by their master (manumission). This was not strange because at that time, on a population of one million people in Rome, there were 400,000 slaves. The position of the slave may be similar to the position of the robot in a future society although declaring them "free men" might be a step too far. Maybe robots could also hold peculium in the sense of a "financial resource to be used without human control" or a general fund enacted by AI industry, to pay for any damage resulting of AI when it occurs and is not attributable to an identified part. It is particularly crucial to determine to what extent it is desirable that robots will perform legal acts. Regarding a "Roomba" that position is clear. More complicated is the abovementioned example of a social robot that performs several functions with legal effect. For instance, when it decides upon needed medical products for a needy person or orders them and decides when and which medications should be administered. To hold a robot liable will only be efficient if the act cannot be tracked back to the original actor or "master" and to see in what legal capacity this robot is performing a task, just as a representative of a legal person or on its own account. In that case, and maybe other cases when it is not completely clear an obligatory insurance, financed by a general fund could be a solution as also proposed in the EP Motion.[131]

18.9.3 Legal Acts

Why is it so important to define the shape of a certain legal personality for robots? If the robot acts with the intention to change the legal circumstances, be it autonomous and sentient, be it instrumental as instructed by another legal or natural person, they must also have a certain legal status beyond that of a legal object. In addition, we will need to find some form of liability that will ultimately best suit the practical qualifications and role of the robot in society. It must be deemed likely that robots in the surveillance and security areas as well as in the advisory and in the health sector, as well as in more exotic services, will play an important role without direct control by natural persons. The acts have to be recognized by other legal subjects based on trust and acceptance.

The responsibility of persons who are performing legal acts for others will ultimately rest with legal persons, a group or single identifiable individuals, the government, the official, political leaders, and representatives accredited to a natural person. With the use of robots in those areas, that same responsibility will usually be traced to the same group and the robot will play a preparatory policy role or even a representative role.

It is conceivable that the robot will also be given a certain mandate attributed to them by authorities in the public sector to perform certain specified duties. Responsibility has to be determined. The arrest of a suspect by a "Robocop" has also to be secured legally. Legal and natural persons may be represented by robots in the future. This is a different situation than the legal representation by natural persons. This is only possible when it is established which specific competencies are relevant to the performance of the task of the robot. The attribution of competences has to be recognized by law. Only then there will be a legally credible acceptance of the legal effect of the performed acts by the robot.

Already the actions of an automated system may have legal implications. The advanced search robot meets other bots and will exchange some codes which can result in an agreement to reserve a seat or buy a product or service. The robot will enter a possible electronic agreement to be accepted by both electronic "parties" without any intervention or even confirmation by a natural person. Can this "Crawler Bot" still be considered an object if it has a kind of

[131] An obligatory insurance scheme, which could be based on the obligation of the producer to take out insurance for the autonomous robots it produces, should be established. The insurance system should be supplemented by a fund in order to ensure that damages can be compensated for in cases where no insurance cover exists. RR\1115573EN.docx, p. 20.

legal subjectivity?[132] As long as it is seen as an instrument used by other legal persons, it fits in the existing legal framework. But what if the instructions are vague and the action is mainly performed by a self-learning algorithm? This requires a clear explanation of the legal circumstances, preferably in the law and the contract, general terms, and conditions.

Up until today, the fact that individual machines and devices were used for a purpose made the question of legal personhood irrelevant. Several times, warnings were issued by concerned scholars and captains of industry concerning the dangers of autonomous AI weapons—so-called killer robots—recently in an open letter by the Future of Life Institute to the UN Convention on Certain Conventional Weapons.[133]

What will be the qualification when a surgeon does not perform the surgery, but has recourse to sophisticated data supplied by a laser instrument that includes all medical information, including patient documentation? Or, if the computer or the social robot determines which drugs a patient requires, based on the patient records in the database? Or if an AI will issue a death certificate. Is there a distinction between an independently operating electronic system as an autonomous player and the use of this system as a tool? After all, in both cases the systems perform activities that have legal consequences.

Legal acts will be performed by persons, being legal entities. Automated systems, electronically or otherwise, are increasingly used in all kinds of relationships within our global society. Algorithms command the trading of the stock market and buy and sell within milliseconds. The fact that these systems, robots, and other devices can act independently and will create changes in legal relations will eventually have an effect on the position of legal persons, parties, or third parties.

What is, ultimately, the difference between the agent in human form, the natural person, and the robot representative?

Even in the case of natural persons, as an attributed representative who loses their reason and sanity, the proceedings may be annulled as a nondeliberate disturbance of the system. One can draw a parallel with the robot in the latter cases; it can reduce the liability of the initiating individual in the use of this system or can exculpate all parties of the legal action, maybe even the robot itself, if the robot has legal responsibility.

This view I share with Voulon, in the sense that any legal effect which is caused by an autonomous and less autonomous system must be attributed to the natural or legal person who has made the decision to commission the system in its service operations.[134] This reasoning is based upon the functioning of electronic agents, described as:

> A computer program, or electronic or other automated means used independently to initiate an action, or respond to electronic messages or performances, on the person's behalf without review or action by an individual at the time of the action or response to the message or performance.[135]

One would apply the level of liability of the person or entity related to the degree of control exercised over the autonomous system, thereby also taking the legal effect into account. However, this would only be the case with regard to liability and accountability to the natural or legal person. The malfunction or failure of the autonomic system can be significant with regard to the recognition of the actor's legal liability. The autonomous system itself, however, can never bear any legal responsibility until there is a degree of legal personality and a certain acceptance of a legal position to perform legal actions with legal effect. A public register where the scope of legal competence of this entity is to be consulted would be a solution to enhance credibility.

[132] Ibidem.

[133] Future of Life Institute [48] An Open Letter To The United Nations Convention On Certain Conventional Weapons. Available at: https://futureoflife.org/autonomous-weapons-open-letter-2017/. Accessed 21 August 2017.

[134] Voulon [43], concluding his dissertation.

[135] Section 102 (a) (27) Uniform Computer Information Transaction Act (UCITA).

Moreover, it would be helpful, in order to find a solution for this omission, to draw a parallel with the liability regulations as arranged in international regulations for electronic agents: the Uniform Electronic Transaction Act (UETA), the Uniform Computer Information Transaction Act (Ucita), and the Electronic Signatures Act (ES-ign). This could provide a model legal framework for autonomous entities to close agreements in a legally acceptable manner.

Ugo Pagallo presented the logical connection to existing forms of legal personhood for AI entities depending on their position and function, be it that more precise specifications of robot and their tasks can result in more specified legal subjectivity and legal competence:

1. "Independent legal personhood to robots with rights and duties of their own;
2. Some rights of constitutional personhood, such as those granted to minors and people with severe psychological illnesses, i.e., personhood without full legal capacity;
3. Dependent, rather than independent, personhood as it occurs with artificial legal persons such as corporations; and,
4. Stricter forms of personhood in the civil law field, such as the accountability of (some types of) robots for both contractual and extra-contractual obligations."[136]

As Ugo Pagallo concludes in another book concerning contracting capability: "artificial agents should be able to qualify for independent legal personality" based on the task they have to perform.[137]

18.10 Where to Go from Here?

Already projects are started with attempts to implant neural networks in robots to create biological robots. There are even speculations about the possibility for humans to donate their neurons or their whole brain in the future to live on in some way creating a form of immortality.[138] Envisioning possibilities like this is quite scary. Maybe some of the memories could actually be preserved that way, as there is evidence that the implanted neurons do really take on different roles like motor neurons and sensory neurons similar to how it works in a human body.

Moreover, this also entails the question if biological robots or robots with a simulated brain could develop consciousness and emotions like us humans. Anyhow, some researchers are skeptical that it is possible for robots to develop consciousness just because of a small selection of living cultural networks they have implanted. This is due to the many different types of cells in the human brain while nobody actually knows which cells in which combination are important for developing consciousness.

For a start AI algorithms can predict more and more of the human behavior, evolving from assisting to replacing and steering several economic and societal processes.[139]

Although an autonomous system or robot, even with an independent intelligence and emotion to function in our society, would not need to have a legal status that is similar to the rights and obligations of natural and legal persons in the positive law, change is imminent. The contours have to be defined. Even as an autonomous system passes the Turing test, this would not create any legal responsibilities per se. It is, however, advisable that certain forms of acting by autonomously functioning intelligent systems, such as social robots or legal enforcement robots,

[136] Going back to Teubner's analysis in the Rights of Nonhumans?, the entry of new actors on the legal scene concerns all the nuances of legal agenthood, such as "distinctions between different graduations of legal subjectivity, between mere interests, partial rights and full-fledged rights, between limited and full capacity for action, between agency, representation and trust, between individual, group, corporate and other forms of collective responsibility." Pagallo [19], p. 153 (referring to Teubner [49]).

[137] Hildebrandt and Gaakeer [50], p. 60.

[138] Randal Koene (http://rak.minduploading.org/ and https://read.bi/2lKAMqS).

[139] David C. Parkes and Michael P. Wellman, Economic reasoning and artificial intelligence, Science 17 July 2015: Vol. 349, Issue 6245, pp. 267–272. DOI: https://doi.org/10.1126/science.aaa8403.

may be conceivable to obtain a certain form of attributed legal personhood to carry out their tasks. This is based on the essential requirement that there is a social and legal necessity justifying such an attribution. This conception would also surpass the legal qualification of the integration between AI entities and human entities.

Already human enhancement is taking place, improving resistance and controlling physical processes according to the defense (DARPA) programs in the USA.[140] It is conceivable that in the near future robots could get human neurons or whole brains implanted. They would then be partially composed of human material making it more probable to get legal personhood attributed, as they would bear more similarity to humans than robots today making treatment equality more likely.

The legal positioning of robots could be selected for an amendment of the law or possibly even a sui generis standard for certain autonomous robots. This legal positioning will be dependent on the degree of autonomy and social need. For the qualification of the robots, the grading of the ISO standards can serve as an example.[141] In the International Standardization Organization already a development can be seen to treat the role of the robot differently (in security) and to accept a standard for robot/human collaboration.[142]

One might also imagine that certain changes are made to the existing law in order to create a practical system representation of autonomous systems for the initial legal actor, the natural or legal person. These changes in the law will depend on a correct description of the reliability and trust of the representation by the robot, the purpose of the actions, and the legal consensus of the legal entities involved. If these concepts are agreed upon, it will then be necessary to obtain the acceptance by the government and parliament to create or adapt a legal framework. As to how difficult and time-consuming this process will be, reference can be made to the acceptance of the non-natural person in the positive law. The comparison with the rational, responsible actor as presented by Naffine probably will result in too many problems but certainly elements of this reasoning could be of help.

Currently, many AI systems are very difficult for users to understand. This is also increasingly true for those who develop the systems. In particular, neural networks are often “black boxes,” in which the (decision-making) processes taking place can no longer be understood and for which there are no explanatory mechanisms.[143] This could necessitate a legal requirement to create a form of transparency as to how the systems work, to enhance trust and credibility of the acts leading to legal effect as also proposed in the EP motion on civil law rules on robotics.

AI and autonomous robots will be part of our future society. Integration of AI inside the human body will also occur. Our physical and informational integrity will be invaded, with or without our knowledge or consent. We already share a substantial part of our personal data with third parties and appear not really concerned by it. On top of that, governments and industries are forcing us to share even more personal information to regulate or protect the social system or to lower risks and costs of services and products. More knowledge about the brain and its functioning could also lead to ways to improve memory for instance. This brings with it another problem: To whom would such human enhancement options be open? Probably only to the richest part of the population. These are all points that need to be taken into account by politicians, lawyers, and scientists working in the field of robotics, artificial intelligence, and neuroscience.

The European General Data Protection Regulation (GDPR) describes the protection of

[140] See i.e. the “PREPARE” project (https://www.darpa.mil/news-events/2018-05-25).

[141] See, e.g., ISO 13482: 2014 Specifies requirements and guidelines for the inherently safe design, protective measures, and information for use or personal care robots, in particular the following three types of personal care robots: mobile robot servant, physical assistant robot, and person carrier robot.

[142] Human and robot system interaction in industrial settings is now possible thanks to ISO/TS 15066, a new ISO technical specification for collaborative robot system safety.

[143] Hildebrandt and Gaakeer [50], p. 7.

personal data during processing in outdated terminology concerning AI.[144] Due to the non-technological orientation and the hinge on conventional directions of thinking, it is hard to consider the GDPR sufficient to protect personal data in the age of AI.

Informational rights for the data subject and transparency of the process cannot be applied to integrated AI, certainly not if this is integrated into the physical functions of the human being. There is a significant risk of chilling effects for the development of AI and robotics if the GDPR has to be enforced on all AI applications.

In a report of the Science and Technology Committee of the UK Parliament, the need for unhindered but controlled applications of AI technology is stressed:

> It is important to ensure that AI technology is operating as intended and that unwanted, or unpredictable, behaviors are not produced, either by accident or maliciously. Methods are therefore required to verify that the system is functioning correctly. According to the Association for the Advancement of Artificial Intelligence: it is critical that one should be able to prove, test, measure and validate the reliability, performance, safety and ethical compliance—both logically and statistically/probabilistically—of such robotics and artificial intelligence systems before they are deployed.[145]

> The "Big Brother Watch" has a rather naïve point of view on the possibility of transparency of AI in the GDPR as proclaimed in the earlier mentioned UK House of Lords document on AI.[146]

For this reason, it will be necessary to develop some form of certification to determine whether the autonomously functioning robot can be accepted to process data of third parties and perform acts with legal capacity. Which interaction would be considered acceptable between parties will vary, depending on the function and of course the requirements of technological measures of protection of the robot as described above.

Horst Eidenmüller gave the following principles for a legal structure for AI (robots):

> (i) Robot regulation must be robot- and context-specific. This requires a profound understanding of the micro- and macro-effects of 'robot behaviour' in specific areas. (ii) (Refined) existing legal categories are capable of being sensibly applied to and regulating robots. (iii) Robot law is shaped by the 'deep normative structure' of a society. (iv) If that structure is utilitarian, smart robots should, in the not too distant future, be treated like humans. That means that they should be accorded legal personality, have the power to acquire and hold property and to conclude contracts. (v) The case against treating robots like humans rests on epistemological and ontological arguments. These relate to whether machines can *think* (they cannot) and what it *means* to be human.[147]

It is essential that we, as people, maintain control of the system as long as this has an added value. We would not want to be confronted with

[144] Regulation (EU) 2016/679.

[145] Interesting is the concluding recommendation of the Science and Technology Committee: "73. We recommend that a standing Commission on Artificial Intelligence be established, based at the Alan Turing Institute, to examine the social, ethical and legal implications of recent and potential developments in AI. It should focus on establishing principles to govern the development and application of AI techniques, as well as advising the Government of any regulation required on limits to its progression. It will need to be closely coordinated with the work of the Council of Data Ethics which the Government is currently setting up following the recommendation made in our Big Data Dilemma report. 74. Membership of the Commission should be broad and include those with expertise in law, social science and philosophy, as well as computer scientists, natural scientists, mathematicians and engineers. Members drawn from industry, NGOs and the public, should also be included and a programme of wide ranging public dialogue instituted." Available at: https://publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/14506.htm#_idTextAnchor014. Accessed 25 October 2017.

[146] "Because AI can fundamentally impact a person's life, moves should be undertaken to ensure that the transparency of AI programs is standard, particularly when AI is used to make a decision affecting people or impact how people live their lives. The public must always be fully aware of when they are subject to, or affected or impacted by a decision made by AI. Increased transparency and accountability of public-facing AI, including the methods behind the system, and the reasons for decisions, will not only benefit society as a whole in terms of open source information but will increase public trust and confidence and subsequently, public engagement with AI systems." (https://www.parliament.uk/documents/lords-committees/Artificial-Intelligence/AI-Written-Evidence-Volume.pdf), p. 140.

[147] https://www.law.ox.ac.uk/business-law-blog/blog/2017/04/rise-robots-and-law-humans

autonomous systems, which use the collection of all kinds of personal information and other available data for their own purposes. But, on the other hand, AI technology can only develop without chilling effects if it is commercially admitted to the consumer's daily life without too much legal constraint. The existence of a sui generis structure, comparable with the case of the artificial legal person in corporate law, may provide a solution. The Naffine definition of the Cheshire Cat combined with Rational Actor can form a rational basis for a legal framework comparable with the existing position of artificial legal persons.

At least, the following requirements of the AI entity have to be fulfilled to acquire a sui generis legal personhood:

1. "Necessity in the 'human' society, socio-economic relevance, need for legal certification;
2. Determination of autonomous intelligence, Turing test like, 'human impression' level;
3. Sufficient social intelligence; The AI entity must be able to understand the socio-emotional and moral value of statements by other parties to respond appropriately so that there is an equivalent basis for consensus;
4. Being able to respond to changing circumstances; this aspect I would call 'adaptive or dynamic' intelligence;
5. Acceptance by other legal persons by creating trust and reliance for other legal and natural persons to integrate in economic, social and legal interactions;
6. A public register that specifies which robots will have specific legal competences for specified roles and tasks."

On top of this, an ethical code has to be developed on the basis of the EP motion that should also consider the use of different categories of robots, as well as the default rules needed for developers and producers of robotics.[148]

We are better off using our electronic, or better, technology-based servants to help us with the practical performance of our duties. The more intelligent the system is, all the more reliable the functionality will be. Give the robot a place in our legal system, maybe even with a form of digital peculium as proposed by Pagallo, giving them a limited resource that could also be used as a guarantee for possible mistakes or damages, and open the possibility of accountability for their autonomous acts. In a more extensive elaboration of this idea, one could establish a fund financed by a certain percentage of the earnings by robots to guarantee any losses or damages. Though it will have to be a select group of AI entities that qualify for a new form of legal personhood and economic personality. In that respect, the robot will be active in the social and economic functioning of society. This can also concern the public sector. A certain trust in the acts of robots and recognition of their identity will prove to be essential.

On top of that a utilitarian, sui generis legal position does not result in a comparable legal status comparable with natural persons. The protest against the European Parliament motion on legal status of electronic persons in an open letter by a number of experts seems to be a bit "over the hill" because robots will not be human(yet):

> A legal status for a robot can't derive from the Natural Person model, since the robot would then hold human rights, such as the right to dignity, the right to its integrity, the right to remuneration or the right to citizenship, thus directly confronting the Human rights. This would be in contradiction with the Charter of Fundamental Rights of the European Union and the Convention for the Protection of Human Rights and Fundamental Freedoms.[149]

But we have to keep in mind that we still have to control the developments and not end up with the rather pessimistic post-human idea described by Yuval Noah Harari in his famous book Homo Deus. In this account, science will move in the direction that all organisms are algorithms, life is data processing, intelligence will

[148] The proposed code of ethical conduct in the field of robotics will lay the groundwork for the identification, oversight, and compliance with fundamental ethical principles from the design and development phase. EP motion, PE582.443v03-00, p. 21.

[149] https://bit.ly/2xfMToe

be separated from consciousness, and the hyper-intelligent algorithms will know us better than we know ourselves.[150] Even if super-intelligent algorithms will decide how society and humans develop we must not forget we will be part, integrated with AI or not, of the development of our own future.

References

1. Vinge V. The coming technological singularity: how to survive in the post-human era. NASA, Lewis Research Center, Vision 21. 1993. p. 11–22. Available at: https://edoras.sdsu.edu/~vinge/misc/singularity.html. Accessed 25 Oct 2017.
2. Clarke AC. Profiles of the future: an inquiry into the limits of the possible. New York: Harper & Row; 1973.
3. Miller CA, Bennett I. Thinking longer term about technology: is there value in science fiction-inspired approaches to constructing futures? Sci Public Policy. 2008;35(8):597–606.
4. Burkhart L. Symposium – governance of emerging technologies: law, policy, and ethics. Jurimetrics. 2016;56:219–22. Available at: https://www.americanbar.org/content/dam/aba/administrative/science_technology/2016/governance_in_emerging_technologies.authcheckdam.pdf. Accessed 12 Sept 2017.
5. Tjong Tjin Tai, TFE. Private law for homo digitalis, use and maintenance. Preliminary Advice for NVJ. 2016. p. 248.
6. Geldart WM. Legal personality. Law Q Rev. 1911;27:90–108.
7. Richards NM, King JH. Big data ethics. Wake Forest Law Rev. 2014;49:393–432.
8. Berriat Saint-Prix J. Rapport et Recherches sur les Procès et Jugemens Relatifs aux Animaux. Paris: Imprimerie de Selligue; 1829.
9. Delvaux M. Report PE582.443v01-00 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)). 2017. Available at: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+REPORT+A8-2017-0005+0+DOC+PDF+V0//EN. Accessed 8 Dec 2017.
10. Bertolini A. Robots as products: the case for a realistic analysis of robotic applications and liability rules. Law Innov Technol. 2013;5(2):214–27.
11. Solum LB. Legal personhood for artificial intelligences. North Carol Law Rev. 1992;70(4):1238–9.
12. Ohlin JD. Is the concept of person necessary for human rights? Columbia Law Rev. 2005;105:209–49.
13. Fagundes D. What we talk about when we talk about persons: the language of a legal fiction. Harv Law Rev. 2001;114(6):1745–68.
14. Naffine N. Who are law's persons? From Cheshire Cats to responsible subjects. Mod Law Rev. 2003;66(3):346–67.
15. Hutchinson A. The Whanganui River as a legal person. Altern Law J. 2014;39(3):179–82.
16. Brownlie I. Principles of public international law. London: Clarendon Press; 1990.
17. Crawford JR. Brownlie's principles of public international law. 8th ed. Oxford: Oxford University Press; 2012.
18. Hobbes T. Chapter xvi: of persons, authors, and things personated. In: Hobbes T, editor. Leviathan. London: Andrew Crooke; 1651.
19. Pagallo U. The laws of robots: crimes, contracts, and torts. Dordrecht: Springer; 2013.
20. Bodin J. Les Six Livres de la Republique (Translation by MJ Tooley). Oxford: Blackwell; 1955.
21. Dewey J. The historic background of corporate legal personality. Yale Law Rev. 1926;35(6):655–73.
22. Minors D. Can you read my mind? 2017. Available at: https://www.wits.ac.za/news/latest-news/research-news/2017/2017-09/can-you-read-my-mind. Accessed 11 Oct 2017.
23. The Global Slavery Index. 2016. Available at: https://www.globalslaveryindex.org/findings/. Accessed 12 Oct 2017.
24. Descartes R. Principia philosophiae. Paris: Vrin; 1973.
25. Gardner H. The theory of multiple intelligences. New York: Basic Books; 1993.
26. Wechsler D. The range of human capacities. Baltimore: Williams & Wilkins; 1955.
27. Turing AM. Computing machinery and intelligence. Mind, New Series. 1950;59(236):433–60.
28. Darling K. Electronic love, trust, & abuse: social aspects of robotics. Workshop "We Robot" at the University of Miami. 2016.
29. Mayer CJ. Personalizing the impersonal: corporations and the bill of rights. Hastings Law J. 1990;41(3):577–667.
30. Shoyama. Intelligent agents: authors, makers, and owners of computer-generated works in Canadian copyright law. Can J Law Technol. 2005;4(2):129.
31. Russell S, Norvig P. Artificial intelligence: a modern approach. 3rd ed. Upper Saddle River, NJ: Pearson Education; 2010.
32. Kurzweil R. The age of intelligent machines. Cambridge: The MIT Press; 1990.
33. Bostrom N. Superintelligence: paths, dangers, strategies. Oxford: Oxford University Press; 2014.
34. Good IJ. Speculations concerning the first ultraintelligent machine. In: Alt FL, Rubinoff M, editors. Advances in computers, vol. 6. New York: Academic Press; 1965. p. 31–88.

[150] Harari [51], last sentences.

35. Giliker P. Vicarious liability or liability for the acts of others in tort: a comparative perspective. J Eur Tort Law. 2011;2(1):31–56.
36. Von Bar C, Clive E, editors. Principles, definitions and model rules of European private law: Draft Common Frame of Reference (CDFR). Munich: Sellier. European Law Publishers GmbH; 2009.
37. Schaerer E, Kelley R, Nicolescu M. Robots as animals: a framework for liability and responsibility in human-robot interaction. In: Robot and Human Interaction Communication. RO-MAN 2009 – The 18th IEEE International Symposium on Robot and Human Interactive Communication. 2009. p. 72–7.
38. Erven D onder de Linden en zoon. Boekzaal der geleerde wereld: en tijdschrift voor de Protestantsche kerken in het koningrijk der Nederlanden. 1831. p. 201–3.
39. Mori M. The uncanny valley: the original essay by Masahiro Mori. 2012. Available at: https://spectrum.ieee.org/automaton/robotics/humanoids/the-uncanny-valley. Accessed 15 Oct 2017.
40. Safi M. Ganges and Yamuna rivers granted same legal rights as human beings. 2017. Available at: https://www.theguardian.com/world/2017/mar/21/ganges-and-yamuna-rivers-granted-same-legal-rights-as-human-beings. Accessed 13 May 2017.
41. Lovejoy AO. The great chain of being: a study of the history of an idea. Cambridge: Harvard University Press; 1936.
42. Wiener N. The human use of human beings. London: Eyre & Spottiswoode; 1950.
43. Voulon MB. Automatisch contracteren. Dissertation, Leiden University. 2010.
44. Raskin M. The law and legality of smart contracts. Georgetown Law Technol Rev. 2017;304(1). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2959166. Accessed 20 Oct 2017.
45. Segrave K. Vending machines: an American social history. Jefferson, NC: McFarland; 2002.
46. Carlile R. To the republicans of the Island of Great Britain. Republican. 1822;16(V), (see also chapter 10, digital version [https://bit.ly/2CduCY1]).
47. Chopra S, White LF. A legal theory for autonomous artificial agents. Ann Arbor, MI: The University of Michigan Press; 2011.
48. Future of Life Institute. An open letter to the United Nations convention on certain conventional weapons. 2017. Available at: https://futureoflife.org/autonomous-weapons-open-letter-2017/. Accessed 21 Aug 2017.
49. Teubner G. Rights of non-humans? Electronic agents and animals as new actors in politics and law. Florence: European University Institute; 2007.
50. Hildebrandt M, Gaakeer J, editors. Human law and computer law: comparative perspectives. Dordrecht: Springer; 2013.
51. Harari YN. Homo Deus: a brief history of tomorrow. London: Random House; 2017.
52. Asimov I. The bicentennial man and other stories. London: Victor Gollancz; 1976.
53. Asimov I, Silverberg R. The positronic man. New York: Doubleday; 1993.
54. Bryson JJ, Diamantis ME, Grant TD. Of, or, and by the people: the legal lacuna of synthetic persons. Artif Intell Law. 2017;25:273–91.
55. Science and Technology Committee. Robotics and artificial intelligence. 2016. Available at: https://publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/14506.htm#_idTextAnchor014. Accessed 25 Oct 2017.

The Role of an Artificial Intelligence Ecosystem in Radiology

19

Bibb Allen, Robert Gish, and Keith Dreyer

19.1 Defining Business Ecosystems

Fueled by the ever-increasing amount of data generated by the healthcare system and the recent exponential advances in computing power detailed in previous chapters, artificial intelligence (AI) applications for healthcare, especially within diagnostic imaging, are rapidly proliferating [1]. Artificial intelligence promises the transformation of massive volumes of generated data, which exceeds the capacity of the human mind, into actionable data usable by healthcare stakeholders. However, currently no well-defined framework exists for determining how great ideas for AI algorithms in healthcare will advance from developmentto integrated clinical practice. For the most part, individual AI software developers are currently working with individual radiologists within single institutions to create AI algorithms focused toward the targeted needs of those individual health systems. One central challenge for radiology informatics is generalizing these single site and limited AI applications to routine clinical practice across a wide range of patient populations, electronic health records, imaging equipment systems, and imaging protocols [1]. Healthcare stakeholders including physicians, patients, medical societies, hospital systems, software developers, the health information technology industry, and governmental regulatory agencies all comprise a community that will have to function as a system in order for AI algorithms to be deployed, monitored, and improved in widespread clinical practice. The community of interacting stakeholders is defined here as an "Artificial Intelligence Ecosystem" for healthcare and the radiological sciences. In a recent report, the JASON Advisory Group identified several key recommendations for advancing computation technology into routine clinical practice as follows [2]:

- New technologies should address a significant clinical need.
- Technology must perform at least as well as the existing standard approach, i.e., demonstration of statistical non-inferiority.

B. Allen (✉)
Department of Radiology, Grandview Medical Center, Birmingham, AL, USA

American College of Radiology Data Science Institute, Reston, VA, USA

R. Gish
Diagnostic Radiology, Brookwood Baptist Health, Birmingham, AL, USA

K. Dreyer
American College of Radiology Data Science Institute, Reston, VA, USA

Department of Radiology, Massachusetts General Hospital, Boston, MA, USA

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_19

- Substantial clinical testing must validate the new technology under the wide range of clinical situations and varying patient populations.
- New technology should provide improvements in patient outcomes, patient quality of life, practicality in use, and reduced medical system costs.

Achieving these goals will require a coordinated approach between multiple stakeholders to move safe and effective AI tools into clinical practice; defining and developing a cohesive artificial intelligence ecosystem will facilitate AI implementation into clinical practice.

In biology, an ecosystem is defined as "the complex of a community of organisms and its environment functioning as an ecological unit" [3]. In 1993, James F. Moore used the term ecosystem to describe the complex business interactions when companies work cooperatively and competitively to develop capabilities around new innovations that support new products, satisfy customer needs, and set the stage for the next round of innovations [4, 5]. He later defined a "business ecosystem" as

> An economic community supported by a foundation of interacting organizations and individuals—the organisms of the business world. The economic community produces goods and services of value to customers, who are themselves members of the ecosystem. The member organisms also include suppliers, lead producers, competitors, and other stakeholders. [4]

In both of these definitions, defining a community of oftentimes disparate stakeholders and understanding the role each play are critical to the success of the community as a whole. Nowhere in business is the term ecosystem more applicable than in the technology and software development industries. In their book, *Software Ecosystem*, Messerschmitt and Clemmons define the community for software development around six groups: users, software developers, managers, industrialists, policy experts and lawyers, and economists [6]. At the beginning, end users of the software products must define what it is that they want the software to accomplish for them. Software developers and engineers then translate the users' needs to program code, and then a group of managers must coordinate resources to bring the software product into the end users' workflow. Companies must be formed to mass distribute the software product, and policy experts and legal teams must ensure there are no legal or other barriers to software implementation. Economists then offer insights into how the software market works. In modern terms, software developers also find themselves within a subecosystem where the software they are writing is being built on top of platforms such as high-level coding languages and operating systems or below platforms such as web pages where their software outputs are designed to be inputs consumed by other software products. In almost all cases, the final software product employed by end users is not the code written by the developer but the results of the output of a compiler taking the instructions written by the software developer, which are then converted to lower-level machine-readable code that becomes the program executed by the computer. All of these additional interactions are continually expanding the community within the software development ecosystem [6].

For software development to be effective in healthcare, another ecosystem must be considered and that is the healthcare community itself. The healthcare ecosystem is an incredibly complex reciprocal network with a seemingly innumerable number of categories of human actors interacting within a similarly vast number of electronic resources and existing software tools (Fig. 19.1).

In addition, the healthcare industry is highly regulated by national and international governing bodies worldwide [7]. Much of this regulation, designed to promote quality and ensure patient safety and privacy, is often not encountered in other fields of software development. Finally, most of healthcare worldwide is not paid for directly by the patients themselves but rather by governmental or other third-party payers such as commercial insurance companies. In the United States, payments to providers are

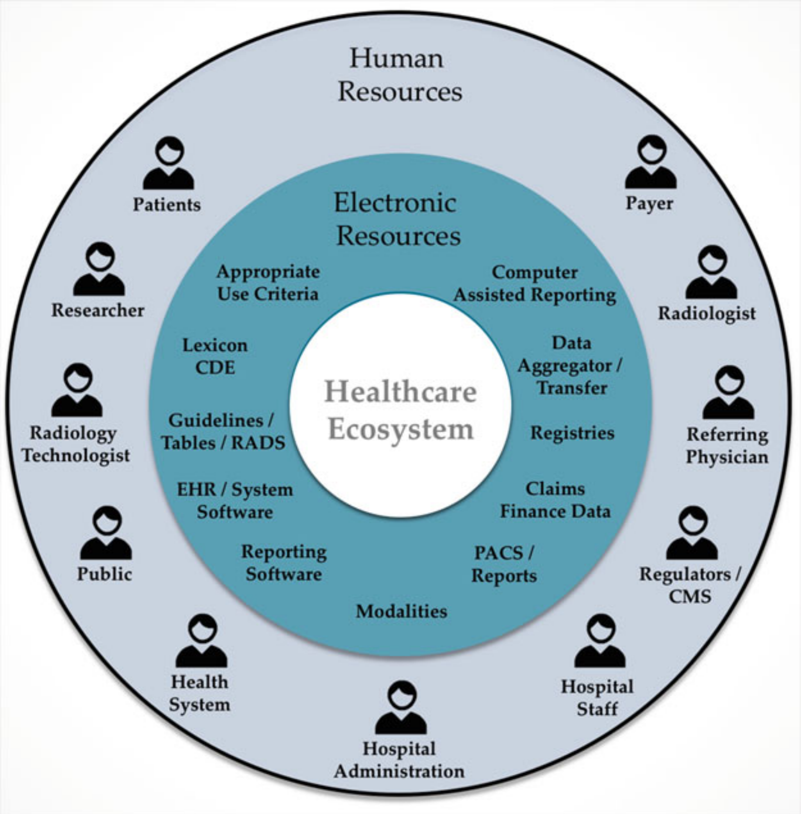


Fig. 19.1 Healthcare ecosystem (Credit authors)

typically made on a fee-for-service basis; however, there is a growing percentage of the population covered under alternate payment models such as accountable care organizations and other forms of population-based health management. Internationally, many countries have public health systems paid for from tax revenue and furnished at no cost to permanent residents. In these systems, physicians and other providers are paid salaries from the government. Many countries including the United States, have developed a hybrid system of both publicly and privately funded healthcare. However, in the United States, federally funded healthcare programs such as Medicare and Medicaid only cover about 36% of the population, whereas employer-based private insurance plans cover approximately 47% of the population [8]. Although variable internationally, governmental programs cover the vast majority of the population in developed nations.

19.2 Artificial Intelligence Ecosystem for Healthcare and Diagnostic Imaging

In order to develop and maintain an ecosystem for artificial intelligence in medicine, both the software development and healthcare stakeholder communities must be considered. While developers of artificial intelligence applications may be well acquainted with the software development ecosystem, they may be unfamiliar with the numerous idiosyncrasies of the healthcare ecosystem. On the other hand, in order for the healthcare system to become a viable market for AI applications, the healthcare community must provide developers with the clinical and technical challenges and proposed solutions to enable the generation of ideas, tools, and pathways for clinical integration that will make AI algorithms

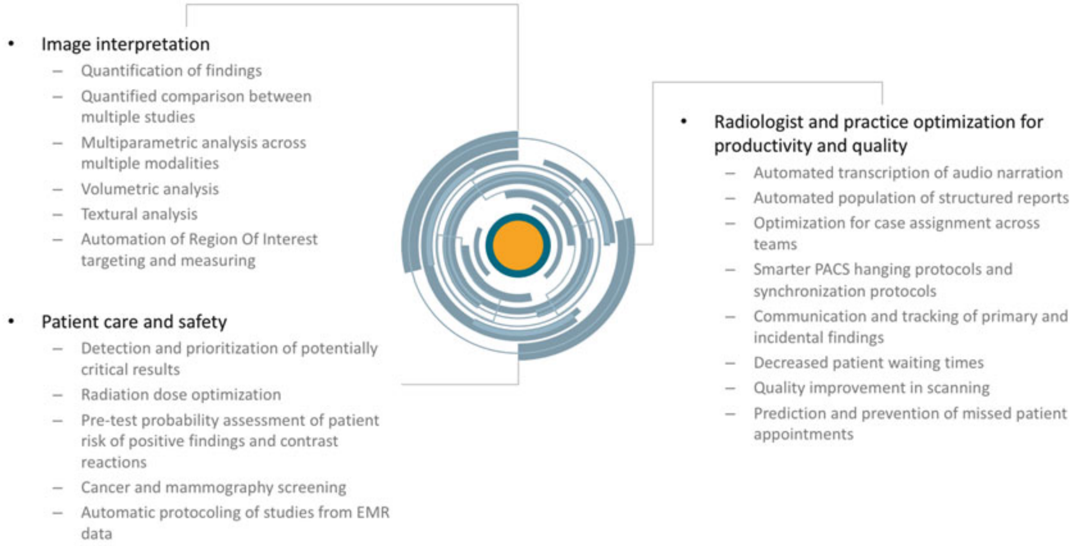


Image interpretation
– Quantification of findings
– Quantified comparison between multiple studies
– Multiparametric analysis across multiple modalities
– Volumetric analysis
– Textural analysis
– Automation of Region Of Interest targeting and measuring
Patient care and safety
– Detection and prioritization of potentially critical results
– Radiation dose optimization
– Pre-test probability assessment of patient risk of positive findings and contrast reactions
– Cancer and mammography screening
– Automatic protocoling of studies from EMR data
Radiologist and practice optimization for productivity and quality
– Automated transcription of audio narration
– Automated population of structured reports
– Optimization for case assignment across teams
– Smarter PACS hanging protocols and synchronization protocols
– Communication and tracking of primary and incidental findings
– Decreased patient waiting times
– Quality improvement in scanning
– Prediction and prevention of missed patient appointments

Fig. 19.2 Use cases for radiology (Source: American College of Radiology Data Science Institute. Used with permission)

indispensable in clinical practice. In 2008, Berwick and colleagues introduced the concept of the Triple Aim for healthcare in the United States [9]. The Triple Aim, which has been widely adopted as a model for improvement in US healthcare, proposed three areas for performance improvement in medicine. The healthcare system should work to improve the overall health of the population while at the same time improve the individual experience of care and reduce per capita costs. Subsequently, Bodenheimer and Sinsky expanded these goals to include a fourth aim, which is improving the work life of healthcare providers [10]. Similarly, Sikka, et al., from the United Kingdom, proposed a similar fourth aim, improving the experience of providing care [11]. The Quadruple Aim for delivering high-value care is now recognized worldwide as a guiding set of principles for health system reform. Artificial intelligence applications are poised to assist health systems meet the goals of the Quadruple Aim across all of healthcare, and in medical imaging the possibilities for AI to improve radiological practice are almost endless. Radiologists are uniquely positioned to be at the forefront of the coming benefits of AI applications [12]. As shown in Fig. 19.2, these AI applications will not only help improve diagnostic accuracy by "seeing" the relevant features human radiologists already extract from images, these algorithms will perform quantitative analysis on imaging features that may be beyond the scope of human vision and difficult for a radiologist to convey using natural language.

However, AI applications for medical imaging will not be limited to image interpretation. AI algorithms will be able to improve patient safety by prioritizing patient imaging worklists and enhancing the communication of critical findings. Imaging protocols can be automated based on information gathered from the electronic health record (EHR) and tailored to optimize radiation exposure [13]. AI could directly optimize the reading radiologist's experience by mining the EHR for patient data including patient problem lists, clinical notes, laboratory data, pathology reports, vital signs, prior treatments, and prior imaging reports and generating a relevant summary to give the reading radiologist the most pertinent contextual information during the interpretation of a study. Another example of a seemingly simple application would be the optimization of hanging protocols. Hanging protocols are currently often disrupted by sequence and plane acquisition order as well as the order of manual entry into PACS by the radiology technologist. AI could be developed to classify image sequences, planes, and contrast phases, and then place them into the preferred order of the individual radiologist. Artificial intelligence tools will also be able to improve radiology department efficiency by optimizing workflow, automation of structured reporting systems, and improving patient experience by decreasing patient wait times and avoiding missed appointments [13, 14]. For all of this to happen, an artificial intelligence ecosystem specific for diagnostic imaging must be developed and supported by all stakeholders including the medical imaging community, the software development community, and the governmental agencies through regulatory processes providing an appropriate balance between fostering innovation, moving new products to market, and patient safety [1].

19.3 Defining an Artificial Intelligence Ecosystem in Healthcare with a Focus on Diagnostic Imaging

19.3.1 Establish Realistic Goals

As the modern ecosystem to support the advancement of artificial intelligence in medical imaging is developed, consideration must be given to how AI has evolved over the years within the research and business communities. John Huffman, the Chief Technology Officer for Informatics Solutions and Services at Phillips Healthcare, presented a plot of research and entrepreneurial activity over the last 70 years at the Healthcare Information and Management Systems Society [HIMSS] March 6, 2018 meeting (Fig. 19.3) [15]. The current rise in enthusiasm for artificial intelligence is actually the third AI Spring; however, each of the two previous AI Springs was followed by an AI Winter.

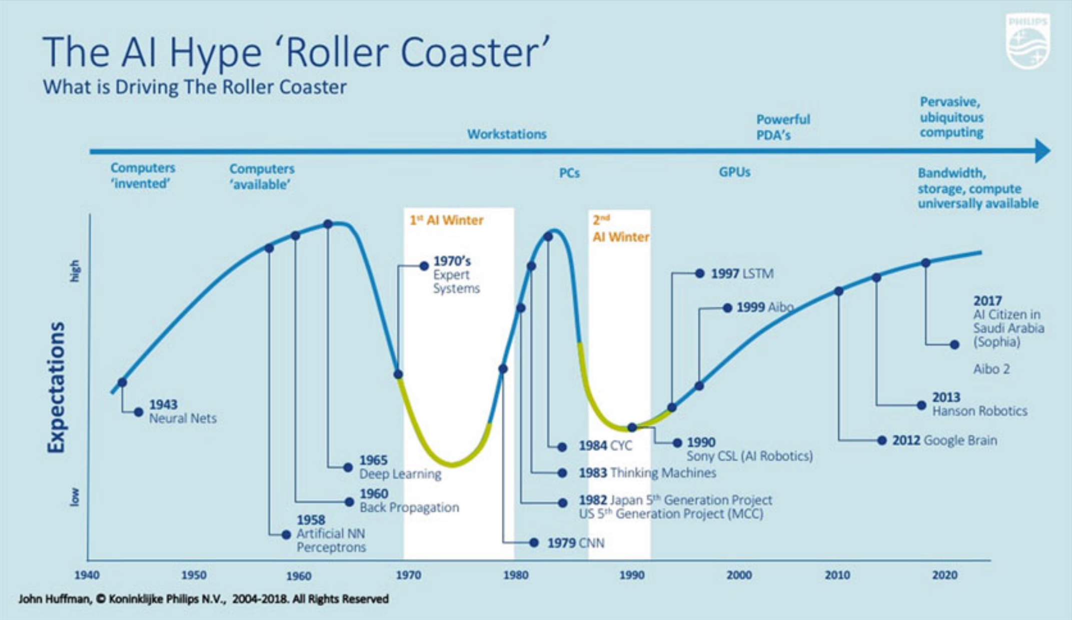


Fig. 19.3 Timetable graphic of the interest in artificial intelligence over the past 70 years. Two golden ages of artificial intelligence have each been followed by "winters" during which little progress was made in AI. The third golden age is now underway (John Huffman, © Koninklijke Philips N.V., 2004–2018. All Rights Reserved. Used with permission)

When research in artificial intelligence began in the 1940s and 1950s, the goal was to create all-knowing computers that could ingest the entirety of world knowledge and totally duplicate the cognitive and reasoning activities of humans. This excitement and anticipation lead to the work of Alan Turing in the 1950s resulting in the famous Turing test, which is an assessment of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human [16]. Fueled by the work of Marvin Minsky [17] and John McCarthy [18], the expectations that computers would able to mimic the tasks of the human brain were high; however, as it became clear that computer processing power was woefully inadequate to support this research, the number of investigators and interest in AI research began to wane. A second spike in AI entrepreneurialism occurred in the 1980s with early companies like Thinking Machines Corporation, founded by Minsky, were quite profitable through the early 1990s [19]. However, despite having the highest level of processing power in the industry, these companies also failed to develop significant AI products, and enthusiasm over AI again waned. As the third AI Spring continues to gain momentum, the AI community must learn from its predecessors in order to avoid another AI Winter. However, the renewed enthusiasm for AI over a large and diverse number of industries has once again caused expectations for AI to soar. Over the last 10 years, success in these decades-old technologies, such as multilayered neural networks, has been fueled by advances in fast hardware parallel graphical process unit (GPU) computing [20] allowing training of more and progressively deeper neural networks [21]. The combination of advances in technology and availability of large annotated datasets for testing has given rise to the concept of deep learning on layered neural networks termed convolution neural networks [2]. The use of these and other modern techniques has once again escalated claims about the imminent rise of all knowing computers duplicating the cognitive activity of humans. There are also a number of additional factors that have increased the enthusiasm for AI in healthcare. Previously there were few ways applications of artificial intelligence touched the daily lives of the population in general. However, developments such as self-driving cars, mechanized manufacturing robotics, and wearable personal health monitors are paving the way for broader acceptance and applications of AI in healthcare with the ability to not only better the lives of the population as a whole but also to impact the lives of individuals [22]. Finally, informatics solutions that can bend the cost curve in healthcare will be readily accepted as the cost of healthcare continues to rise (Fig. 19.4).

Despite these compelling reasons for AI to have a major impact on healthcare worldwide, expectations must be realistic. However, seemingly unrealistic promises about the capabilities of AI in healthcare abound. Much of what appears in the lay media around the promise of AI for healthcare, including replacing radiologists, seems to focus almost entirely on what has been termed artificial general intelligence (AGI). Artificial general intelligence (also called broad AI or strong AI) is an application of AI where machines are designed to fully duplicate the cognitive activity of humans, that is, perform the same intellectual tasks in the same context as humans. Speculation that advances in general AI will soon create systems that will replace radiologists abound. At a meeting of the American College of Radiology in 2016, Zeke Emmanuel told an audience of nearly 1000 radiologists, radiation oncologists, and medical physicists, including many trainees and medical students, that artificial intelligence will take away radiologists' jobs and will be the biggest threat to the specialty over the next decade [23]. Former US president Barrack Obama speculated that radiologists would be replaced by radiologist robots in the same way driverless cars will replace Uber drivers [24], and perhaps most notably Geoff Hinton, considered by many to be the father of deep learning, told a large audience at a Creative Lab Conference in 2016 that:

> If you work as a radiologist you are like Wile E. Coyote in the cartoon. You're already over the edge of the cliff, but you haven't yet looked down. It's just completely obvious that in five years deep learning is going to do better than radiologists.

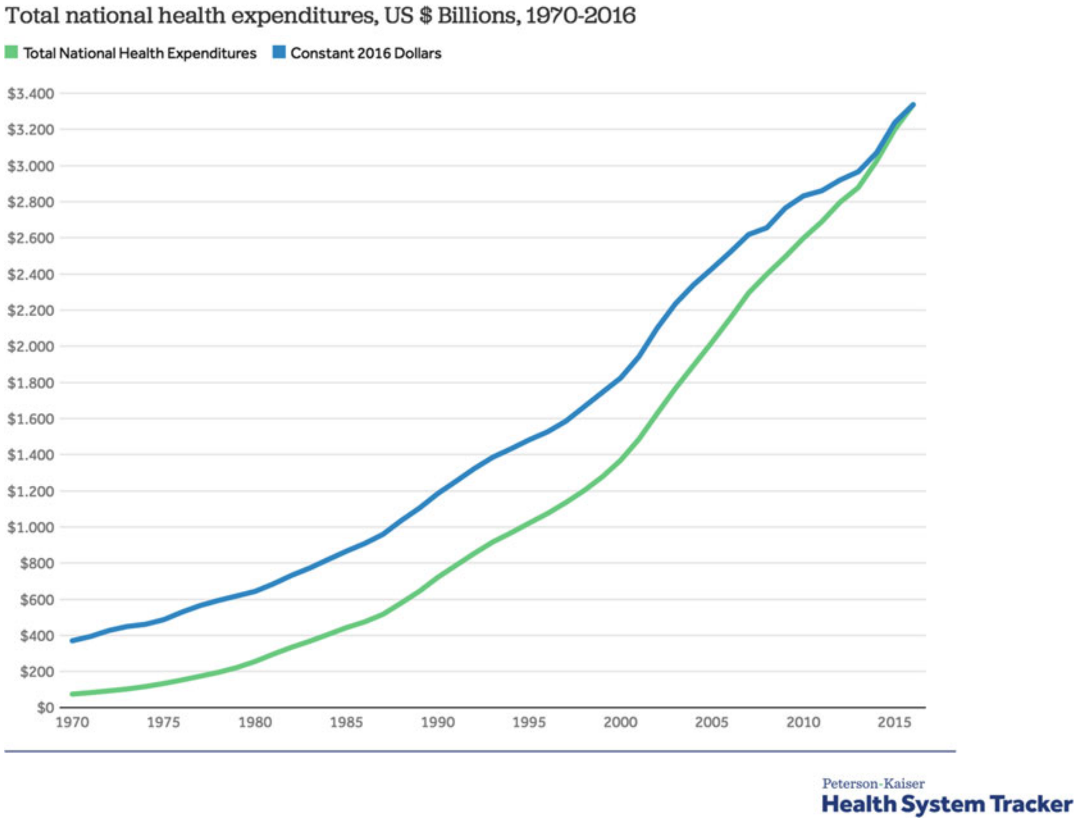


Fig. 19.4 Healthcare costs in the United States (Credit Kaiser Family Foundation analysis of National Health Expenditure (NHE) data from Centers for Medicare and Medicaid Services, Office of the Actuary, National Health Statistics Group. Used with general reuse permissions from the US Centers for Medicare and Medicaid Services). Source: Kaiser Family Foundation analysis of National Health Expenditure (NHE) data from Centers for Medicare and Medicaid Services, Office of the Actuary, National Health Statistics Group, Get the data, PNG

> It *might* be ten years. We should stop training radiologists now. [25]

These are just a few examples of where some in the informatics community continue to overpromise. At this point, there is nothing to suggest that artificial intelligence will replace physicians, cure cancer, or even prolong life expectancy, but to ensure AI algorithms that will help physicians provide better patient care are adopted into clinical practice, developers should focus on specific narrow use cases with readily defined and achievable outcomes. Algorithm training and validation must occur using methods ensuring the results of the algorithm will demonstrate interoperability in widespread clinical practice, and physicians and other end users must be able to understand how the algorithm reached its conclusions in order for the efficacy of the algorithm inferences to be evaluated and communicated on a patient by patient basis.

19.3.2 Maintain a Targeted Focus

As radiologists consider defining an ecosystem to support the development of artificial intelligence applications in medical imaging, the radiology community must consider how algorithms will be developed and trained. Although broad AI with unsupervised learning gets most of the hype, it still remains best suited for science fiction movies. To date, broadly applied AI in healthcare has had mixed results. Notably, IBM Watson's collaborative project with MD Anderson Cancer Center, Oncology Expert Advisor [26], which is a cognitive computing platform providing decision support to cancer care providers, had to be put on hold because the goals and efficiencies that were expected were never realized [27]. Many cited the data quality used in this general AI model with generally continuous unsupervised learning from the medical records as the main problem, and it is likely that training the algorithms on unstructured data with no specific use cases made developers believe the timeline for creating a fully trained and reliable product could be accelerated. Others have asserted that for AI to have high impact in healthcare in general and radiology in particular, developers should focus on narrow AI with structured use cases using supervised learning with training on high-quality structured and carefully annotated data [1, 12]. While intuitively one might believe the natural evolution of AI would be from narrow AI to general AI, the actual progression has been just the opposite. As shown in Fig. 19.5, until the last decade, most AI applications have been generally considered broad AI.

Increasing computing power and modern AI techniques such as deep neural networks have increased the ability to rapidly develop specific uses for AI that can be implemented into physician workflows. In order for AI to be successful in healthcare and medical imaging, development should continue to be focused on producing high-quality, clinically useful AI use cases where algorithms can be trained on high-quality structured data in order to assist radiologists solve specific problems.

Although a detailed discussion of the application of specific artificial intelligence techniques for medical imaging is beyond the scope of this chapter, it is important to understand some of the ways AI inference models for medical imaging will be created to inform how an AI ecosystem for medical imaging can support the development of robust AI tools for the medical imaging community. Built on a foundation of artificial neural networks, deep learning is emerging as the predominant tool for artificial intelligence applications in healthcare [28]. Machine learning has

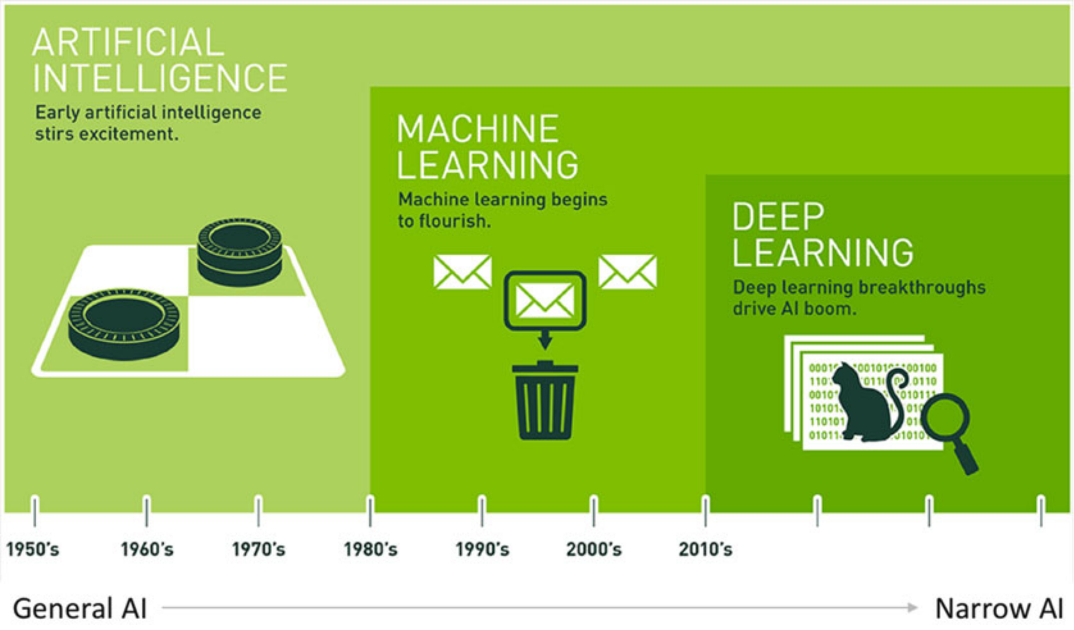

ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's
General AI
Narrow AI

Fig. 19.5 Evolution from general AI to narrow has been built on increasing computing power and modern AI techniques such as deep learning. Combine with high-quality structured data, narrow AI is beginning to produce high-quality results in medical imaging (Source: American College of Radiology Data Science Institute. Used with permission)

traditionally been divided into three categories: supervised learning, reinforced learning, and unsupervised learning [29]. In supervised learning, the goal of the machine learning algorithm is a known output, and the algorithm has been given data that has been labeled with a certain outcome or diagnosis. A widespread, familiar application of supervised learning in healthcare is the automated interpretation of an electrocardiogram to determine the presence or absence of a myocardial infarction. Examples of supervised learning in diagnostic imaging from recent machine learning competitions include lung nodule detection [30] and pediatric bone age determination [31], but the potential number of applications for these narrow AI models for segmentation, detection, quantification, classification, workflow improvements, and risk assessment is almost endless. Unlike supervised learning, reinforced learning models are not presented with a set of predetermined input/output pairs. In reinforced learning, the model determines the most effective pathway toward a goal by being rewarded for choosing different sets of actions. The system is rewarded when it achieves a certain outcome and then finds the path to the highest reward [29]. In unsupervised learning, machine learning models are given data that has not been labeled with a specific outcome, and there are no specific outputs to predict. Instead, the model separates input source data into naturally occurring groups or patterns based on the data. While both unsupervised learning and general AI will inevitably be applied to medical imaging using untagged or only loosely tagged training data, currently, unsupervised learning is best used for clustering, feature extraction or dimensionality, and variable reduction in the analysis of large datasets. One specific application of unsupervised learning in healthcare will be in advancing precision medicine initiatives focused on the various omics-based strategies including radiomics, genomics, proteomics, metabolomics, and epigenomics. These data patterns may be able to subdivide patients into prognostic categories and moreover may predict whether an individual patient would respond to various therapies. Particularly, these various omics-based strategies show promise within the domains of oncology and autoimmune conditions in predicting whether an individual patient would benefit or not from the various emerging targeted agents [29].

19.3.3 Use High-Quality Data for Training and Testing

In creating datasets for training AI algorithms, a robust source of accurate information, often referred to as "ground truth," is required for the training data. In supervised learning, the AI algorithms are trained on known cases. The source of this ground truth can come from a variety of sources but typically includes carefully annotated datasets done by expert radiologists and should be explicitly stated for each AI model. Other possibilities for establishing ground truth include pathology results or specific clinical outcomes [29]. While using high-quality data for algorithm training data is critical in order for algorithms to be effective, the datasets used for algorithm training data must also be diverse. Tremendous variability in the methods of diagnostic imaging such as equipment manufacturer, field strength in magnetic resonance imaging, imaging protocols, and radiation dose in computed tomography exists from institution to institution, and it cannot be assumed that AI models developed by training algorithms on data from a single institution will be effective more broadly. This problem is broadly characterized within the software development ecosystem as the problem of generalizability. Therefore, in bringing applications to market, the technical diversity of the training datasets must be considered. Additionally, patients are diverse as well, and the patient populations are likely to be quite different from institution to institution. In addition to general geographic diversity, patient populations from institution to institution may be variable due to race, gender, socioeconomic background, body habitus, and prevalence of disease processes. Recent reports indicate facial recognition algorithms demonstrate considerable variability in accuracy based on skin color and highlight potential sources of bias in algorithm development [32]. Both developers and consumers of AI applications in healthcare, and diagnostic imaging in particular, must be cognizant of the broad diversity in patient populations so that there will be similar diversity in training data so that algorithms will be free of unintended bias.

While there is critical need to provide high-quality, technically, and geographically diverse data to developers for testing and training, patient privacy must be carefully maintained. In the United States, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) [33] required the Secretary of the US Department of Health and Human Services (HHS) to develop regulations protecting the privacy and security of certain health information. The HIPAA Privacy Rule [34] defines standards and safeguards that protect patients' health records as well as personal health information that apply to all healthcare providers, insurers, and other healthcare entities. The rule sets limits and conditions on the uses and disclosures that may be made of such information without patient authorization. The HIPAA Security Rule [35] establishes national standards to protect individuals' electronic personal health information that is created, received, used, or maintained by a covered entity. The Security Rule requires "appropriate administrative, physical and technical safeguards" to protect the confidentiality, integrity, and security of electronic patient information. While the details of data privacy and other ethical considerations are beyond the scope of this chapter, it is clear that the ethical issues around data ownership, robustness of deidentification algorithms, and transparency in how patient information is shared with AI researchers and developers will play a crucial role in the development of a robust AI ecosystem [36, 37].

19.3.4 Develop Consistent Methods for Validation and Monitoring Algorithm Performance

While algorithms can be developed and used in single institutions without regulatory approval, in order to bringing new AI tools to widespread clinical use, developers will have to develop methods that ensure their products are generalizable to and reproducible in the wide variety of practice settings and patient populations that exist in the healthcare system. Inevitably some degree of governmental regulation for each algorithm will be necessary for AI to become broadly adopted. Algorithm validation standards must be developed that ensure that algorithms produce consistent results across the broad range of technical, geographic, and patient population diversity seen in clinical practice. Developers must be able to show that their algorithms can achieve the expected results on novel and diverse datasets, and there should be standardized statistical methods for comparing various algorithms that purport to have a similar function. Considering the thousands of algorithms that will likely be developed, governmental regulatory agencies could become overwhelmed further slowing the deployment of AI algorithms into clinical practice. An important role of the AI ecosystem for radiology will be to develop methods that support the validation of AI algorithms that can efficiently move AI products to market while ensuring patient safety. Establishing public-private partnerships between regulatory agencies and honest broker private groups, such as medical specialty societies, could play an important role in validation of AI algorithms.

However, regulatory approval of AI algorithms need not be entirely based on a premarket approval process. If methods can be developed that provide monitoring of the algorithms performance after deployment in community practice, these data can be used not only to ensure algorithms function as expected but also to provide information back to developers so that the algorithms can be improved. In radiology, these data should include not only information about the utility of the algorithm based on the radiologist input but also metadata about patient characteristics and technical parameters of the examination so that poor algorithm performance can be correlated with specifics about the examination.

19.3.5 Build Public-Private Partnerships for Safety and Efficacy

In the United States, the Food and Drug Administration (FDA) is charged with protecting the public health by "ensuring the safety, efficacy and security" of a wide range of healthcare products including medical devices [38]. As software has begun to play an increasingly important role in medical device technology, the US FDA's Center for Devices and Radiological Health has assumed a primary role in developing pathways for premarket review of medical device AI algorithms [39]. As a participant in the International Medical Device Regulators Forum (IMDRF)—a group of medical device regulators from around the world working to reach harmonization on medical device regulation—the US FDA has chaired IMDRF's Software as a Medical Device Working Group, which is developing guidance to support innovation and timely access to safe and effective "Software as a Medical Device" (SaMD) globally [40]. SaMD, defined as "software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device," has unique issues that make it worthy for consideration of its own regulatory approval processes [41, 42]. The FDA is working with the International Medical Device Regulators Forum [43] to ensure the US guidance on SaMD encompasses global issues around the regulation of software for medical purposes.

In the meantime, the US FDA has worked on several initiatives that will likely impact the regulation of AI products in the United States. In August 2017, the US FDA proposed the Medical Device Development Tools (MDDT) program [44], which is a pathway where the US FDA can qualify tools that medical device sponsors can use in the development and evaluation of medical devices. Qualification means that the FDA has evaluated the tool and has determined that the tool "produces scientifically-plausible measurements and works as intended within the specified context of use" [45]. FDA anticipates these tools, which can be developed by sponsors or private entities, will be useful in the approval process for AI algorithms and other SaMD.

Another US FDA program, the National Evaluation System for Health Technology (NEST) is intended to shorten the time to market for new technology healthcare products by developing a system for more robust post-market surveillance [46]. The US FDA NEST strives to generate better evidence for medical device evaluation more efficiently and enhance regulatory decision-making across the total product lifecycle of medical devices by "strategically and systematically leveraging real-world evidence and applying advanced analytics to data tailored to the unique data needs and innovation cycles of medical devices" [38]. Stated goals include moving medical devices to market more quickly, improving the ability to detect safety issues by moving to more active surveillance and to "efficiently harness data from the diverse set of real-world evidence—digital information collected from clinical experience in registries and similar tools—creating the necessary infrastructure for a national evaluation system for medical devices" [46, 47]. The US FDA believes that the NEST program, "by leveraging real world data collected as part of routine clinical care, our nation and the public will more fully realize the potential of the digital revolution for the device space" [46, 47].

The US FDA NEST program has established a number of demonstration projects to provide proof of concept for scalable approaches to generate safety and efficacy data across the entire medical device product life cycle using real-world evidence. These projects include methods to develop, verify, and operationalize methods of evidence generation and data use in the pre- and post-market space and demonstrate scalability across healthcare systems, device types, and manufacturers [47]. The NEST Coordinating Center (NESTcc) has chosen Lung-RADS Assist: Advanced Radiology Guidance, Reporting and Monitoring as one of their early demonstration projects for artificial intelligence algorithms. This project, sponsored by the American College of Radiology Data Science Institute (ACR DSI), is a method for validating and monitoring artificial intelligence algorithms built for detection and classification of lung nodules in lung cancer screening programs according to the ACR Lung-RADS classification system. The demonstration will use real-world data to assess the end-to-end workflow from deployment of an AI algorithm in a radiology reporting system through capture of performance metrics within a national registry [48]. This example of a public-private partnership may serve as a model for how AI algorithms can be monitored in clinical practice to ensure ongoing patient safely while establishing a pathway to increase the efficiency of the US FDA premarket review process.

Finally, the US FDA has also been working to develop and pilot the "Software Precertification Program" which focuses on the assessment of organizations that perform high-quality software design, testing, and monitoring based on demonstration of a "culture of quality and organizational excellence and a commitment to monitor ongoing performance" [49, 50]. The Software Precertification Program is envisioned to evaluate a developer's capability to respond to real-world performance and provide qualified developers with a more efficient premarket regulatory pathway for certain SaMD applications. SaMD developers would need to establish mechanisms for AI algorithm validation and post-market surveillance, and the program is expected to be synergistic with the US FDA MDDT and US FDA NEST programs.

While these US FDA programs are planned for the future, a number of solutions leveraging artificial intelligence algorithms have obtained premarket US FDA approval using the current US FDA processes. The US FDA classifies and regulates medical devices based on the degree of risk to the public with the least risky Class I devices subject to the lowest level of regulatory controls and Class III devices subject to the highest level of regulatory controls. Class I devices include simple medical supplies such as gloves. Class II devices include CT scanners and other radiological equipment, and Class III devices include intravascular balloon catheters and stents [51]. Class I devices and certain Class II medical devices do not require formal premarket notification or 510(k), but the vast majority of Class II devices require premarket notification, also called 510(k). The 510(k) clearance process is the path to market for the majority of medical devices but requires that the device be substantially equivalent to a legally marketed device termed a "predicate" by the US FDA. Class III devices require a more robust premarket approval process than a 510(k) clearance. This approval process typically requires the sponsor submits clinical data showing reasonable safety and efficacy of the medical device [51]. Medical devices with no legally marketed substantially equivalent predicate would be automatically classified as Class III regardless of risk; however, the US FDA has recently revamped the de novo request process that allows a developer of a low-to-moderate risk device without a predicate to submit a request to the US FDA to make a risk-based classification of the device into Class I or II, without first submitting a 510(k) and receiving a not substantially equivalent (NSE) determination. Once a device is cleared under the de novo process, this device may then serve as a predicate for 510(k) premarket approval of similar devices in the future [52]. A number of US FDA approvals for artificial intelligence software have been granted based on this de novo process [53–55].

The US FDA also classifies computer software intended for lesion detection and diagnosis. The computer-aided detection (CADe) is defined as "computerized systems that incorporate pattern recognition and data analysis capabilities intended to identify, mark, highlight or in any other manner direct attention to portions of an image, or aspects of radiology device data, that may reveal abnormalities during interpretation of patient radiology images or patient radiology device data by the intended user" [56, 57]. Computer-aided diagnosis (CADx) is defined by the FDA as "computerized systems intended to provide information beyond identifying, marking, highlighting or in any other manner directing attention to portions of an image, or aspects of radiology device data, that may reveal abnormalities during interpretation of patient radiology images or patient radiology device data by the clinician." Both CADe and CADx utilize highly complex algorithms. A primary distinction between CADe and CADx is that CADe is intended as merely an adjunct detection tool for the radiologist who, per device labeling, is expected to fully review the images and not rely on the software. Although initially regulated as Class III, more recently FDA has approved CADx under its 510(k) process. Because of the relatively higher risk associated with CADx, FDA has previously been slower to move CADx toward the 510(k) process.

However, on July 19, 2017, the US FDA granted developer QuantX de novo approval and Class II status to computer-aided diagnosis software (CADx) for breast cancer detection [58]. This appears to represent a relaxation of the US FDA's premarket approval process requirements. CADx software and may become the basis for clearance for some artificial intelligence applications.

Although these US regulatory programs seem somewhat disjointed, in all of its activities, the US FDA seems to be working to streamline the review process for artificial intelligence applications in healthcare, and they are demonstrating a high level of cooperation with international regulatory bodies. However, even with the streamlined premarket processes described above, developers will still need to demonstrate efficacy, patient safety, and a process for post-market surveillance of ongoing effectiveness using real-world data. Regulatory agencies are ill-equipped to perform these tasks internally. Additionally, the sheer number of algorithms that will likely be submitted for regulatory approval could place considerable burdens on the regulatory reviews process in the United States and internationally as well. Public-private partnerships between regulatory agencies and trusted organizations such as medical specialty societies can facilitate both the premarket review and the collection of real-world evidence that support ongoing efficacy and safety of AI algorithms in clinical practice.

19.3.6 Establish Standards for Interoperability and Pathways for Integration into Clinical Workflows

In concert with the Quadruple Aim for increasing value in healthcare [9, 10], the American College of Radiology's Imaging 3.0 initiative [59–61] is a call to action for how radiologists can play a leadership role in shaping the future of healthcare by systematically providing value to patients and the healthcare system beyond image interpretation beginning when diagnostic imaging is first considered until the referring physician and ultimately the patient fully understand the imaging results and recommendations. This imaging cycle has been described as the imaging value chain and describes how radiologists can impact appropriateness, quality, safety, efficiency, and patient satisfaction at each step in the cycle (Fig. 19.6), and at each step, there are software tools available to radiologists to help them provide higher-value care [59].

Imaging 3.0 identifies specific ways radiologists can enhance the value they provide to patients through a number of initiatives: imaging appropriateness, quality, safety, efficiency, and satisfaction [59–61]. For this effort to succeed, radiologists must have the necessary informatics tools available to them at the point of care throughout the imaging value chain. These tools, depicted in Fig. 19.7, have been successfully implemented in clinical workflows in the United States and worldwide, and the success of the Imaging 3.0 initiative has been dependent on using informatics tools that work in concert with the imaging interpretation workflow.

Imaging 3.0 informatics tools promote appropriate use of imaging services, the use of structured reporting so that critical data can be easily extracted from imaging reports, clinical decision support for radiologist interpretation, image sharing solutions to provide access to patient electronic access images within the enterprise and across sites, and communication enhancements

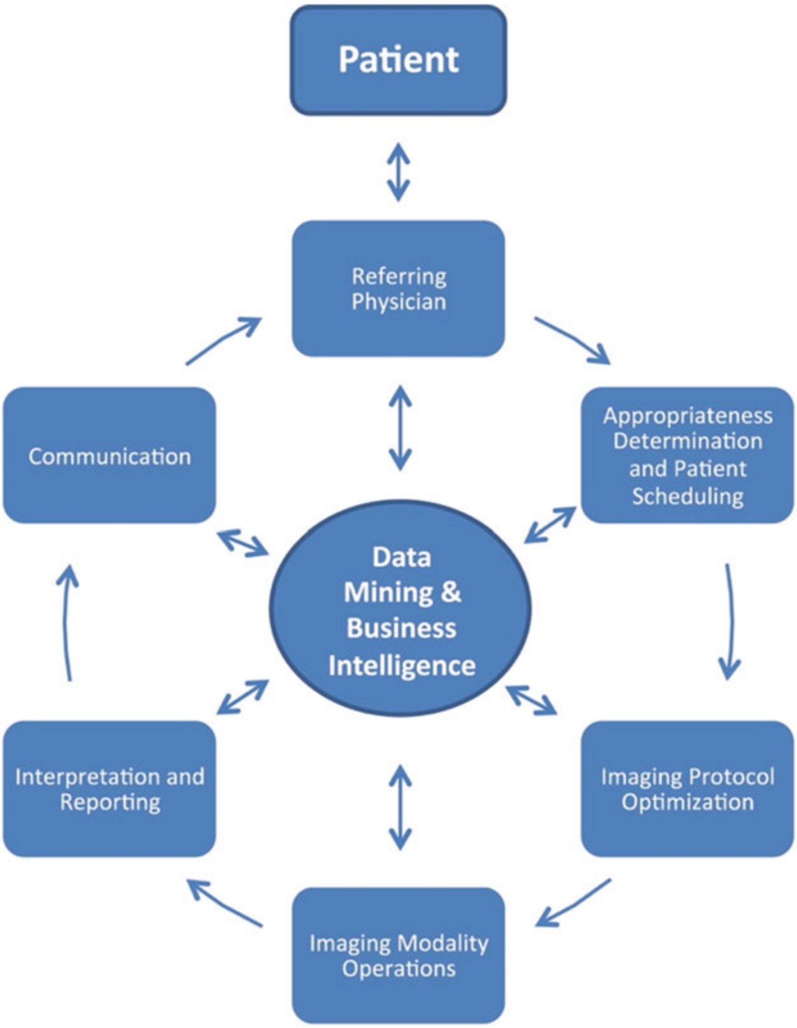

Patient
Referring Physician
Appropriateness Determination and Patient Scheduling
Imaging Protocol Optimization
Imaging Modality Operations
Interpretation and Reporting
Communication
Data Mining & Business Intelligence

Fig. 19.6 The imaging value chain. Source: Ref. [59]. Boland GW, Duszak R, McGinty G, Allen B. Delivery of appropriateness, quality, safety, efficiency and patient satisfaction. Journal of the American College of Radiology. 2014 Jan 1;11(1):7–11. Used with permission

using registry reporting to benchmark patient radiation exposure, patient outcomes, and quality assessment. Artificial intelligence algorithms are poised to become radiology professionals' next important Imaging 3.0 informatics tool and will continue to increase radiologists' value to patients and their health systems.

Just as with the informatics tools for Imaging 3.0, in order for radiologists to effectively use artificial intelligence algorithms in routine clinical practice, developers must pay careful attention to how algorithms will capture data for analysis and how output from the algorithms will integrate back into the clinical workflow. Seamless interoperability with the healthcare systems' numerous electronic resources will be necessary for optimal clinical integration. Inputs for the algorithm may come from data from the imaging modalities, the picture archiving and communication systems (PACS), the electronic health record (EHR), and an array of data sources including pathological information, radiology information systems, patient monitoring systems, or even wearable health monitoring devices. Standard interfaces must be developed so that similar algorithms can import this information in the same way, and proprietary solutions must be avoided. For instance, it is inevitable that in robust clinical use, radiology departments will be using innumerable algorithms for a wide

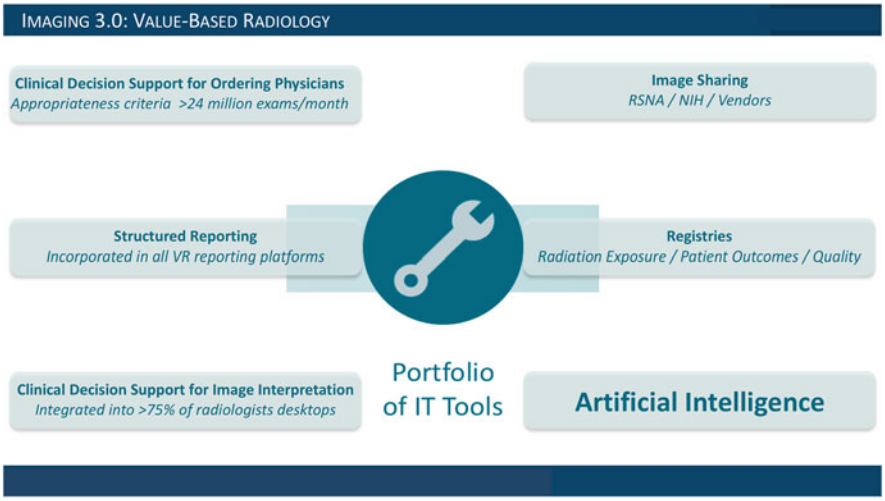

IMAGING 3.0: VALUE-BASED RADIOLOGY
Clinical Decision Support for Ordering Physicians
Appropriateness criteria >24 million exams/month
Image Sharing
RSNA / NIH / Vendors
Structured Reporting
Incorporated in all VR reporting platforms
Registries
Radiation Exposure / Patient Outcomes / Quality
Clinical Decision Support for Image Interpretation
Integrated into >75% of radiologists desktops
Portfolio of IT Tools
Artificial Intelligence

Fig. 19.7 Imaging 3.0 informatics toolkit (Credit authors)

variety of clinical applications. Some may run on premises, and in those instances, it would be much more efficient for algorithms with similar hardware requirements to run on the same on-premises server and acquire input data using the same interfaces. Cloud-based solutions, even if developer specific, will also benefit from standardized input interfaces, and the developer community should work in concert with physicians and the health information technology (HIT) industry to set interoperability standards for these interfaces. By developing standardized methods for communications between platforms, different vendors can focus various different tool development areas within an infrastructure that allows them all to connect together ultimately giving the physicians and other end users access to a wider array of solutions without concern for compatibility. For instance, the Logical Observation Identifiers Names and Codes (LOINC), developed by the Regenstfrief Institute, in the mid-1990s, is a universal standard endorsed by the American Clinical Laboratory Association and the College of American College of Pathologists. It also contains a database of standard terms and has been expanded to include nursing diagnoses, interventions, and outcome classifications [62].

Equally important will be standardization of output interfaces. Radiologists, referring physicians, and other providers use an array of electronic resources throughout the imaging cycle. Output from AI algorithms will eventually interface with existing clinical decision support tools for selecting the most appropriate radiological examination as well as existing decision support tools for radiologist interpretation. Standardized interfaces for algorithm output into PACS worklists and at the modality will be necessary as well, and for optimal workflow integration, artificial intelligence algorithms will have to seamlessly interface with all of these resources. Developing open sources for coding and standardized interfaces for data transfer will ultimately affect the entire health information technology ecosystem, and developers of AI applications must avoid proprietary interfaces. An example of an open-source interface for bringing evidence-based guidelines to the point of care is the American College of Radiology's Computer Assisted Reporting Data Science (CARDS) platform [63]. The CARDS authoring and reporting system includes a definition format for representing radiology clinical guidelines as structured, machine-readable Extensible Markup Language (XML) documents with a user-friendly reference implementation to test the computer language with the clinical guideline. The CARDS output has open-source standards for delivering the CARDS output to voice recognition software (VRS) platforms.

There will be numerous other electronic resources in healthcare that developers must consider, and interoperable data transfer standards are critical. Communications with PACS and imaging modalities must include interfaces with the Digital Imaging and Communications in Medicine (DICOM) which is the standard for storing and transmitting medical images. The DICOM communication standards, developed through a collaboration between the American College of Radiology and National Electrical Manufacturers Association (NEMA), facilitate the integration of medical imaging devices such as scanners, servers, workstations, printers, network hardware, and PACS from multiple manufacturers [64]. However, this standard is generally limited to use within radiology, and as AI evolves, other mechanisms for data transfer must be considered to allow input of patient information from sources outside radiology. Additionally, as AI evolves in other specialties in medicine such as pathology, ophthalmology, and dermatology, expanding image digitalization and transfer standards to other areas in the healthcare system will be necessary so that outputs from AI algorithms can interface with these resources as well.

AI algorithms will also be expected to interface with electronic health records and other primarily text-based systems. Data transfer in these systems is predominantly via Health Level Seven (HL7) protocols which are designed to facilitate "the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services" [65]. More recently, the Fast Healthcare Interoperability Resource (FHIR) [66] is showing tremendous promise for joining disparate systems together. FHIR's resource-based modular components allow the development of an application-based approach to interoperability and health information exchange. FHIR supports interoperability over a variety of architectures including representational state transfer (REST), messaging, documents, and services. FHIR has the ability to be used over a variety of platforms including cloud communications, EHR data sharing, radiology information systems (RIS), server communications, and mobile platforms, among others [57]. Artificial intelligence interfaces will need to be cognizant of these communication platforms to optimize input from and output to the patients' health records outside of the radiology department.

Another requirement for interoperability and clinical integration of artificial intelligence algorithms will be the development of pathways to increase the production of more structured data in our health systems in general and more specifically in radiological reporting. Narrative radiological reports, designed for human consumption, contain a wealth of information that, if extractable by automated systems, will be invaluable not only for the clinical care of that specific patient but also useful for clinical quality improvement activities, population health management, and research [67]. The creation of common data elements, which define the attributes and allowable values of a unit of information, are "data elements that are collected and stored uniformly across institutions and studies and are defined in a data dictionary" [67, 68]. CDEs allow machine-readable representation of imaging findings including anatomic location and dimensions and can store computational features including density and textural metrics. CDEs allow reports to be built from tiny collections of information that contain not only words but also context, meaning, and relationships [67, 68]. In order to optimize standardization and interoperability of artificial intelligence applications in radiology, use case definitions need to use standardized definitions (CDEs) for algorithm input and output, not only to be interoperable with other electronic resources, but also to ensure that algorithms built around similar clinical applications have standardized inputs and outputs, so they can be compared and integrated into clinical workflows in a similar manner. Standardized use cases are critical to interoperability and integration of AI into clinical practice, and these use cases can only be developed using CDEs.

The radiology community has long recognized the need for developing CDEs for radiology reporting and data aggregation [67]. CDEs are important for populating data registries such as the American College of Radiology Dose Index Registry [69] and other data registries [70] and in the development of radiological lexicons and structured reporting templates as well as in the integration of outputs from decision support tools using appropriate use criteria such as ACR Select [71–73] or evidence-based guidelines such as CARDS [63, 74]. A similar effort in the radiology community will be necessary to create the CDEs necessary for the implementation of AI in clinical practice. The Radiological Society of North America and the American College of Radiology are collaborating to advance the CDEs for radiology. RadElements for radiology [75] are standardized names and attributes of data elements to support clinical and translational research, patient care, and performance improvement in diagnostic and interventional radiology. These will be the key data elements to support standardized pathways for algorithm inputs and outputs as well interactions with data registries (Fig. 19.8).

19.3.7 Promote Explicability of Algorithm Output

In order to gain physician acceptance of algorithm output, developers must be able to ensure that their work is reproducible and that algorithm output can be verified by physicians as they incorporate the results into clinical care [76]. Recently adopted American Medical Association policy states that developers have a responsibility to ensure their work is transparent and can be reproduced by others [77]. Some guidelines suggest that developers publish reports of their predictive machine learning algorithms that include not only rationale and objectives but also the setting, prediction problem, relevant data, and a description of the building of the predictive model and its limitations [78, 79]. Some have gone as far as to recommend creating open repositories for long-term storage, archiving, and public access to datasets and code to enable replication of published findings [79]. Whether these or similar requirements are ultimately adopted remains to be seen, but from a practical perspective, radiologists will need to have explicability of algorithm results, typically in the form of saliency maps, that demonstrate the finding identified in AI inference models. For instance, an algorithm identifying cardiomegaly from chest radiographs should not rely on the presence or absence of sternal wires to make that determination, and algorithms trained to characterize pulmonary nodules into the Lung-RADS classification scheme should not just provide a Lung-RADS classification but also locate, identify, and report the characteristics of the nodule that led to that particular inference [80] (Fig. 19.9).

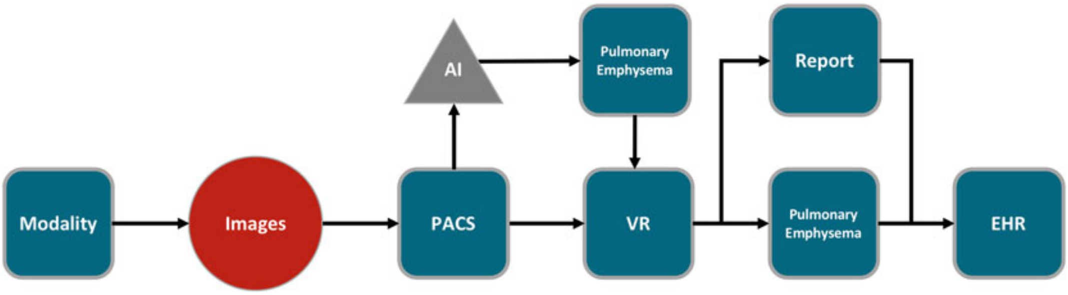


Fig. 19.8 AI has the potential to bring quantitative information about underlying disease, often incidentally detected, directly to the top level of the EHR. For instance, a patient where significant emphysema is detected in a surveillance CT examination after chest trauma could have that information transmitted directly to and recorded in the problem list of EHR. Full integration and interoperability of electronic resources including PACS and voice recognition reporting software (VR) are required (Credit authors)

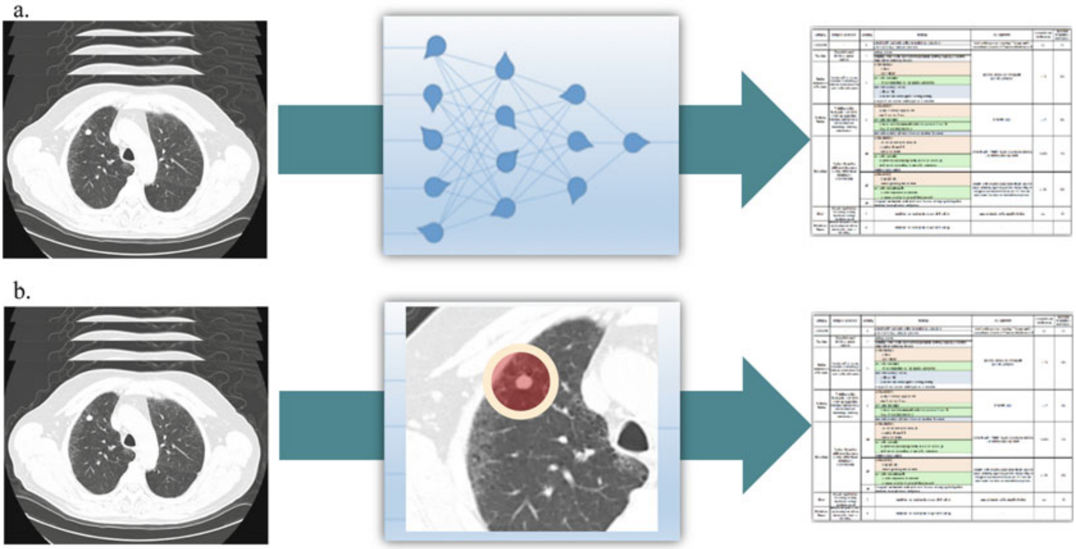


Fig. 19.9 Although an AI algorithm could provide a direct answer regarding assignment of a lung nodule in the Lung-RADS classification, physicians will want to "see" the abnormality in order to assess the reliability of the algorithm's output. In evaluating patients in a lung cancer screening program, an AI algorithm might be expected to be able to directly classify patient as having a Lung-RADS 4 lung nodule (**a**); however, without providing information to the radiologist about location, size, and imaging characteristics of the nodule, the radiologist will not be able to verify the result for clinical use (**b**) (Credit authors)

Use cases for pediatric bone age determination have typically specified that the algorithm output displays a radiographic of the result of the inference model along with radiographs of bone ages 6 months on either side of the inference so that the radiologist is able to choose the standard in best agreement with the patient's radiographs [81]. While the ability to provide saliency maps will be possible when the algorithm inference is the detection of radiographic findings, explicability determination for many other AI applications will need to be established as the specific use case is developed.

19.3.8 Facilitate Radiologist Input into Development, Validation, and Implementation

Development of a robust AI ecosystem where there is widespread adoption and clinical implementation of artificial intelligence is dependent on active radiologist involvement in the development, validation, and implementation of AI algorithms. Creation of AI tools at single institutions does not ensure that the validity of the algorithm will be the same in widespread clinical use. Furthermore, specific needs of various institutions and implementation pathways could be significantly different from one institution to another. In order to ensure the development of AI tools is generalizable to widespread clinical use, there should be general agreement among the end users, that is, radiologists and their health systems, regarding selecting use cases for AI that impact a significant clinical or administrative need and can be seamlessly integrated into the workflow. Radiologists should work collectively to define these important use cases for AI. Additionally, radiologists will be necessary to develop datasets for training and testing of algorithms, and standards should be developed to help them create datasets for algorithm training and testing that are accurately annotated and well curated so that developers will have robust and diverse data sources for training and testing. Radiologists will also play a significant role in ensuring algorithms are effective and safe for clinical use. Datasets used for algorithm validation prior to general clinical deployment should have higher standards for ground truth than the datasets used for algorithm training, and radiologists should play a significant role in creating the standards for algorithm validation including not only in ensuring validation datasets are as close to ground truth as possible but also in defining the metrics used for algorithm validation so that similar algorithms can be compared to one another in a similar fashion. Finally, radiologists will be the best source of ensuring the safety and efficacy of AI algorithms in clinical practice. Mechanisms for capturing input about algorithm performance from radiologists should be built into the clinical workflow, and radiologists must recognize the importance of their role in assessing the performance of the algorithm in clinical practice. Collaborations between individual radiology professionals, their medical specialty societies, and the developer community will be necessary to facilitate the advancement and clinical use of artificial intelligence in clinical practice.

19.4 Bringing Artificial Products to Widespread Clinical Use: Challenges, Opportunities for Radiologists, and the Role of Medical Specialty Societies

Artificial intelligence challenges such as those hosted by ImageNet and now Kaggle are showing that for certain narrow AI use cases machine learning can achieve results for detection and characterization of radiological imaging findings on par with humans [82]. A growing number of studies have also shown that AI can accurately diagnose certain medical conditions as well as physicians [83–85]. Additionally, the last 2 years have seen a growing number of artificial intelligence tools cleared for marketing in the United States [86]. A new emphasis at the US FDA to decrease the time for premarket review for medical software and artificial intelligence products includes de novo classification of some medical image analyzers such as computer-assisted detection mammography software (CADe) as US FDA Class II devices, which significantly lowers the premarket requirements for developers [87]. The number of publications on machine learning in healthcare is increasing steadily with medical imaging the most popular topic [28] (Fig. 19.10).

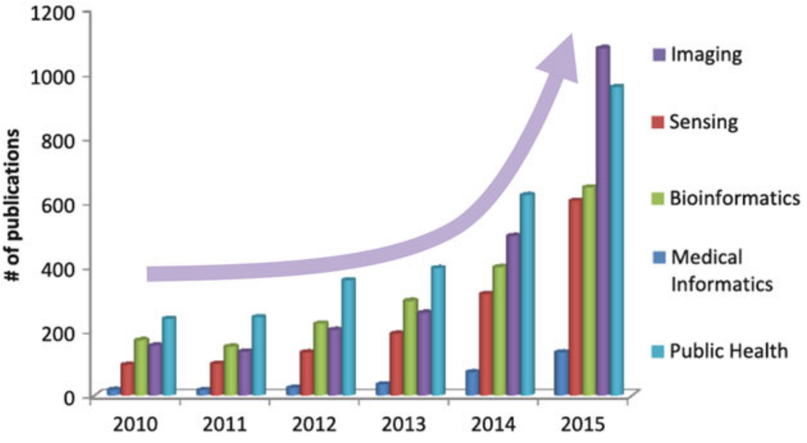


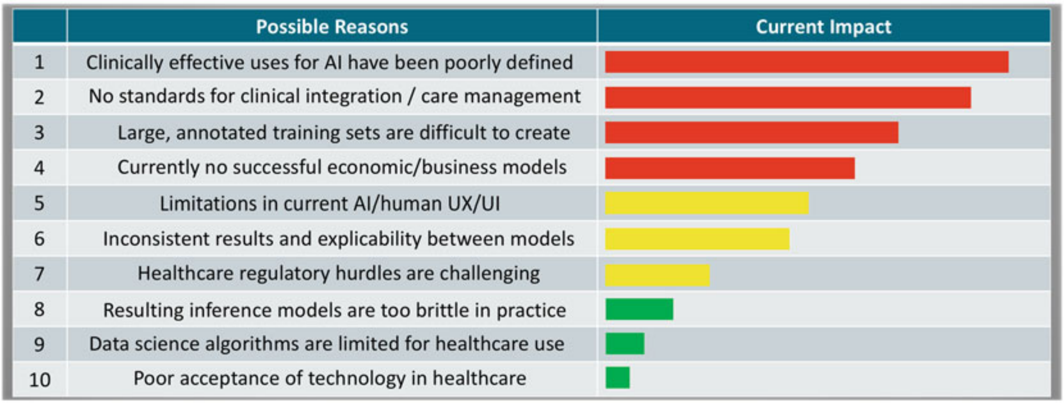
Fig. 19.10 Source: Ref. [28]. Deep Learning for Health Informatics—Daniele Ravi et al., IEEE Journal of Biomedical and Health Informatics, Vol. 21, No. 1, January 2017 (Used with permission)

The potential market for AI solutions in healthcare has attracted a growing number of developers in the United States and worldwide, and diagnostic imaging has been a major focus of their attention [88]. Despite these potential incentives from the FDA and extensive marketing from developers [89], the current penetrance of AI algorithms in clinical practice is quite modest and generally limited to implementation of algorithms developed at single institutions and integrated into their specific workflows [90]. Understanding the challenges facing developers as they move their AI solutions from concept to implementation provides opportunities for radiology professionals, radiology specialty societies, and the entire healthcare community to facilitate an ecosystem that will allow rapid deployment of clinically useful, effective, and safe algorithms into routine clinical practice. As shown in Fig. 19.11, the challenges being faced by developers include a lack of general availability of large, diverse, and deeply annotated data for training and testing AI algorithms, inconsistent results and explicability between AI models, ensuring AI algorithms are valid across a wide variety of practice settings, navigating the US FDA process and other regulatory hurdles for marketing a new AI application, a lack of defined standards for clinical integration, interoperability, a lack of mechanisms for using real-world data to monitor AI effectiveness in clinical practice, and most significantly a lack of well-defined clinically relevant use cases for AI in healthcare that are able to address dataset annotation and curation, algorithm validation, integration into clinical practice, and monitoring of clinical effectiveness [1].

Radiology professionals can help mitigate these challenges by playing a leading role in the use case development process, and radiology professionals' medical specialty societies can serve as a convener, coordinator, and honest broker to facilitate the process.

19.4.1 Creating Clinically Effective Artificial Intelligence Use Cases

A software "use case" is much more than just an idea for what a software application, including artificial intelligence algorithms, should do. In software development terms, a use case is a prose description of a computing system's behavior when interacting with the outside world. First proposed by Jacobson in 1987 [91], a use case defines a list of actions or events between the end users (actors) and the computing system and describes the system's behavior under various conditions as the system responds to requests from the primary actors [92]. The actors may be humans, or in the case of healthcare, the electronic resources used by the healthcare team



	Possible Reasons	Current Impact
1	Clinically effective uses for AI have been poorly defined	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥
2	No standards for clinical integration / care management	🟥🟥🟥🟥🟥🟥🟥🟥🟥🟥
3	Large, annotated training sets are difficult to create	🟥🟥🟥🟥🟥🟥🟥🟥
4	Currently no successful economic/business models	🟥🟥🟥🟥🟥🟥🟥
5	Limitations in current AI/human UX/UI	🟨🟨🟨🟨🟨🟨
6	Inconsistent results and explicability between models	🟨🟨🟨🟨🟨
7	Healthcare regulatory hurdles are challenging	🟨🟨🟨
8	Resulting inference models are too brittle in practice	🟩🟩
9	Data science algorithms are limited for healthcare use	🟩
10	Poor acceptance of technology in healthcare	🟩

Fig. 19.11 Challenges to getting AI tools into clinical practice (Credit authors)

in daily interactions. For AI development purposes, an artificial intelligence use case defines exactly how an AI algorithm takes in information (images, EHR, genetic, structured data, or unstructured data) from the clinical workflow and then provides a specific output (detection, classification, quantification, prioritization, etc.) to the end user within the clinical workflow [1]. To help move AI algorithms into clinical practice, AI use cases can also include parameters for how the algorithms are trained, tested, and validated for regulatory approval and clinical use, how they are deployed into clinical workflows, and how their effectiveness can be monitored in clinical practice.

Use case creation is an opportunity for radiologists to play a leading role in assisting developers create algorithms that will be useful, effective, and safe in clinical practice and enhance the value radiology professionals provide to their patients and health systems. Radiology subspecialty societies are uniquely positioned to convene multiple stakeholders, ensure patient safety, promote diversity in algorithm development, and collaborate with regulatory agencies to facilitate the introduction of AI algorithms into clinical practice. A result of single institution development of AI algorithms is that in the aggregate, specific use cases for artificial intelligence (AI) in diagnostic radiology are broadly and inconsistently defined with variation in how algorithms will be developed, validated, adopted, and monitored in clinical practice. There has been little validation of algorithms across more than a few sites, and whether the effectiveness of these algorithms will be generalizable to widespread clinical practice and how they will be integrated into clinical workflows across a variety of practice settings remains uncertain. The American College of Radiology's Data Science Institute has developed a standardized process for AI use case development to help achieve the goal of widespread use of clinically relevant, safe, and effective AI algorithms in routine radiological practice [93]. Technology-Oriented Use Cases in Healthcare AI (TOUCH-AI) is an open framework authoring system for defining clinical and operational AI use cases for the radiological sciences that intersect high clinical value with problems solvable by AI. TOUCH-AI provides a framework that includes narrative descriptions and flowcharts that specify the goals the algorithm should meet, the required clinical inputs, how it should integrate into the clinical workflow, and how it should interface with both human end users and an array of electronic resources, such as reporting software, PACS, and electronic health records. Combined with the ACR's existing open framework for authoring and implementing computer-assisted reporting tools in clinical workflows, CARDS (Computer Assisted Reporting Data Science) and TOUCH-AI provide an end-to-end AI use case authoring platform for the development of ACR DSI use cases for the AI developer community.

Using the guidelines and open specifications in authoring tools such as TOUCH-AI and CARDS, AI use cases can be developed in an environment that creates uniform data elements that allow standardization of data elements for creation and annotation of datasets for algorithm testing and training, data elements and statistical metrics for algorithm validation, application programming interfaces (APIs) for algorithm deployment into clinical and departmental workflows, and data elements for monitoring the algorithm's performance in widespread clinical practice. This process helps ensure patient safety by creating use cases that have data elements for algorithm validation and regulatory review and for monitoring real-world performance of the algorithm after deployment in routine clinical practice. This process also ensures AI use cases have data elements for effective clinical integration using workflow tools such as reporting software, the modalities, PACS, and EHR. The TOUCH-AI development platform takes advantage of the array of common data elements being created under the joint ACR RSNA RadElements project to optimize clinical interoperability and implementation by ensuring standardization of input and output elements from the algorithms [75]. While ACR DSI use cases begin as narratives and flowcharts describing the use case, this human language is then converted to machine-readable format (Extensible Markup Language—XML).

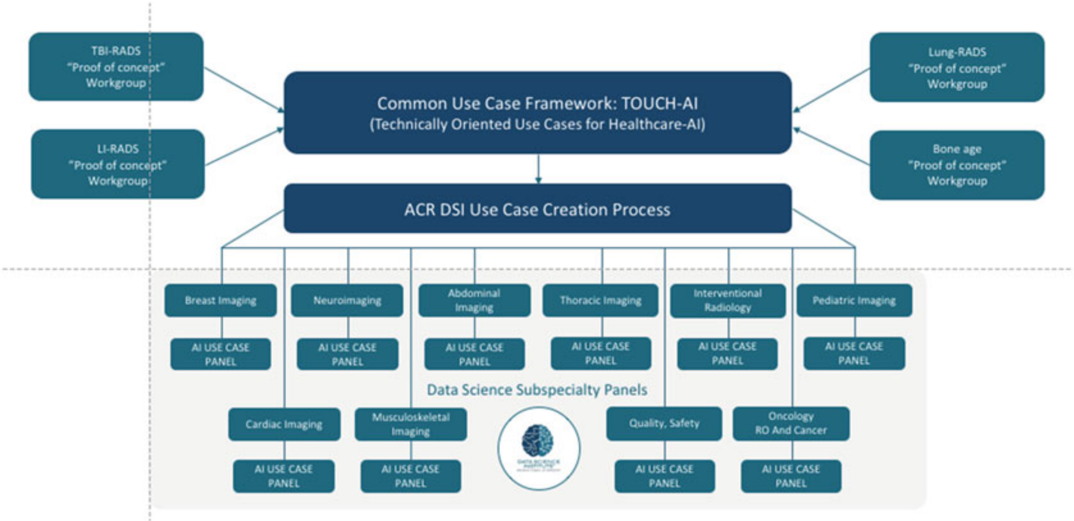


Fig. 19.12 The ACR DSI Data Science Subspecialty Panels are composed of radiologists in the various subspecialties of radiology and are tasked with developing AI use cases that will find the intersection of problems in radiology and those problems that are potentially solvable by AI (Credit authors)

To facilitate the involvement of radiology professionals in the AI use case development process, the ACR DSI has established ten Data Science Subspecialty Panels, composed of clinical experts, many with data science backgrounds, to evaluate and choose the highest value use case proposals for development. These panels include all of the subspecialty areas of diagnostic radiology, a panel for oncology and radiation oncology and a panel for non-interpretive AI use cases. Additionally, ACR DSI Data Science Workgroups are developing proof of concept use cases to be used in concert with AI developers, the HIT industry, and US FDA regulators to demonstrate the ACR DSI use case development concepts (Fig. 19.12).

While many of the ACR DSI AI use cases will be developed by the panel members, crowdsourcing in AI development, particularly in the form of AI competitions, has been a key to rapid dissemination of knowledge and technical information [2]. These concepts should be applied to use case development as well. Radiologists can collaborate through specialty societies to develop larger pools of thought regarding the highest priority for use cases for the radiological sciences. Additionally, individual developers and institutions can take use cases they are working on and have them encoded with data elements specifying broader standardized annotation of training sets, validation, integration, and monitoring in clinical practice [75].

Crowdsourcing has been a helpful tool for engaging the developer community around the development of AI algorithms, and Kaggle has hosted a number of competitions related to healthcare and medical imaging [94–97]. These competitions have engaged thousands of researchers and developers to focus their attention on healthcare use cases; however, participants in many instances are not healthcare or diagnostic imaging experts, which creates a lack of information about how physicians and other stakeholders will interact with an AI algorithm. Additionally, the sponsors have often created use cases that are generally unstructured with broad rather than specific goals for outputs that can be integrated into clinical practice. For instance, the 2017 Data Science Bowl sponsored by Kaggle and the Memorial Sloan Kettering Cancer Center used a public dataset from the US National Institute of Health and asked participants to "develop algorithms that accurately determine when lesions in the lungs are cancerous" in order to "dramatically reduce the false positive rate that plagues the current detection technology" [94]. While the algorithms developed for this competition were impressive from a data science perspective, the clinical utility of these algorithms is difficult to determine. There was no structured mechanism for detection, localization, and characterization of the lesions defined in the use case, and as a result the output of the algorithms was variable. Most of the algorithms reported a percent cancer risk for an individual nodule was reported, but the information was in many ways not useful in routine clinical practice. For instance, while the risk of cancer in a nodule could be classified as 95% or 15%, the ultimate medical treatment for both nodules is still tissue sampling [98]. A better output for the algorithm might have been to assign a Lung-RADS score [99] along with the additional features radiologists would use in reporting lung cancer screening examinations such as lesion location, lesion size, and lesion characteristics such as solid or subsolid and smooth, lobulated, or spiculated. For these reasons, AI use cases developed by the end users in concert with an understanding of available guidelines and electronic resources for clinical integration are likely to gain more widespread clinical adoption than those developed from more broadly based unstructured use cases. The American College of Radiology (ACR) and the Medical Image Computing and Computer Assistance Intervention (MICCAI) Society recently announced that MICCAI will be using ACR DSI use cases in the MICCAI imaging AI competitions in order to foster the development artificial intelligence algorithms that will better meet the clinical needs of radiologists [100].

There are nearly endless opportunities for use case development for the radiological sciences (Fig. 19.2). Examples of clinically useful AI use cases based on image analysis include detection of critical findings in order to prioritize radiologist work lists, classification of detected abnormalities based on clinically accepted evidence-based guidelines, pre-analysis of images to mitigate observer fatigue, and extracting information from images that is not visually apparent [12]. Non-interpretive use cases will be useful in enhancing image quality, optimizing radiation exposure, improving departmental workflows, and enhancing patient experience [93]. As in the case of the data science competitions where crowd-sourcing has typically resulted in publicly and freely available code for advancing the field, standardized structured use cases for developing AI algorithms should also be made publicly available at no cost. The ACR DSI intends to make all of its structured use cases freely available to the developer community. Creation of structured use cases can be a key component of the AI ecosystem. By helping mitigate some of the challenges facing developers such as aggregation of training datasets, streamlining the validation process, specifying pathways of clinical integration, and providing mechanisms for monitoring in clinical practice, development of structured use cases has the potential to accelerate the process of moving high-quality AI algorithms into clinical practice (Fig. 19.13).

Radiology specialty societies such as the American College of Radiology are uniquely positioned to facilitate the development of an AI ecosystem that convenes multiple stakeholders, ensures patient safety, promotes diversity in algorithm development, and collaborates with federal regulatory agencies and even the Congress to facilitate the introduction of AI algorithms into the market that will enhance the care radiology professionals provide for their patients [1].

19.4.2 Enhancing the Availability of High-Quality Datasets for Algorithm Testing and Training

Use cases that standardize definitions of data elements, tools, and instructions for annotating these datasets will enable a common framework for multiple institutions and developers to use for algorithm training and testing. Using multiple sites as data sources for these datasets provides technical, geographic, and patient diversity to prevent unintended bias in algorithm development and

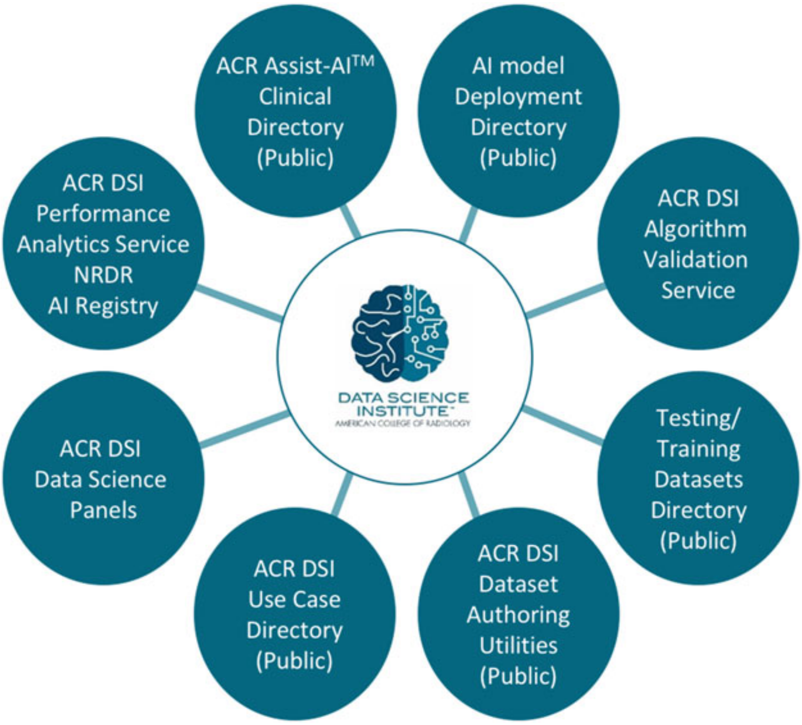

ACR Assist-AI™ Clinical Directory (Public)
AI model Deployment Directory (Public)
ACR DSI Performance Analytics Service NRDR AI Registry
ACR DSI Algorithm Validation Service
DATA SCIENCE INSTITUTE™
AMERICAN COLLEGE OF RADIOLOGY
ACR DSI Data Science Panels
Testing/ Training Datasets Directory (Public)
ACR DSI Use Case Directory (Public)
ACR DSI Dataset Authoring Utilities (Public)

Fig. 19.13 (Ref. [93]). The ACR Data Science Institute has proposed a process where AI use cases developed by clinical experts and translated to machine-readable language can include data elements for training and testing AI algorithms, metrics for algorithm validation, specifications for clinical integration, and mechanism for monitoring algorithm in performance clinical practice (Credit authors)

allows more individual radiology professionals and institutions to participate in the AI development process. Public directories of institutions that have created these datasets around structured use cases can inform the developer community about sites that have training datasets available. Compared to unsupervised learning or the use of only loosely annotated datasets for algorithm training, the cost of creating well-curated, deeply annotated datasets will be high. Expert analysts, including radiologists, and methods to analyze health records for pathology data will be needed to create high-quality datasets, and this process will be costly [2]. However, if the datasets created around multiple use cases are widely available from multiple developers, the aggregate cost of training and testing AI algorithms could be substantially reduced. The current practice and associated costs of developers obtaining data from single institutions have led some developers to require practices and institutions providing data to developers to sign noncompete agreements. If developers are expected to work together, then the AI ecosystem will need support mechanisms to protect intellectual property while fostering the sharing of annotated datasets and tools.

Another challenge to be addressed will be the integration of multiple healthcare datasets that will be complex, heterogeneous, and inconsistently structured. An aspirational goal is to amass large datasets to facilitate novel disease correlations in order to match patients to the best treatments based on their specific health, life experiences, and genetic profile [2]. AI holds the promise of integrating all of these data sources with imaging data to promote population health management. However, the availability of high-quality data and the ability of AI algorithms to integrate between a narrow AI use case for image recognition and a more general AI use case interacting with unstructured data from non-imaging data sources have to be considered.

A collaborative approach between institutions with annotated datasets built according to specific AI use cases and AI developers working on algorithms around those use cases can be enhanced by involvement of honest-broker third parties such as medical specialty societies who can house directories of institutions with available datasets. This could become a key function of the radiology AI ecosystem to facilitate the advancement of AI tools to clinical practice.

19.4.3 Maintaining Patient Data Privacy in Developing and Validating Artificial Intelligence Algorithms

Both healthcare culture and public law require physicians to closely protect patients' health data, but the development of large patient datasets incorporating wide ranges of radiologic, clinical, and pathologic information across multiple institutions for the development of AI algorithms will necessitate a thorough re-examination of issues surrounding patient privacy, confidentiality, and informed consent. The same tools that are anticipated to be useful in the characterization of a patient's disease may eventually extract information in a manner that makes any image identifiable to a specific patient, similar to a fingerprint. The integration of patient data from multiple sites and sources in the development of AI use cases likely enhances the risk of large-scale leakage of protected information. Routine disclosure of patient information care, at least within a given institution, is widely accepted within direct patient care, while otherwise identical disclosures for research and development require informed consent [101]. This model raises a number of questions for how patient data in radiology AI can be perceived. Will informed consent be required only for patient data in the development of deeply annotated AI datasets? How will conformed consent be addressed if a patient's data is used in assessing an algorithm in routine clinical practice is then used to refine/retrain the algorithm? If the data is used to develop applications sold for profit, are patients entitled to compensation? What mechanisms are in place to protect individuals who do opt out? These questions will have to be addressed as clinical AI becomes routine [102].

One key in managing the use of patient data will be transparency. In general, the public is willing to share personal data if they believe there will be downstream benefits, but they want to be confident it will not be shared in ways they do not understand. In an interview with the Harvard Business Review, MIT professor Alex Petland contradicts the notion that organizations collecting the data actually own the data. He goes on to say that without developing rules for who does, the public will revolt, regulators will get involved, and there will inevitably be restrictive overreaction, and as such, applications such as AI, which are dependent on these data, will fail to reach their potential. Petland's "New Deal on Data" proposes that transparency depends on allowing the public to see what is being collected and then allowing individuals opportunities to opt in or out [103]. The AI community should work together to create an infrastructure that allows responsible use of patients' health data to facilitate the development of AI tools that will improve population health. The industry should welcome structure around responsible data use, and having defined rules for data use will ultimately facilitate the development AI tools and hopefully prevent data breaches and other data disasters which could set the industry back decades.

Nonetheless, providing developer access to the large datasets will create the opportunity for large leaks of protected information, and new cryptography techniques should be considered [2]. Blockchain methodologies use a distributed database consisting of continuously updated (augmented) "blocks" which contain a linked list of all previous transactions [102]. In the case of healthcare, this encompasses all previous records of access to an individual data record including information about how the data was used and any additions or changes to the data record [104]. Blockchain technology can also be used to validate the provenance of data and facilitate the distribution of data without compromising the quality of data. Pilots are underway assessing the ability of blockchain type ledgers to function within Health Level 7 and FHIR standards for electronic health records. In health systems, blockchain technology may solve some problems for researchers such as localizing the most current record and tracking patient activity across a health system. Development of Merkle tree technology for health systems [104], which uses a hash function and hash values to track changes to the database, may be one way to ensure security in a distributed data system. This type of data structure allows verification of users who made changes and what changes were made making it difficult to corrupt the database since changes in the data cause changes in the hash codes. No matter which technologies are ultimately considered most effective in protecting data privacy, the AI ecosystem must embrace standards for data security and patient privacy in both centralized and distributed models for algorithm development and implementation. This will help ensure there are no systematic data breaches or other data disasters that would almost certainly impede the development and implementation of AI algorithms in healthcare [105].

19.4.4 Enhancing Algorithm Validation

In addition to enhancing the supply of datasets available for training and testing, a robust AI ecosystem should also focus on creating rigorous testing and validation approaches for the clinical use of AI algorithms in order to identify and mitigate any problems in implementation to provide confidence to the medical community. The 2017 JASON Report *Artificial Intelligence for Health and Health Care* further recommends that work to prepare and assist developers of promising AI applications navigate the regulatory and other approval processes needed for acceptance in clinical practice should be supported and include "testing and validation approaches for AI algorithms to evaluate performance of the algorithms under conditions that differ from the training set" [2]. One such approach is development of a centralized program that allows assessment of algorithm performance using novel validation datasets and the statistical metrics specified in structured AI use cases. By specifying the elements in the AI use case, algorithms can be readily compared and assessment for clinical deployment standardized. These validation datasets could be developed from an amalgam of datasets created at multiple institutions which when used in the aggregate would ensure geographic, technical, and patient diversity within the validation dataset. In addition to ensuring diversity within the validation datasets, these datasets must be held to the highest ground truth reasonably achievable by using data labeled at levels that exceed standard assessments when possible including the use of biopsy results to label dermatological images [2]. Multiple readers and guidelines for data quality should be used to ensure consistency between sites and consistent metrics for measuring performance of different algorithms built around the same use case. Internal standards to protect developers' intellectual property and to ensure patient privacy and diminish potential unintended bias in algorithm performance should also be developed. With these fundamentals in place, these validation centers could then prepare reports for developers about their algorithm's performance for use in the regulatory approval processes such as US FDA clearance. As discussed previously, the US FDA is looking for tools within the MDDT program that developers can use to facilitate the regulatory approval process. While these have not been officially established as "special controls," in the FDA's proposal to reclassify many SaMD products as Class II (special controls), the AI community should welcome the opportunity to develop a streamlined process that can move AI products expeditiously into clinical practice.

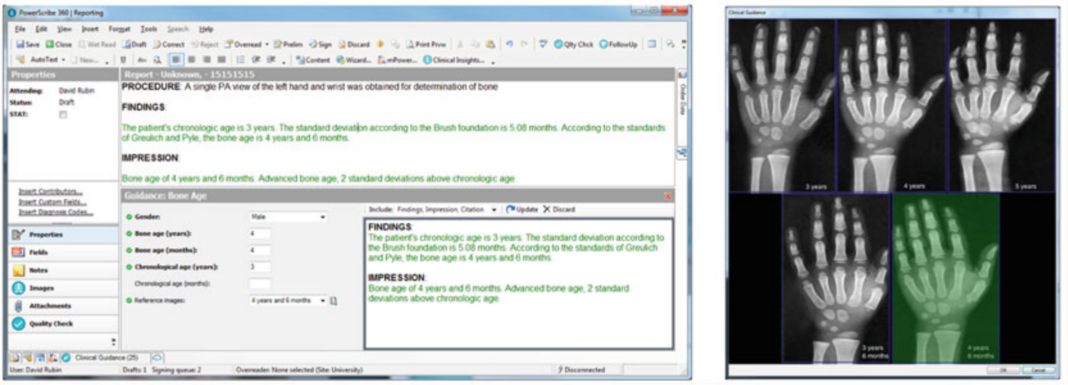
Acceptance of AI in clinical practice will be dependent on the believability and explicability of the algorithm output. The JASON 2017 report *Artificial Intelligence For Health and Healthcare* highlighted this issue by summarizing a series of studies demonstrating the value of quantitative information from cardiac fluid flow reserve computed tomography (FFRCT) for identifying patients with clinically significant coronary artery disease at less cost than invasive coronary angiography [2, 106]. The favorable results shown by these studies as well as an independent review by United Kingdom's National Institute for Health and Care Excellence (NICE) resulted in NICE issuing guidance FFRCT into the NICE pathway on chest pain [107]. Because the FFRCT technology is based data than can be readily verified in clinical practice, physician acceptance may be better than for less-understood outputs of general AI algorithms. For the medical community to develop trust in AI-based tools, assessments at least as rigorous as the FFRCT technology will be needed [2].

19.4.5 Enhancing Clinical Integration

The use of structured AI use cases will also enhance the integration of algorithms into clinical practice by defining standards for inputs and outputs (I/O) into the algorithm. The use of common data elements and specifications within the use case for how application programming interfaces (API) can ingest algorithm output allow deployment of AI models in a vendor neutral environment into clinical and operational workflows. Figure 19.14 shows how the standardized output from a pediatric bone age algorithm is incorporated into reporting software using the CARDS platform along with saliency maps to ensure algorithm transparency; however, output from AI algorithms can also be incorporated in a vendor neutral environment into existing HIT tools including reporting software, the modalities, and the EHR.

19.4.6 Mechanisms for Assessing Algorithm Performance in Clinical Practice

As methods for assessing algorithm performance in clinical practice are established, data elements in each structured AI use case can specify the appropriate data elements that should be captured in order to monitor an algorithm's performance in clinical practice. Radiologist input is gathered as the case is being reported, and if the radiologist does not incorporate the algorithm inferences into the report, this change is captured in the background by the reporting software. If the radiologist agrees with the output of the algorithm, this is also noted and transmitted to a data registry. Radiology specialty societies are also uniquely positioned to host these registries. Metadata specified in the AI use case about the examination such as equipment vendor, slice thickness, and exposure are also transmitted to the registry. Algorithm assessment reports provide a summary of the algorithm's real-world performance across a wide variety of practice settings. Areas where algorithm performance is low are correlated with examination metadata to look for patterns that will allow improvements to


PowerScribe 360 | Reporting
File Edit View Insert Format Tools Speech Help
Save Close Wet Read Draft Correct Reject Overread Prelim Sign Discard Print Prvw Qlty Chck FollowUp
AutoText New... Content Wizard... mPower... Clinical Insights...
Properties
Report - Unknown, - 15151515
Order Data
Insert Contributors...
Insert Custom Fields...
Insert Diagnosis Codes...
Properties
Fields
Notes
Images
Attachments
Quality Check
Guidance: Bone Age
Include: Findings, Impression, Citation
Update
Discard
Clinical Guidance (25)
Clinical Guidance
3 years
4 years
5 years
3 years 6 months
4 years 6 months
OK
Cancel

Attending: David Rubin
Status: Draft
STAT: □

PROCEDURE: A single PA view of the left hand and wrist was obtained for determination of bone

FINDINGS:

The patient's chronologic age is 3 years. The standard deviation according to the Brush foundation is 5.08 months. According to the standards of Greulich and Pyle, the bone age is 4 years and 6 months.

IMPRESSION:

Bone age of 4 years and 6 months. Advanced bone age, 2 standard deviations above chronologic age.

Gender: Male
Bone age (years): 4
Bone age (months): 4
Chronological age (years): 3
Chronological age (months):
Reference images: 4 years and 6 months

FINDINGS:
The patient's chronologic age is 3 years. The standard deviation according to the Brush foundation is 5.08 months. According to the standards of Greulich and Pyle, the bone age is 4 years and 6 months.

IMPRESSION:
Bone age of 4 years and 6 months. Advanced bone age, 2 standard deviations above chronologic age.

User: David Rubin Drafts: 1 Signing queue: 2 Overreader: None selected (Site: University) Disconnected

Fig. 19.14 Standardized output from AI algorithms in clinical workflow (Courtesy Nuance, RSNA 2017. Used with permission)

the algorithm through additional training. These reports will also be useful to developers in reporting real-world performance to regulatory agencies such as the US FDA and to the clinical sites to ensure their algorithm performance is in line with national benchmarks.

The American College of Radiology National Radiology Data Registry (NRDR) [70] is an example of how radiology specialty societies are helping the specialty capture and benchmark information about quality, patient safety, and other improvement activities. AI data registries can potentially capture both radiologist assessment and metadata about the examination without hampering clinical workflow. The results can be collated centrally and provided to developers and the clinical sites to ensure patient safety and improve algorithm effectiveness.

19.4.7 The Economics of AI and Business Models for Moving AI to Clinical Practice

A key ingredient in moving artificial intelligence (AI) algorithms for healthcare into routine clinical practice will be ensuring our healthcare system supports the fair compensation for the development of these algorithms and other AI tools, but developing a process for how that will happen may not be as simple as it might seem. Costs in the US healthcare system are already at unsustainable levels, and so developers and the physician community will have to demonstrate the value and cost savings that each artificial intelligence algorithm brings to our patients and our healthcare system before reimbursement from third-party payers can be considered. The value to patients may be in earlier and more accurate diagnoses and treatments. The value to physicians may be in improved efficiency in data management and integration, and the value to our health systems may be in improved quality of care, overall efficiency, and decreased length of stay.

Developers will need understanding of current and future payment models to develop sustainable business models and has to begin with the current US fee-for-service (FFS) model. In this system, specific medical services, procedures, and supplies are reimbursed using the Center for Medicare and Medicaid Service's (CMS) Healthcare Common Procedure Coding System [108]. Level I of the HCPCS system is based on Current Procedural Terminology^{TM} (CPT), which is a numeric coding system developed and maintained by the American Medical Association. The CPT system identifies and describes medical services and procedures commonly furnished and billed by physicians and other healthcare professionals. However, CPT does not include the codes needed to separately report medical items or services for patients that are provided outside of the physician office setting, such as durable medical equipment and supplies. The Level II HCPCS was established to provide codes for the non-physician providers to submit claims for these items to Medicare and private health insurance programs. Each HCPCS code is assigned a value by Medicare and other payers, and claims are submitted by providers based on these codes. When medical equipment and supplies are used in the physician office setting, the reimbursement for these items is included in the CPT code payment to the physician as "Direct Practice Expense"; however, when the same services are performed by physicians in the hospital or site of service other than a physician office, the payments for equipment and supplies payments are made directly to the facility. As such, each CPT code in the Medicare Physician Fee Schedule (PFS) has different payments to physicians based on whether the service was provided in a physician's office (non-facility) or hospital (facility) setting [108, 109]. Finally, a portion of the payment for each physician service ("Indirect Practice Expense") is designed to cover the costs of operating a practice including office rent, utilities, computers, and billing costs. The Medicare PFS uses the resource-based relative value scale (RVRVS) to assign relative value units (RVUs) for each physician service, and then all of the practice expenses are then converted to RVUs. RVUs for physician work and compensation for professional liability insurance are added to the direct and indirect practice expense RVUs to comprise the total RVUs for each physician service in the Medicare PFS, which is then multiplied by a conversion factor set by CMS to give the dollar payment to physicians. Hospitals are reimbursed under two separate payment systems, the inpatient prospective payment system (IPPS), which uses diagnosis-related groups (DRG) as its fundamental coding system, and the hospital outpatient prospective payment system (HOPPS), which uses ambulatory payment classification (APC) as its fundamental coding system. Each of these systems accounts for the payments for medical equipment, devices, and supplies in different ways. And while some private payers base their payment systems on the Medicare PFS, each private insurer has their own way assigning reimbursement for medical equipment, devices, and supplies to each service.

While the various US payment systems are complicated in their own right, the process is made even more complicated because there will not be a one-size-fits-all payment scheme for reimbursing the use of AI in healthcare. Some algorithms will affect payments to physicians, perhaps making their work more efficient or perhaps more time consuming as we bring in more and more patient information into our care of complex patients. Some algorithms will improve the overall quality and efficiency of our practices and health systems but cannot be attributed or assigned to a specific service or procedure, and while some algorithms may be directly reimbursable by third-party payers, many will not. Finally, all algorithms that are adopted by physicians and our health systems must be able to document that they are providing demonstrable value to our patients in a safe and bias-free environment.

The US CMS Quality Payment Program (QPP) [110] is the next step in the development and adoption of alternate payment models (APMs) in US healthcare. The QPP includes the Merit-based Improvement Payment System (MIPS) and Alternate Payment Models (APM). The MIPS uses four categories—quality, clinical practice improvement, resource use, and advancing care information—to adjust Medicare FFS payments to physicians, up or down by as much as 9% in 2022, based on their performance in each category. Measures for quality, clinical practice improvement activities, and advancing care information are reported to CMS by physicians, and if certain AI algorithms are able to provide documented value and improved quality to our patients, the use of the algorithms to improve patient care, quality, and value can be included as MIPS measures. While APMs are much less prevalent in the United States, algorithms that increase overall efficiency for health systems will be welcomed as the medical community strives to do more for our patents at a lower overall cost. In the alternate payment models, assigning and attributing a per unit cost of an AI algorithm to an individual CPT code will be much less important than ensuring the algorithm functions in a way that augments the care provided to patients without taking away the commonsense decisions of physicians and our patients.

Finally, the economics of AI in healthcare will have to include a discussion about potential disparities if AI is available to some patients and not available to others. While market leaders will likely emerge touting that their services include the latest AI innovations, the global healthcare system should not devolve into a two-tier system where some can afford AI, while others cannot. The reimbursement system has a duty to protect our patients by ensuring all physicians have access to these potentially revolutionary tools.

Radiology specialty societies such as the American College of Radiology have always been strategically involved in the federal regulatory and payment policy issues around the radiological sciences. Reimbursement issues for moving artificial intelligence into clinical practice will have to be considered in the payment policy arena. Specialty societies can function in the AI ecosystem to provide education around regulatory payment policy issues around AI, and these policy issues were discussed with developers, physicians, and the AI community at the ACR Data Science Institute's Data Science Summit: The Economics of AI in conjunction with Society for Imaging Informatics in Medicine (SIIM). The proceedings of this summit are freely available to the community [111].

Medial specialty societies also play important roles in interacting with regulatory agencies including the US FDA, the International Atomic Energy Agency, and the World Health Organization (WHO), all of which may play an eventual role in regulating healthcare AI. Radiology specialty societies are uniquely positioned to serve as honest brokers with these regulatory agencies facilitating processes that advance the use of AI in clinical practice while protecting patients by ensuring algorithms are safe and effective in clinical use.

19.4.8 Facilitating the Development of Non-interpretive Use Cases for Artificial Intelligence in Radiological Practice

Non-interpretive AI algorithms will also be important for the radiology professionals [12]. Use cases that promote quality, safety, protocol optimization, patient experience, and many others will be valuable to both radiology professionals and hospital systems. End users will not only include radiologists but also technologists, hospital administrators, hospital quality team, and hospital finance team. As with the interpretive-based AI use cases, development of appropriately curated data will be necessary for algorithm training, and demonstration projects will be needed to demonstrate the clinical utility. While these types of algorithms may not require regulatory approval, processes to ensure algorithms are effective and free of unintended bias will be important. Radiologists and radiological specialty societies can play an active role in facilitating the development of AI tools for non-interpretive uses by developing use cases for researchers and developers that address important workflow, patient access, and numerous non-interpretive issues in the radiological community. Developing standards for interoperability for using AI across the entire health enterprise will be even more important for developing non-interpretive uses for AI than the interpretive uses. Not only will data from imaging studies be necessary, but data from a variety of electronic resources will also be needed to bring in the additional information to accomplish these uses of AI. Radiologists and radiology specialty societies should play leading roles in working with all of medicine and the HIT community to develop these interoperability standards with artificial intelligence in mind. Additionally, specialty societies can coordinate piloting demonstration projects that can be used to establish utility and effectiveness of AI in using the abundance of data in the health systems, and the AI community should look for methods that can continuously monitor the effectiveness of these tools as they are deployed in clinical practice.

19.4.9 Educating Non-radiologist Stakeholders About the Value of AI

Fostering collaborations between stakeholders requires education demonstrating the value of establishing an AI ecosystem. Radiology specialty societies can foster collaborations between organizations establishing joint educational programs and other defined collaborations. These same organizations as well as governmental agencies and the developer community can provide venues that bring all stakeholders together. The Machine Learning Showcase at RSNA 2017 gathered AI developers into a common location and also provided a venue for education [89]. There have been a number of events that included industry at meetings such as the Society for Imaging Informatics in Medicine and the American College of Radiology [111]. The American College of Radiology DSI also hosted FDA representatives in the Fall of 2017. Finally, technology companies such as NVIDIA have hosted educational meetings where radiology specialty societies were invited to provide their perspectives on AI. These collaborative meetings should continue [112].

Additionally, radiology specialty societies are working to prepare the profession for the opportunities AI will bring. Rather than opposing AI initiatives as a threat to the specialty, radiology organizations have been providing educational activities that demonstrate how AI will help radiology professionals take better care of their patients and in turn be more valuable to their health systems [113].

19.5 Summary of the Proposed AI Ecosystem for the Radiological Sciences

To ensure AI applications in the radiological sciences are implemented in ways that add to the value radiologists provide to their patients' medical care, radiology professionals have an opportunity, if not an obligation, to be involved in the development and implementation of this new technology. Radiologists and developers can develop a synergistic relationship that promotes widespread adoption of AI in clinical practice. Radiology professionals should collaborate to create structured AI use cases addressing specific clinical needs across the profession and ensure there are pathways for workflow integration during clinical deployment. AI researchers and commercial developers should use these standardized use cases to create AI models with vendor neutral interoperable outputs that allow widespread use in clinical practice. Ideas for AI use cases can come from the societies themselves, academic institutions, developers, or individual radiologists but should be built using a standardized process using common data elements, vendor neutral inputs for the algorithm, and interoperable outputs from the algorithm. No matter the source of the use case idea, radiology specialty societies can facilitate these collaborations and provide an infrastructure that supports standardized and robust use case development (Fig. 19.15a).

Another opportunity for radiologist participation in the AI development process is in the production of well-annotated datasets for algorithm testing and training. While many radiologists have begun working with individual developers to annotate data for use in algorithm development, by using structured use cases as the basis for this effort, many practices can create training data based on the specifications in the AI use case that can then be used in aggregate by developers for algorithm development, training, and testing. The aggregated data provide a diverse mix of technical differences and variability in patient populations typical of widespread clinical practice. The healthcare ecosystem including physicians, healthcare administrators, government regulators, and patient advocates should support these efforts by offering standardized methodologies for deidentifying sensitive patient information across the health system so that development of AI in healthcare can proceed at a reasonable pace. A potentially important consideration is that if developers can use the datasets created by individual radiology practices on-prem at the institution rather than a centralized offsite location out of the institution's direct control, the patient information is much better controlled and protected than if contained in a centralized repository or completely under the control of individual developers. This model allows development of a large data pool with technical and geographic diversity while avoiding the risks associated with a large centralized repository of patient data (Fig. 19.15b).

In a robust AI ecosystem, there should be many opportunities for radiological practices to participate in validation of AI algorithms for the radiological sciences. Structured AI use cases should contain the data elements and statistical metrics necessary to ensure algorithms will be safe and effective in clinical practice, and developers should be aware in advance how the algorithms will be evaluated. Furthermore, to facilitate comparison, similar algorithms should be evaluated using similar statistical metrics. Since the algorithm performance against validation datasets may be used to obtain premarket regulatory approval for many AI applications, the validation process must be robust, standardized, and statistically valid. This means that standards for ground truth must be higher than those for creating training datasets and should even exceed the standards for routine clinical practice.

In contrast to the training datasets, the validation datasets should be held centrally. A

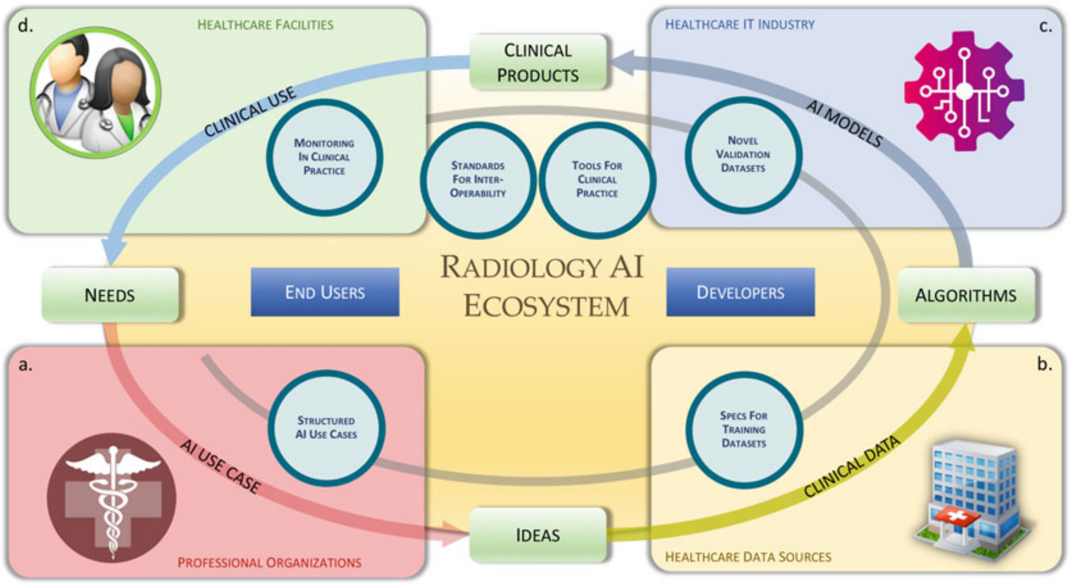


Fig. 19.15 (**a**) Radiologists can play a leading role in the creation of AI use cases that address clinical needs and are readily integrated into clinical workflows by working with specialty societies to develop structured and standardized use cases; (**b**) annotated datasets created based on the specifications from a structured use case allow multiple practices to contribute data to algorithm training and testing which provides developers with more technically and geographically diverse data than when working with single institutions. By working with practices on-prem, institutions can protect patient privacy by maintaining better control of patient data than in a centralized repository; (**c**) validation of AI algorithms to ensure they will be safe and effective in clinical practice will be key to general acceptance of AI in healthcare. When AI algorithms are built around structured use cases, robust and technically and geographically diverse validation datasets can be developed and hosted by a third-party honest broker such as medical specialty societies. Standardized statistical metrics can become the basis of reports of algorithm performance that can be used by developers in the regulatory process. (**d**) Structured AI use case can also specify data elements that will allow real-world performance of AI algorithms in clinical practice. These data, including algorithm performance and metadata about the examination, can be collected and housed in data registries. Radiologists will play a key role in assessing algorithm performance in clinical practice, and radiology specialty societies can play a leadership role in hosting data registries. The radiology AI ecosystem is a critical collaboration to harmonize clinical needs and ideas to marketable AI products that will be safe and effective in clinical practice (Credit authors)

centralized repository of validation datasets maintained by a “third-party” honest broker promotes confidentiality so that the validation data cannot used for algorithm training. Additionally, safeguards can be in place to ensure protection of patient information as well as developer intellectual property. A natural host for algorithm validation would be radiological specialty societies. For example, the American College of Radiology (ACR) has developed an infrastructure designed to support multicenter clinical trials using imaging data. This infrastructure includes the ability to transmit DICOM, HL7, and other data sources from clinical sites to a central repository along with tools for data curation and aggregation to combine the results from individual sites into a combined result [114, 115]. Demonstrating the effectiveness of third-party validation of AI algorithms will be important in order to convince regulatory bodies such as the US FDA that these processes could be used in an AI algorithm’s premarket approval process. Once again, professional medical societies have a role to play in interacting with governmental agencies to facilitate a review process that facilitates AI development while ensuring the safety of patients and the public. The ACR has a long history of interacting with the US FDA to promote radiological quality and safety, particularly implementation of the Mammography Quality Standards Act (MQSA) [116] and radiation safety issues. The US FDA adopted the ACR Accreditation Program as a means to demonstrate MQSA compliance. Therefore, it seems that public-private partnerships between governmental regulatory agencies and medical specialty societies could be developed for AI in healthcare as well (Fig. 19.15c).

Once an AI product has received regulatory clearance for marketing, developing pathways for clinical implementation of AI models will be necessary. Structured use cases should contain data elements that specify how the output of the algorithm should interact with other electronic resources. Standardization of algorithm output so that the data can be used to inform the electronic resources used by physicians is necessary, and more robust standards for communicating between the array of electronic healthcare resources should be developed as well. Physicians and professional societies should also play a role in this process as well. The ACR and NEMA created DICOM to move image data between the electronic interfaces used by radiology. Professional organizations should be involved in development of standards that allow movement of AI inference model outputs to the most usable locations in a patient's medical records.

Finally, to gain wide acceptance in the healthcare markets, being able to assure end users and the public that the AI applications used in medical practice perform as expected cannot be overstated. Physicians and patients will expect nothing less, but collection of real-world performance data will not be trivial. Physicians will not want to be distracted from their clinical workflows in order to complete and submit forms or other data designed to monitor performance in practice, and even data collected in that manner is likely to be unhelpful in systematically monitoring the real-world performance of AI algorithms. To mitigate these challenges, structured AI use cases can contain data elements specifying pathways for how AI algorithms will be monitored in routine clinical practice. For instance, AI algorithms designed to assist radiologists in lesion characterization could display the AI inference in the PACS and prepopulate a radiologists' report. If the radiologist does not change the report, then the algorithm is considered to have worked as expected. If the radiologist changes the report beyond a predefined tolerance, then the algorithm is considered to have failed for that examination. To help understand potential reasons for algorithm failure, the transcription or other system can collect metadata about the examination specified in the use case in the background. For each instance of algorithm use, radiologist agreement and metadata can be transmitted to a registry for aggregation and collation. Reports regarding algorithm performance can be generated for developers to ensure compliance with any post-market regulatory requirements. By correlating algorithm performance with examination data, developers can understand which examination parameters may be associated with poor algorithm performance and expand training and testing to include those circumstances. These data can be collected and housed in data registries. Currently many medical specialty societies offer the collection and benchmarking of practice data [61]. In some instances, data registries housed by specialty societies have dramatically improved the cost of premarket review for FDA clearance [117]. If these processes can be implemented on a widespread basis, radiologists will be in the center of ensuring the development and use of AI in clinical practice reaches its potential, and feedback from clinical use will be the best way to assist developers improve software and expand algorithms into more and more clinical problems (Fig. 19.15d).

19.6 Conclusion

The development and implementation of AI algorithms for use in routine clinical practice will benefit from the establishment of an AI ecosystem that leverages the value of radiologists and radiology specialty societies from the development of AI use cases to assessing the use of AI in routine clinical practice. Such an ecosystem includes not only physicians, researchers, and software developers but also regulatory agencies, the HIT industry, and hospital administrators. By developing structured AI use cases based on the needs of the physician community, developers can create the tools that will advance the practice of medicine. If these use cases can specify how datasets for algorithm training, testing, and validation can be developed including statistical metrics for validation, parameters for clinical integration, and pathways for assessing algorithm performance in clinical practice, the likelihood of bringing safe and effective algorithms to clinical practice will increase dramatically. Additional challenges for the community such as respecting patient privacy, technical and geographic diversity, as well as decreasing unintended bias in algorithm development will be best solved with collaboration between all stakeholders. Finding a balance between promoting innovation and respecting and protecting confidential patient information will also require a consensus between the healthcare community and the public, and finally the healthcare community must come together to promote interoperable standards so that data from AI algorithms can be delivered to the electronic resource where it can be most useful to physicians and their patients. The development of an active AI ecosystem will facilitate the development and deployment of AI tools for healthcare that will help physicians solve medicine's important problems.

Take-Home Points

- Moving artificial intelligence tools in diagnostic imaging to routine clinical practice and avoiding another AI winter will require cooperation and collaboration between developers, physicians, regulators, and health system administrators.
- Radiologists can play an important role in promoting this AI ecosystem by delineating AI use cases for diagnostic imaging and standardizing data elements and workflow integration interfaces.
- Medial specialty societies can play a leading role in protecting patients from unintended consequences of AI through involvement in algorithm validation.
- AI registries will be useful in monitoring the effectiveness and safety of AI tools in clinical practice.

References

1. Allen B, Dreyer K. The artificial intelligence ecosystem for the radiological sciences: ideas to clinical practice. J Am Coll Radiol. 2018; https://doi.org/10.1016/j.jacr.2018.02.032.
2. JASON 2017. Artificial intelligence for health and heath care. JSR-17-Task-002.
3. Definition of Ecosystem. [Internet]. Merrian-webster.com. 2018 [cited 10 June 2018]. Available from: https://www.merriam-webster.com/dictionary/ecosystem
4. Moore JF. Predators and prey: a new ecology of competition. Harv Bus Rev. 1993 May 1;71(3):75–86.
5. Moore JF. The death of competition: leadership and strategy in the age of business ecosystems. New York: HarperBusiness; 1996 May.
6. Messerschmitt DG, Szyperski C. Software ecosystem: understanding an indispensable technology and industry, vol. 1. London: MIT Press Books; 2005.
7. Seddon JJ, Currie WL. Cloud computing and trans-border health data: unpacking US and EU healthcare regulation and compliance. Health Policy Technol. 2013 Dec 1;2(4):229–41.
8. Barnett, JC, Berchick, ER. Current population reports, P60–260, Health Insurance Coverage in the United States: 2016, U.S. Washington, DC: Government Printing Office; 2017.
9. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff. 2008 May;27(3):759–69.
10. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. Ann Fam Med. 2014 Nov 1;12(6):573–6.
11. Sikka R, Morath JM, Leape L. The Quadruple Aim: care, health, cost and meaning in work. BMJ Qual Saf. https://doi.org/10.1136/bmjqs-2015-004160.
12. Thrall JH, Li X, Li Q, Cruz C, Do S, Dreyer K, Brink J. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. J Am Coll Radiol. 2018 Mar 1;15(3):504–8.
13. Lakhani P, Prater AB, Hutson RK, Andriole KP, Dreyer KJ, Morey J, Prevedello LM, Clark TJ, Geis JR, Itri JN, Hawkins CM. Machine learning in radiology: applications beyond image interpretation. J Am Coll Radiol. 2017 Nov 17;15(2):350–9.
14. Erdal BS, Prevedello LM, Qian S, Demirer M, Little K, Ryu J, O'Donnell T, White RD. Radiology and Enterprise Medical Imaging Extensions (REMIX). J Digit Imaging. 2018 Feb 1;31(1):91–106.

15. Huffman J. Healthcare Information and Management Systems Society. 2018 March 6.
16. Turing AM. Computing machinery and intelligence. Mind. 1950 Oct;59(236):433.
17. Minsky M. Steps toward artificial intelligence. Proc IRE. 1961 Jan;49(1):8–30.
18. McCarthy J. From here to human-level AI. In Proc. of principles of knowledge representation and reasoning (KR 1996).
19. Taubes G. The rise and fall of thinking machines. Inc. 1995;17(13):61–5.
20. Yang Z, Zhu Y, Pu Y. Parallel image processing based on CUDA. In Computer Science and Software Engineering, 2008 International Conference on 2008 Dec 12 (vol. 3, pp. 198–201). IEEE.
21. Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. In Computer vision and pattern recognition (CVPR), 2012 IEEE conference on 2012 Jun 16 (pp. 3642–3649). IEEE.
22. Mobile Fact Sheet. Pew Research Center: Internet, Science & Tech. 2018 [cited 10 June 2018]. Available from http://www.pewinternet.org/fact-sheet/mobile/
23. Chockley K, Emanuel E. The end of radiology? Three threats to the future practice of radiology. J Am Coll Radiol. 2016 Dec 1;13(12):1415–20.
24. Remnick D. Obama reckons with a Trump presidency. The New Yorker. 2016 Nov;28:28.
25. Hinton G. Geoff Hinton on Radiology. Machine Learning and Market for Intelligence Conference, Creative Disruption Lab Toronto, Canada. 2016. Viewable at: https://www.youtube.com/watch?v=2HMPRXstSvQ
26. Oncology Expert Advisor [Internet]. MD Anderson Cancer Center. 2018 [cited 10 June 2018]. Available from: https://www.mdanderson.org/publications/annual-report/annual-report-2013/the-oncology-expert-advisor.html
27. Herper M. MD Anderson benches IBM Watson in setback for artificial intelligence in medicine. Forbes. Zugriff im Juli. 2017 Feb.
28. Ravì D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, Yang GZ. Deep learning for health informatics. IEEE J Biomed Health Inform. 2017 Jan;21(1):4–21.
29. Deo RC. Machine learning in medicine. Circulation. 2015 Nov 17;132(20):1920–30.
30. Valente IR, Cortez PC, Neto EC, Soares JM, de Albuquerque VH, Tavares JM. Automatic 3D pulmonary nodule detection in CT images: a survey. Comput Methods Programs Biomed. 2016 Feb 1;124:91–107.
31. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in X-ray images. Med Image Anal. 2017 Feb 1;36:41–51.
32. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency 2018 Jan 21 (pp. 77–91).
33. Health Insurance Portability and Accountability Act of 1996 (HIPAA.)Pub. L. 104–191, 110 Stat. 1936 (1996)
34. The HIPAA Privacy Rule. 45 CFR 160, 162, and 164. 28 Dec 2000.
35. The Security Rule. 45 CFR Part 160 and Subparts A and C of Part 164. 20 Feb 2003.
36. Artificial Intelligence For Health and Health Care. https://www.healthit.gov/sites/default/files/jsr-17-task-002_aiforhealthandhealthcare12122017.pdf
37. AI has no place in the NHS If patient privacy isn't assured. Wired. http://www.wired.co.uk/article/ai-healthcare-gp-deepmind-privacy-problems
38. US Food and Drug Administration. What we do. https://www.fda.gov/AboutFDA/WhatWeDo/
39. US Food and Drug Administration. Medical Devices.
40. The 21st Century Cures Act. Pub. L. 114–255.
41. US Food and Drug Administration. Response To 21st Century Cures Act. https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM587820.pdf
42. US Food and Drug Administration. Software as a medical device. Do. https://www.fda.gov/MedicalDevices/DigitalHealth/SoftwareasaMedicalDevice/default.htm
43. US Food and Drug Administration. International Medical Device Regulators Forum. https://www.fda.gov/MedicalDevices/InternationalPrograms/IMDRF/default.htm
44. Qualification of Medical Device Development Tools. https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM374432.pdf
45. US Food and Drug Administration. Medical Device Development Tools Program. https://www.fda.gov/MedicalDevices/ScienceandResearch/MedicalDeviceDevelopmentToolsMDDT
46. US Food and Drug Administration. National Evaluation System for Health Technology. https://www.fda.gov/aboutfda/centersoffices/officeofmedicalproductsandtobacco/cdrh/cdrhreports/ucm301912.htm
47. US Food and Drug Administration. National evaluation system for health technology demonstration projects. https://nestcc.org/demonstration-projects/
48. Lund-RADS Assist: Advanced radiology guidance, reporting and monitoring. https://www.acr.org/Media-Center/ACR-News-Releases/2018/FDA-NEST-Program-Names-ACR-DSI-Use-Case-as-Demo-Project
49. Digital Health Software Precertification Program. https://www.fda.gov/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/default.ht
50. US FDA Software Precertification Program. https://www.fda.gov/downloads/MedicalDevices/

DigitalHealth/DigitalHealthPreCertProgram/UCM605685.pdf

51. US FDA Classification of Medical Devices. https://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/Overview/ClassifyYourDevice/
52. US FDA de novo request. https://www.fda.gov/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDRH/CDRHTransparency/ucm232269.htm
53. US FDA de novo approval clinical decision support software for stroke. https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm596575.htm
54. US FDA de novo approval artificial intelligence based device to detect diabetes related eye problems. https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm604357.htm
55. US FDA de novo approval of artificial intelligence algorithm for aiding providers in detecting wrist fractures. https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm608833.htm
56. US FDA CADe and CADx. https://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm187249.htm
57. US FDA CADe and CADx. https://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm187277.htm
58. USFDA approval QuantX as Class II device. https://www.accessdata.fda.gov/cdrh_docs/pdf17/DEN170022.pdf
59. Boland GW, Duszak R, McGinty G, Allen B. Delivery of appropriateness, quality, safety, efficiency and patient satisfaction. J Am Coll Radiol. 2014 Jan 1;11(1):7–11.
60. ACR, Imaging 3.0. http://www.acr.org/Advocacy/Economics-Health-Policy/Imaging-3.
61. Imaging 3.0. https://www.acr.org/-/media/ACR/Files/Imaging3/Imaging3_Overview.pdf
62. LOINC. Available at: http://loinc.org/about/
63. Alkasab TK, Bizzo BC, Berland LL, Nair S, Pandharipande PV, Harvey HB. Creation o an open framework for point-of-care computer-assisted reporting and decision support tools for radiologists. J Am Coll Radiol. 2017 Sep 1;14(9):1184–9.
64. A Brief History of DICOM. In: Digital Imaging and Communications in Medicine (DICOM). Berlin, Heidelberg: Springer; 2008.
65. HL7 protocols. http://www.hl7.org
66. Fast Healthcare Interoperability Resources Specification. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=449
67. Rubin DL, Kahn CE Jr. Common data elements in radiology. Radiology. 2016 Nov 10;283(3):837–44.
68. Winget MD, Baron JA, Spitz MR, Brenner DE, Warzel D, Kincaid H, Thornquist M, Feng Z. Development of common data elements: the experience of and recommendations from the early detection research network. Int J Med Inform. 2003 Apr 1;70(1):41–8.
69. Morin RL, Coombs LP, Chatfield MB. ACR dose index registry. J Am Coll Radiol. 2011 Apr 1;8(4):288–91.
70. ACR National Radiology Data Registry. https://nrdr.acr.org/Portal/Nrdr/Main/page.aspx
71. Langlotz CP. RadLex: a new method for indexing online educational materials. Radiographics. 2006;26(6)
72. Structured Reporting. http://www.radreport.org
73. ACR Select. https://www.acr.org/Clinical-Resources/Clinical-Decision-Support
74. Boland GW, Thrall JH, Gazelle GS, Samir A, Rosenthal DI, Dreyer KJ, Alkasab TK. Decision support for radiologist report recommendations. J Am Coll Radiol. 2011 Dec 1;8(12):819–23.
75. Rad Elements. http://www.radelement.org
76. Miller T, Howe P, Sonenberg L. Explainable AI: Beware of inmates running the asylum. InIJCAI-17 Workshop on Explainable AI (XAI). 2017 (p. 36).
77. American Medical Association Policy. https://www.ama-assn.org/ama-passes-first-policy-recommendations-augmented-intelligence
78. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res. 2016 Dec;18(12)
79. Stodden V. Reproducible research for scientific computing: Tools and strategies for changing the culture. Comput Sci Eng. 2012 Jul;14(4):13–7.
80. Data Science Bowl Lung Cancer Detection. http://blog.kaggle.com/2017/06/29/2017-data-science-bowl-predicting-lung-cancer-2nd-place-solution-write-up-daniel-hammack-and-julian-de-wit/
81. Iglovikov V, Rakhlin A, Kalinin A, Shvets A. Pediatric Bone Age Assessment Using Deep Convolutional Neural Networks. arXiv preprint arXiv:1712.05053. 2017 Dec 13.
82. Kaggle https://www.kaggle.com/c/imagenet-object-localization-challenge
83. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, Lungren MP. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv preprint arXiv:1711.05225. 2017 Nov 14.
84. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016 Dec 13;316(22):2402–10.
85. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Feb;542(7639):115.
86. FDA Announcements. https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/default.htm

87. Reclassification of Medical Image Analyzers. https://www.federalregister.gov/documents/2018/06/04/2018-11880/radiology-devices-reclassification-of-medical-image-analyzers
88. https://www.cbinsights.com/research/artificial-intelligence-startups-healthcare/
89. RSNA Machine Learning Showcase. https://www.rsna.org/Machine-Learning-Showcase/
90. http://www.healthcareitnews.com/news/combination-pacs-and-ai-helps-uncover-what-radiologists-sometimes-miss
91. Jacobson I. Object-oriented development in an industrial environment. ACM SIGPLAN Not. 1987 Dec 1;22 (12):183–191). ACM.
92. Alistair C. Writing effective use cases. Michigan: Addison-Wesley; 2001.
93. ACR DSI. https://www.acrdsi.org/Use-Case-Development
94. Competitions Kaggle Data Science Bowl. https://www.kaggle.com/c/data-science-bowl-2017
95. Competitions Kaggle Lung Cancer Risk. https://www.kaggle.com/c/msk-redefining-cancer-treatment
96. Competitions Kaggle Heart Disease. http://www.datasciencebowl.com/competitions/transforming-how-we-diagnose-heart-disease/
97. Competitions Kaggle Seizure Prediction. https://www.kaggle.com/c/seizure-prediction
98. Personal communication. (soon in press_Andriole, Katherine. MGH and BWI Center For Clinical Data Science.
99. Lung-RADS American College of Radiology. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads
100. ACR MICCAI Collaboration. https://www.acr.org/Media-Center/ACR-News-Releases/2018/ACR-and-MICCAI-to-Leverage-AI-Algorithms-to-Meet-Clinical-Needs-in-Radiology
101. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow PM, Zietz M, Hoffman MM, Xie W. Opportunities and obstacles for deep learning in biology and medicine. bioRxiv. 2018 Jan;1:142760.
102. Balthazar P, Harri P, Prater A, Safdar NM. Protecting your patients' interests in the era of big data, artificial intelligence, and predictive analytics. J Am Coll Radiol. 2018 Mar 1;15(3):580–6.
103. Berinato S. With big data comes big responsibility. Harv Bus Rev. 2014;92(11):20.
104. Merkle RC. A digital signature based on a conventional encryption function. In Conference on the theory and application of cryptographic techniques 1987 Aug 16 (pp. 369–378). Berlin, Heidelberg: Springer.
105. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. Science. 2014 Mar 14;343(6176):1203–5.
106. Clinical trials. https://clinicaltrials.gov/ct2/show/NCT01189331
107. Ekblaw A, Azaria A, Halamka JD, Lippman A. A case study for blockchain in healthcare: "MedRec" prototype for electronic health records and medical research data. In Proceedings of IEEE Open & Big Data Conference 2016 Aug 22 (vol. 13, p. 13).
108. https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/HCPCS_Coding_Questions.html
109. https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/medcrephysfeeschedfctsht.pdf
110. https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/MACRA-MIPS-and-APMs/MIPS-Scoring-Methodology-slide-deck.pdf
111. ACR Data Science Institute Data Science Summit. https://www.acrdsi.org/dsisummit2018
112. NVIDIA GTC. https://www.nvidia.com/en-us/gtc/
113. https://www.acrdsi.org/Resources/Recommended-Reading
114. ACR TRIAD. https://triadhelp.acr.org
115. ACR DART. https://dart.acr.org
116. MQSA public Law. PL 102-539.
117. FDA and Registries. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMA/pma_pas.cfm?t_id=439786%26;c_id=380

Advantages, Challenges, and Risks of Artificial Intelligence for Radiologists

20

Erik R. Ranschaert, André J. Duerinckx, Paul Algra, Elmar Kotter, Hans Kortman, and Sergey Morozov

20.1 Innovation in Radiology

Within the past few decades, radiology has witnessed the introduction of amazing new discoveries such as ultrasound, CT, MRI, and PET and PACS (digital image storage), while more recently hospitals and healthcare systems have developed big electronic health record (EHR) systems. We now also have cloud-based image transfer systems, not only to exchange image information between providers but also to deliver our images and radiology reports directly to our patients. These technologies have vastly improved the care of patients across the world. Nevertheless several improvements still can be made, such as the optimization of the interoperability between all digital systems in the hospital. The patient perception aspect of radiology could also make great progress by further automation of basic tasks, such as streamlining patient appointments, reducing waiting times, and timely delivery of results and images to providers and patients. We have accepted the use of sophisticated computer post-processing for three-dimensional (3D) visualization of complex anatomy or guidance prior to orthopedic surgery. We have embraced automated mapping of dynamic contrast enhancement for breast MRI and other clinical applications. And more recently we have witnessed the use of complex clinically relevant but time-consuming computations such as the calculation of fractional flow reserve (FFR) after a cardiac CT angiogram. On the other hand not all of us have accepted existing systems that can enhance our ability to detect disease or subtle abnormalities because of an added cost exceeding perceived benefit (such as computer-aided detection in mammography, as explained in the chapter on breast cancer).

E. R. Ranschaert (✉) · H. Kortman
ETZ Hospital, Tilburg, The Netherlands

A. J. Duerinckx
Howard University College of Medicine and Howard University Hospital, Washington, DC, USA

P. Algra
Department of Radiology, Northwest Hospital Group, Alkmaar, The Netherlands

E. Kotter
Department of Radiology, Medical Center – University of Freiburg, Freiburg, Germany
e-mail: elmar.kotter@uniklinik-freiburg.de

S. Morozov
Radiology Research and Practical Centre, Moscow, Russia

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2_20

20.1.1 Artificial Intelligence (AI) Is the Next Big Thing

The “next big thing” in radiology is not a new type of scanner or storage device or image delivery system, but technologies to improve or accelerate image data interpretation or to facilitate tasks with less or no input from a radiologist: this is what most people refer to as “artificial intelligence” (AI) in radiology [1]. When talking about AI, a clear distinction should be made however between “narrow AI” and “general AI.” Narrow AI is referring to a machine’s ability to perform one or more well-defined tasks extremely well (sometimes even better than humans) and is applicable to most of the new developments related to medical imaging. The term general AI refers to a higher level of AI in which machines are able to think and function as a human mind, or even better. This level of AI is still beyond our reach at this moment, and certainly not yet applicable for AI applications in radiology. In this chapter we restrict the use of the term “artificial intelligence” to refer to only narrow types of AI, which are mainly based upon machine learning (ML) and deep learning (DL). Whereas ML is a technique using algorithms to parse data, learn from data, and make informed decisions based on what it has learned, DL is a subfield of ML based upon algorithms built in several layers to create an “artificial neural network” that can learn and make intelligent decisions on its own. The recent “hype” about AI in radiology is mainly due to the success of the DL-based tools for analyzing medical images. In less than a decade computers and algorithms based upon DL have gained the power to equal or exceed humans in an increasing number of simple tasks, such as the detection of pneumonia on a chest X-ray or the analysis of white matter lesions on MRI scans of the brain [2].

The astute observers and readers of this book undoubtedly will realize that our generation of physicians is witnessing the beginning of a new revolution in medicine and radiology. However, there is both a lot of excitement and uneasiness about the disruptive potential of DL and artificial AI, certainly in radiology. Numerous news reports would appear to announce an imminent takeover of the profession by computers armed with ML-based software. According to a recent column in The Economist called “Free exchange,” the new AI algorithms would soon become better than radiologists and take over their jobs [3]. Moreover, in this chapter it is stated that, whereas existing AI applications may at first work best as a complement to the radiologist skill, these algorithms will in effect be trained by radiologists to do their task better, and thus may eventually take over parts of their jobs. Eric Topol considers DL as an autodidact—like an outstanding radiology resident, the more images it analyzes, the better it gets [4]. In an article from the New England Journal of Medicine it was predicted that machine learning (ML) will produce big winners and losers in healthcare, with radiologists and pathologists among the biggest losers [5]. According to the authors, Dr. Ziad Obermeyer of Harvard Medical School and Brigham and Women’s Hospital and Ezekiel Emanuel of the University of Pennsylvania, machine learning not only will improve diagnostic accuracy but also displace much of the work of radiologists and anatomical pathologists. Robert Schier, the medical director from RadNet, recently clarified the vastly different opinions about the future of AI in a JACR publication: there is the apocalyptic claim that AI will make all radiologists extinct, as well as the delusional thought that AI and computers will merely assist—and never replace—radiologists [2]. Schier concludes that both extremes are mistaken, but the truth is in the direction of the first. Successes have been achieved with convolutional neural networks (CCNs) able to analyze skin lesions from photos as well as dermatologists, and able to detect diabetic retinopathy from color fundus images as well as ophthalmologists [6, 7]. In radiology, such CNNs are doing well in various tasks such as detecting bone fractures or pneumonias on plain films, quantifying white matter lesions in the brain, and analyzing interstitial pulmonary changes [2, 6, 8, 9].

Some algorithms are able to do certain tasks better than the average radiologist. Although this has only been tested in a research setting, the best systems are performing at a level similar to humans. Publication of preprint articles about successful DL algorithms in open-access journals (e.g., arXiv.org) intended to exchange information about technical developments (typically in computer science and engineering) and thus not meeting the classic requirements of a peer-reviewed radiological journal has led to the speculation that one day radiologists might be replaced by "robots" [7, 10, 11]. Many of the recent publications in lay media and even scientific journals portray computers and AI as a potential threat for radiologists, able to take over their jobs in the long run (15–20 years) [1, 10, 12]. The first FDA-approved computer vision algorithm, in April 2018, that can be utilized for medical diagnosis (to detect certain diabetes-related eye problems) without the input of a human clinician, is considered by many as an important step towards a future in which certain specialties such as radiology and pathology may be radically reshaped or cease to exist [4, 7]. This algorithm will be used in the first device that provides a screening decision without the need for a clinician specialist to also interpret the image or results, which makes it usable by healthcare providers who may not normally be involved in eye care.

Due to these recent events, statements, and publications, radiology trainees find themselves in a vulnerable position and some may have doubts on whether they should have pursued diagnostic radiology as a career if they had known of the potential impact artificial intelligence is predicted to have on the specialty [13]. Is the future for radiology in danger? Are the real risks for radiologists as bad as predicted in these media? Should residents or young doctors be afraid of choosing a career in radiology? In this chapter we try to provide an answer to those questions, as well as suggestions on how to deal with these exciting changes.

20.1.2 Radiologists' Perspective

The successful algorithms referred to in the publications mentioned earlier are mostly based on the analysis of diagnostic images. Indeed, so far the most successful DL algorithms that were developed are based upon the principle of image recognition, which is a task that can be divided into three main categories: *detection* (absence or presence of abnormalities or pathology in an image), *classification* (prediction of properties of regions of an image, such as malignant vs. benign), and *segmentation* (isolate organs or structures, enabling volume measurements or calculation of other properties) [10, 14, 15]. These are all relatively simple single tasks and thus examples of narrow AI, which means that the software only works within a very limited context and is not able to take on tasks beyond image recognition. Narrow AI is the only form of AI that humanity has achieved to develop so far within radiology [16]. Since radiologists are primarily known for their image interpretation skills, which for some types of basic examinations could be considered as a narrow and well-defined task, the early studies and publications about AI algorithms outperforming the radiologists have amplified the misconception about the imminent disappearance of the radiology profession.

Although the interpretation of images plays a central role in the radiologist's workflow, the work also goes beyond image interpretation, including consultations with other physicians about diagnosis and treatment, performance of image-guided minimally invasive procedures, selection of the most appropriate imaging method, integration of image findings with data from the electronic health record (EHR), discussion of findings with the referring physician and patient, quality control, and education. The complexity of such tasks, which require superior clinical or managerial expertise, background information, and different forms of intelligence, goes beyond simply making a statistical analysis and prediction based upon pixels, which is the essence of what most DL algorithms based upon CNNs are currently made for. AI will not carry out those tasks in the short term. This is also one of the reasons why nonradiology specialties (e.g., cardiologists) feel much less threatened by the potential of AI to displace them: their direct patient interaction will be very difficult to replace. AI systems could add more value once efforts focus on tasks that are challenging for radiologists and located beyond the domain of image recognition, e.g., for facilitating the integration of quantitative image findings with EHR data, optimization of patient scheduling, reduction of radiation dose or scanning time, etc. If one day AI would take over the radiologist's image reading function, most radiologists would probably use the extra time to focus on other essential activities, including more interactions with patients and referring doctors. But it's not unthinkable that such additional time might not even become available, since radiologists primarily have to cope with an ever-increasing workload and data load caused by the progressively rising volume in medical imaging procedures (Fig. 20.1). In addition, thousands of algorithms still have to be developed for narrow detection tasks to identify all possible abnormalities and diseases in medical images, whereas at this moment only a few can be done by AI [11].

On the other hand, in parallel to the progress in medical DL, there is a growing interest in the potential risks and downsides of using DL-based software applications in clinical practice. Whereas it may seem attractive from an economic point of view to take the expensive human doctor or radiologist "out of the loop" and replace him or her with a cheap but accurate DL algorithm, the potential damage that might be caused by malignant cyber attacks needs to be considered as well. The so-called adversarial examples have become a very popular area of research in the machine learning (ML) community because they give a better insight into the possible limitations of current DL methods and the vulnerability of such algorithms to cyber attacks. They are based upon inputs engineered to cause misclassification. A clinical system that leverages a ML algorithm for diagnosis, decision-making, or reimbursement could be manipulated or hacked with adversarial examples, even when keeping a human in the loop. Healthcare seems to be very vulnerable to such attacks, and there are many incentives for prospective bad actors to

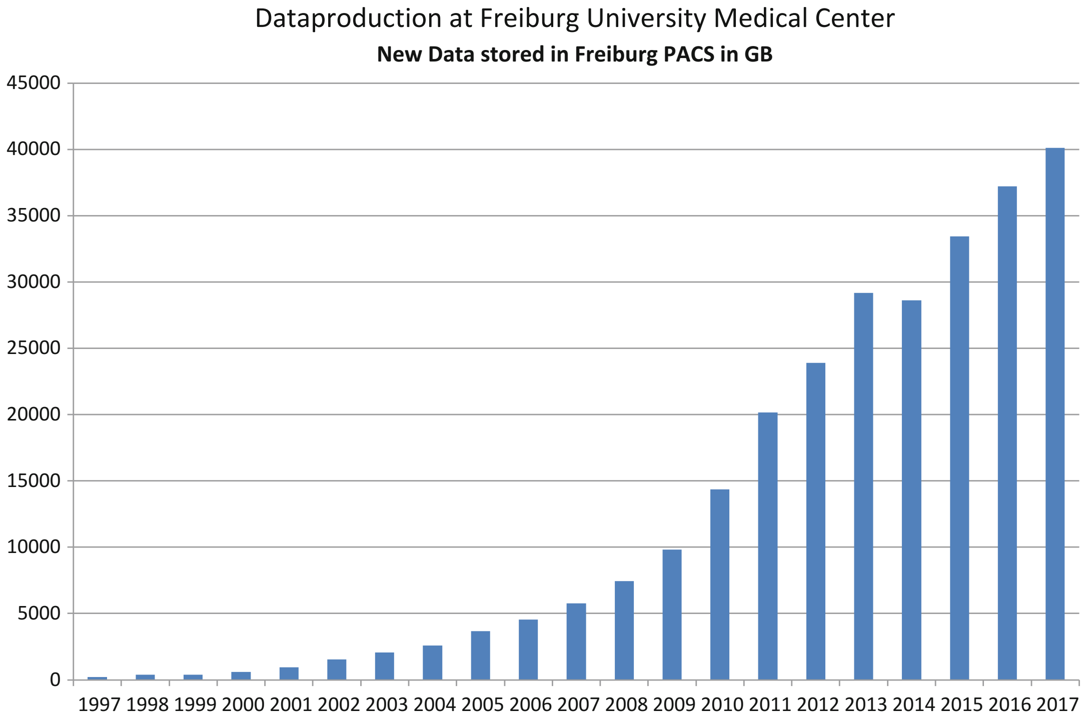


Fig. 20.1 Progressive increase in the amount of radiological data stored in the PACS of the University Medical Center in Freiburg, Germany (GB: gigabyte)

implement them [7]. Some risks and dangers related to using AI applications in medical imaging will be addressed in more detail in the section *Hidden risks and dangers*.

20.2 Level of Expectation for AI in Radiology

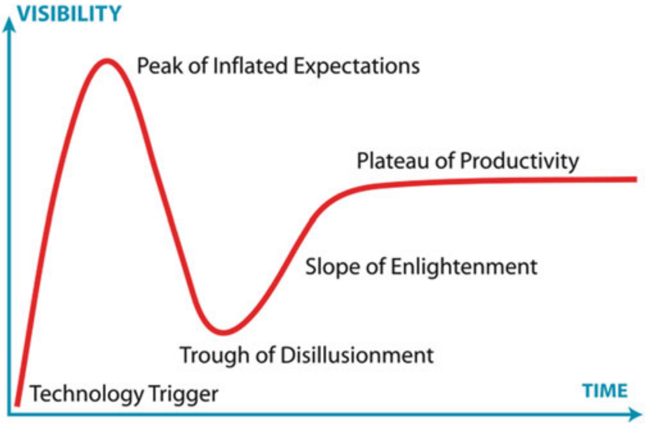
Advances in the field of computer vision, which is also relevant for medical imaging, have aroused the interest of technology giants, venture capitalists, and governments [17]. In 2017 healthcare was ranked as the most important area of AI start-up investments, because healthcare has innumerable opportunities to leverage AI in its pursuit of more accurate, proactive, and comprehensive patient care [18]. With medical imaging being one of the fastest moving areas with technological changes in healthcare, the expectations for radiology are very high and have escalated into a real hype. The real size of the hype became visible at the American and European radiological societies' annual meetings, the RSNA 2017 in Chicago, and the ECR 2018 in Vienna, where there was a massive interest for all AI-related topics [19, 20]. A crucial question that concerns many radiologists on a global level is what expectation level is realistic and on what terms, i.e., how severe will the impact of the new technology be on the radiological profession, and on what terms? The Gartner hype circle is a well-known representation of the evolution of expectations for new technologies (Fig. 20.2). Similar examples of such "hype cycles" can be seen with the introduction of 3D printing in radiology a few years ago and also the self-driving cars.

The "peak of inflated expectations" is usually followed by a "trough of disillusionment," which occurs when the initial expectations are not rapidly met. A "slope of enlightenment," a period during which real progress is being made, is usually followed by a "plateau of productivity" in which the technology is adopted on a more general level. When the level of expectation is set too high, disappointment will inevitably follow, even if the fundamental technology is sound. When setting a level of expectation it is essential to make a distinction between applications that will complement radiologists in fulfilling repetitive, time-consuming narrow tasks such as lung nodule detection, and those that will become supplementary and perform functions that radiologists cannot perform themselves, such as making a treatment proposal or survival prognosis of the patient by integrating the imaging findings with all other data available in the EHR.

20.2.1 AI Will Complement Many Routine Radiology Tasks

It's realistic to assume that very soon AI applications that currently fit under the category of narrow AI will be at least as good as radiologists to deal with simple and well-defined tasks. The automated analysis of chest radiographs could be

Fig. 20.2 The Gartner hype circle representing the evolution of expectations for new technologies



used to perform a triage of patients with abnormal findings and to prioritize readings of those examinations. Detecting, measuring, and characterizing a lung nodule on computed tomography (CT) images is another example of such a well-defined task, the management of which is standardized following the Lung-RADS$^{\mathrm{TM}}$ scoring system [21]. New ML-based applications offer the possibility to detect abnormalities more quickly in emergency situations such as stroke. In patients undergoing a non-contrast-enhanced emergency brain CT to exclude a cerebral infarction, it is critical to detect the ischemic changes in the brain as quickly as possible for guiding the stroke management. The e-ASPECTS algorithm, which has received the European CE Mark, has been demonstrated to be non-inferior to neuroradiologists in assigning the Alberta Stroke Program Early Computed Tomography Score (ASPECTS) to those brain scans [22]. The Viz.ai LVO Stroke Platform is an FDA-cleared algorithm that has been demonstrated to perform well in automated detection of proximal intracranial large vessel occlusions (LVO) from the CT angiography (CTA) images of the cerebral vessels [23]. It immediately notifies the on-call stroke physician about the findings using a secured messaging service, hereby speeding up the decision process regarding the need of performing a thrombectomy and (if necessary) transferring the patient to a specialized treatment center. Another very promising although more complicated clinical use case seems to be the automated calculation of fractional flow reserve (FFR) based upon cardiac CT angiography (CTA) images (FFR-CT). CTA of the coronary vessels is a well-accepted technique for estimating the percentage of obstruction of coronary vessels, but often in the cath lab significant blockages seen on CTA do not greatly impact blood flow according to FFR measurements. A super-computing fluid dynamics algorithm can be used to determine the virtual hemodynamic significance of lesions and thus may offer the ability for CT in the emergency department to be the gatekeeper to the cath lab [24]. This type of service could therefore be usedTM in the benefit of the patient by avoiding an unnecessary interventional procedure, in combination with a potential significant reduction of costs. More information about this subject could be found in the chapter about Cardiovascular Diseases.

Therefore the first step of a useful and successful implementation of AI applications should be the identification of relevant clinical use cases by radiologists in their practice, mainly with the intention of improving the workflow efficiency and improving the efficiency and value of care. As this workflow consists of several other tasks besides image analysis, such use cases could also be found in non-imaging-related radiological tasks such as optimized patient scheduling, including the reduction of patient "no-shows." The topic of non-imaging AI applications is addressed in more detail in the chapter about applications of AI beyond image interpretation.

After identification of the appropriate use cases, a clear definition of the goal of the application should be made for instructing the software developers, preferably following a standardized method or template. The Digital Science Institute of the American College of Radiology (ACR DSI) is currently specifying detailed use cases for AI algorithms that will provide not only a robust narrative description for what a specific algorithm needs to accomplish, but also the mechanisms for training, testing, validation, and monitoring algorithms in clinical practice. They are also promoting a model of sharing approved use case templates that have been submitted by the radiological community [25].

The introduction of AI applications for purposes such as taking over simple or tedious tasks, reducing the radiological workload, facilitating the analysis of emergency examinations, and optimizing the patient scheduling is probably the best way to smoothly introduce AI-based technology in radiology practice. Within a time span of 3–5 years a substantial number of use cases will be available for implementation, and thus in a certain sense will cause a "displacement" rather than a "replacement" of radiologists. In addition, thanks to AI applications developed for simple tasks, it will be easier to provide basic medical care and diagnostic services to patients living in places deprived of physicians or radiologists, such as in remote areas and in developing countries. A good example of an existing type of such clinical use case is the CAD4TB (Computer-Aided Detection for Tuberculosis) software that is being used for screening of tuberculosis (TB) on chest X-rays (CXR) in areas with a lack of skilled readers, usually in less developed countries, such as Ghana. TB suspects undergo both symptom screening and chest radiography, and those with symptoms and/or abnormalities in the chest X-ray undergo further testing. A large retrospective evaluation of the software on a database of 38,961 CXRs with 87 TB cases showed that the software had a negative predictive value of 99.98% with an area under the curve (AUC) of 0.90. The authors concluded that CAD can be used to identify a large proportion of normal CXRs at high sensitivity, and therefore could be used as a cost-effective instrument of triage for radiographic TB screening [26].

20.2.2 Will AI Also Surpass Existing Radiology Tasks?

The next and more difficult, for some even threatening question, is if AI will surpass radiologists in performing tasks they cannot do themselves, and if yes, when it will be able to do this. In our opinion the answer should be approached from two different angles, namely the *data perspective* and the *learning perspective*.

First of all, the clinical practice of radiology involves synthesizing of disparate data sources, such as image findings, lab results, patient history, and clinical findings, a task that is not readily amenable to complete automation. It's highly likely that, as soon as image information can be integrated seamlessly with the data from the EHR and other e-health data sources, it will be possible to approach the image findings from a more holistic perspective and to gradually shift to a more personalized approach of the patient's disease. A new concept in healthcare is the so-called digital twin, which originates from engineering and represents the centralized availability and management of all health-related digital information of a single patient. In the context of healthcare the digital twin is considered as a lifelong digital replica of an individual, comprising all information of the physiological status and lifestyle of that person. Automated updating of the person's status with data from any health-monitoring tool and new medical exam would allow a more dynamic and personalized management of that individual's health condition [27]. The digital twin model thus implies a data-driven approach based upon a centralized availability of all patient-related data. For the development of ML-based applications able to learn from all this information and capable of surpassing the skills of the radiologist a seamless interoperability between all institutional hardware and software solutions is a key component, which can only be achieved through a productive collaboration of all involved stakeholders [10]. For development of such highly advanced and more complex types of AI applications, patient data will have to be made available outside hospitals while respecting all legal and ethical considerations regarding patient data privacy. Although currently partnerships are already being made between vendors, pharmaceutical industry, start-ups, and academic hospitals, it will take many years, even several decades, before enough high-quality and readily exchangeable vetted data for training of AI solutions will be available. Solid and internationally accepted regulations will have to be implemented, not only for the exchange and usage of such data but also regarding the validation and autonomy of the newly developed applications. For this purpose a radiology informatics ecosystem needs to be established [16].

Secondly, whereas currently most DL algorithms for medical imaging are now based upon a supervised learning technique, it is to be expected that when unsupervised learning is applied on a larger scale, novel image patterns and relationships nobody was even aware of will be detected. It will probably even be possible to generate features directly from raw data, which means that radiology images from CT and MRI might become superfluous or unnecessary for developing such algorithms [14]. As soon as such algorithms are proven to be more accurate and faster than radiologists, it might even be questioned if radiologists will have to continue reading images and generating reports. The unsupervised DL models however are often considered to be "black boxes" in which humans can only interact by checking the results at the end of the pipeline, since the algorithms are applied on the raw data on which the learning process is fully automated. They have no explicit declarative knowledge representation, which makes it difficult to generate the required explanatory structures [28]. Even with an understanding of the mathematical theories behind the machine model it is complicated to get a real insight into the internal working of the unsupervised learning model; hence such black box models are raising the question if we can trust them. The high relevance of a high level of trust for acceptance of ML-based applications for clinical use by radiologists was also expressed in a multidisciplinary panel discussion that took place at the MIDL 2018 meeting (Medical Imaging with Deep Learning), in which the panel consisted of a mix of representatives from the industry and academy [29]. To create a substantial level of trust and confidence between radiologists and AI applications, humans and AI applications should work together in a human-cybernetic partnership. Radiologists aren't only indispensable in the identification and definition of the best clinical use cases for which the AI applications are developed, but also in the verification, approval, and validation of algorithms. Such human-cybernetic harmony could for example be achieved by aiming at a model of interactive machine learning (iML), as proclaimed by Andreas Holzinger from the HCI-KDD (Human-Computer Interaction & Knowledge Discovery/Data Mining) research group. The term iML is defined as *"algorithms that can interact with agents and optimize their learning behavior through these interactions, where the agents can also be humans"* [28]. iML approaches are based upon integration of a human-into-the-loop, thereby making use of human cognitive abilities. Because of the quirks in which DL-based algorithms can process data, it is to be expected that they will make errors that are readily apparent to a human observer. Integration of a human agent is useful in making algorithms transparent and explainable.

There is also a legal aspect (and in the United States a malpractice aspect) related to this issue and, in particular, the question of who is liable for the final patient outcome, the one who delivered the training set, the person that developed the AI, the company that sold the AI algorithm, or the radiologist. By following the radiologist-in-the-loop principle and integrating human expertise and long-term experience in the learning phase, the complexity of developing ML-based algorithms could be significantly reduced (Fig. 20.3). Such glass-box approaches foster transparency and trust

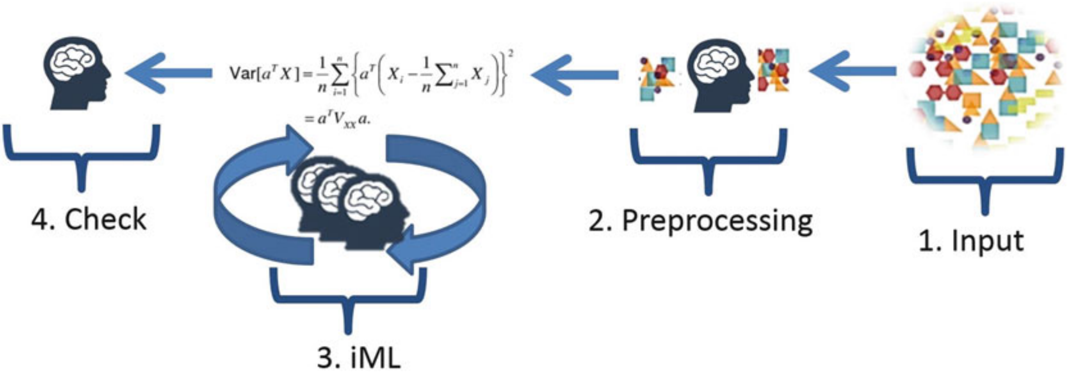

4. Check
$\mathrm{Var}[a^T X] = \frac{1}{n}\sum_{i=1}^{n}\left\{a^T\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)\right\}^2$
$= a^T V_{XX} a.$
3. iML
2. Preprocessing
1. Input

Fig. 20.3 The interactive machine learning (iML) model is based on the integration of a human-into-the-loop for the development of ML algorithms, hereby incorporating human cognitive abilities to make the learning process explainable. The data that are used for training (1) are preprocessed by humans (2). The human is seen as an agent involved in the actual learning phase (3). The human factor is also included for checking the results (4) (figure used with permission of A. Holzinger)

in ML by making the algorithms interpretable and explainable, which is mandatory in the context of the existing legal issues on privacy, data protection, safety, and security [28].

Predicting the time frames applicable to the development of AI applications that will surpass radiologists is very difficult and depends on many factors. As we tried to explain in this chapter, and as was highlighted in several other chapters of this book, there are still many hurdles to be taken, which are not only located in the development of software, but also in the creation of the appropriate infrastructure for exchanging patient data and acceptance of global standards for the development and validation of algorithms.

20.3 Strategies to Prepare for the Future

As clearly stated earlier we have to be careful not to have unrealistic expectations about AI, nor fear AI. We have to accept the fact that in the coming years AI will play a serious role in the practice of medicine and radiology. The development of AI is still in an early stage, offering radiologists the opportunity to get accustomed, adapt the workflow, and change the workplace culture. Undoubtedly there are many potential ways to leverage AI for improving image interpretation and optimizing many other facets of the daily radiological workflow. AI can improve the performance of radiologists, and both radiologists and AI working together will be better than either alone [10]. But what strategy should software developers, radiologists, and other stakeholders in the hospital, including IT directors and hospital managers, follow to prepare for the future and to smoothly introduce AI applications in radiology practice on a wider scale? And what challenges are lying ahead for DL-based systems to make them attractive for regular usage by radiologists?

20.3.1 Multitask Learning

In ML the development of algorithms is mainly focusing on a particular (narrow) problem and on optimization of the outcome of the algorithm. Usually a single model or an ensemble of models is trained to perform one desired task. Multitask learning (MTL) is a learning methodology that estimates models for several tasks in a joint manner. Information coming from the training signals of related tasks is used, which enables a model to perform better on the original task [30]. Moeskops et al. used MTL for training a single convolutional network (CNN) architecture for different medical image segmentation tasks in different modalities, visualizing different anatomical structures [31]. Such an “all-in-one” algorithm would be able to perform multiple tasks in different modalities without problem-specific design of the architecture, i.e., the network would be able to recognize the modality, the anatomy visualized in the image, and the tissues of interest. Secondly they used that single trained CNN for segmentation of six different tissues in brain MRI, the pectoral muscle in breast MRI, and the coronary arteries in cardiac CT angiography (CTA). The results showed that one single system can be used in clinical practice to automatically perform diverse segmentation tasks without task-specific training. Including multiple tasks in the training procedure resulted in a segmentation performance equivalent to that of a CNN trained specifically for the task. The more tasks one single algorithm is able to perform, the more useful and applicable it will be for clinical practice, and the more radiologists will be eager to use it as a complementary tool. Developers should focus on the creation of algorithms able to perform multiple tasks applicable to a range of modalities, which will result in a wider variety of clinical use cases for the same system, and thus a more flexible integration in the radiological workflow.

20.3.2 Swiss Knife for Radiologists

As already mentioned the ML-based algorithms should be fully integrated in the PACS workstations, with the purpose of making the radiologist's workflow run very smoothly and to reduce the number of clicks as much as possible.

The smooth integration of DL-based applications in the PACS interface, attainable through a minimum of clicks, is a prerequisite to motivate radiologists to make use of such applications routinely. Currently most ML-based algorithms are not well integrated in PACS workstations yet. Often a separate workstation or network node is required for sending images out for analysis. Ideally the processing of data should take place in background so that the reading and reporting are not delayed, and the data is offered as soon as the examination is opened. The "availability in one hand" of such algorithms through the workstation could be seen as a Swiss knife for radiologists. The more easy the access is made to the tools, the more often the radiologists will use them and the more data will be available for improving them, e.g., by reinforcement learning or iML, and the smarter the applications will become. One could question however if by doing so, these AI applications, although initially designed to complement the radiologist's work, would then be trained by radiologists themselves until they are ready to replace them. We think that this prediction should be put into the perspective of the continuously growing demand for medical imaging in combination with the ever-increasing workload and the broadening of the spectrum of activities for radiologists. The amount of data to be used by radiologists will become unmanageable and therefore the need for accurate and reliable software applications, ideally developed in a human-cybernetic type of collaboration, becomes indispensable.

20.3.3 Integration of Existing Medical Information Databases

According to Dr. Paul Chang, radiologists first have to build the necessary infrastructure before they can start using ML-based applications. This means that a platform allowing a seamless interaction between all connected data systems in the hospital is needed. Such infrastructure can be based on cloud technology and should be able to handle big data, facilitating the smooth exchange of data between the PACS and EHR. The goal of this strategy is to make a gradual transition to a data-driven "human-cybernetic collaboration," with the ultimate purpose of improving patient care [32]. As was mentioned earlier, a seamless interoperability between all institutional hardware and software solutions is a key component in the implementation of practical AI solutions. Unfortunately the IT architecture of most hospitals is still PACS or even EHR centric, so hospitals will have to start thinking beyond this consolidated centric approach. Paul Chang advocates switching to a service-oriented architecture (SOA) as a solution to escape from the existing data siloes [20]. The main principles of the SOA technology are based upon the integration of distributed separately maintained and deployed software components in a network. It is a mixture of loosely coupled services that are able to interact in a disciplined manner. A central bus or "spinal cord" is created in which information from sources throughout the enterprise is distributed in a well-orchestrated manner [33, 34]. Thanks to such SOA a company like Amazon is able to offer its customers an integrated set of services in a fast and easy-to-use Web-based application, whereas Amazon's backend interacts with dozens of different databases. By implementing an SOA in a hospital environment, as is for example the case at the University of Chicago, information can be extracted from sources throughout the enterprise and seamlessly exchanged in a quick and simple manner [35]. The radiologist reading images in the PACS is able to retrieve information from the EHR with the click of a button, i.e., without having to leave the PACS or to log in in a separate system. It can be questioned however if deploying an SOA model is achievable for large healthcare enterprises, since it's quite challenging from a governance perspective and requiring a significant cultural change in the hospital [33]. Nevertheless this type of change management will be necessary in healthcare, allowing a more profound integration and interoperability of enterprise-based data sources, facilitating the optimal use and training of DL-based applications.

20.3.4 Blockchain Technology

The current techniques for transferring medical imaging data are inconvenient and occasionally wholly inadequate. Despite the widespread availability of digital imaging and high-speed network connectivity, a physical copy (e.g., a CD or DVD) often still needs to be couriered between providers. The existing PACS- or EHR-centric infrastructure of most hospitals makes the cross-site imaging transfer, either digital or physical, often depending on trusted third-party intermediaries. The blockchain concept of "decentralized image sharing" is offering the possibility to develop a framework for cross-domain sharing of medical images and other patient data. Patients are able to delegate electronic access to their medical imaging in a secure manner, and third-party access to protected health information is eliminated [36, 37]. In healthcare organizations the interest for the blockchain architecture is growing slowly but steadily. The FDA and IBM Watson already established a partnership, with the intention to use blockchain technology for enabling a secure, efficient, and scalable exchange of health data coming from several sources, such as EHRs, clinical trials, genomic databanks, mobile, and wearable devices [37]. A blockchain consists of a distributed tamper-proof database that is shared by multiple parties and in which all records are securely stored. It is maintained by a set of "nodes," entities without a preexisting trust relationship that are connected through a peer-to-peer network [36] (Fig. 20.4). Records can only be added to the database, never removed, and each record contains a timestamp and secure links to the previous record. New records can only be added based upon synchronous agreement or "distributed consensus" of the parties maintaining the database [37]. Blockchains thus

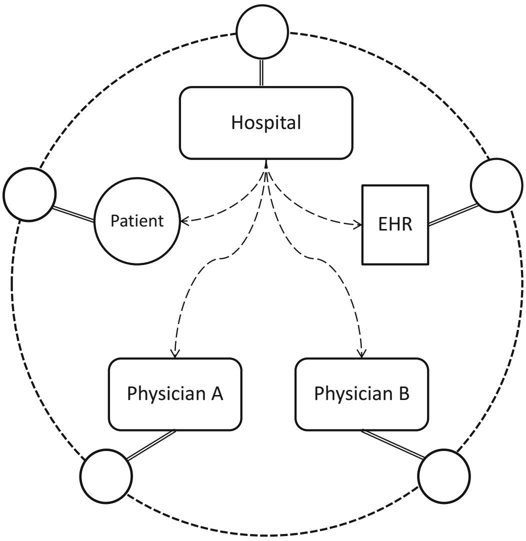


Fig. 20.4 The image sharing blockchain. Each participant operates a node (o) on the network, which establishes a blockchain (Dashed line of outer circle). The patient provides access to chosen entities by posting blockchain transactions. Imaging data are transferred directly from the source to these authorized recipients; no central intermediary is required (figure from [36] used with permission)

enable many separate parties to converge upon a single, immutable record without requiring an authoritative intermediary (Fig. 20.4).

An AI service provider specialized in developing DL algorithms could for example represent a central hub of the blockchain, functioning as "the brain." Associated hospitals would be allowed to run the provider's AI algorithms over their data, e.g., medical images. When the final diagnosis or treatment is funneled back into the EHR, the data are published back to the blockchain. With these data the AI provider will be able to refine the accuracy of its algorithms. In such a blockchain model intelligent algorithms will be able to mine enormous amounts of structured and unstructured data from numerous sources, and provide scientific insight and business intelligence to the members of the blockchain (Fig. 20.5). Ultimately, the large-scale feasibility of such an approach remains to be demonstrated. In healthcare the blockchain architecture is a nascent technology and a thorough discussion of its anticipated benefits and limitations is thus warranted but is beyond the scope of this chapter.

20.4 Hidden Risks and Dangers

Knowing that the development of AI products and certainly their routine use in radiological practice is still in its infancy, we also should be aware of the fact that there are still many known and unknown risks or traps connected to the use of ML-based solutions. Although the currently existing weaknesses and problems with early implementations of algorithms for radiology will progressively become more visible, they potentially might remain hidden for many future

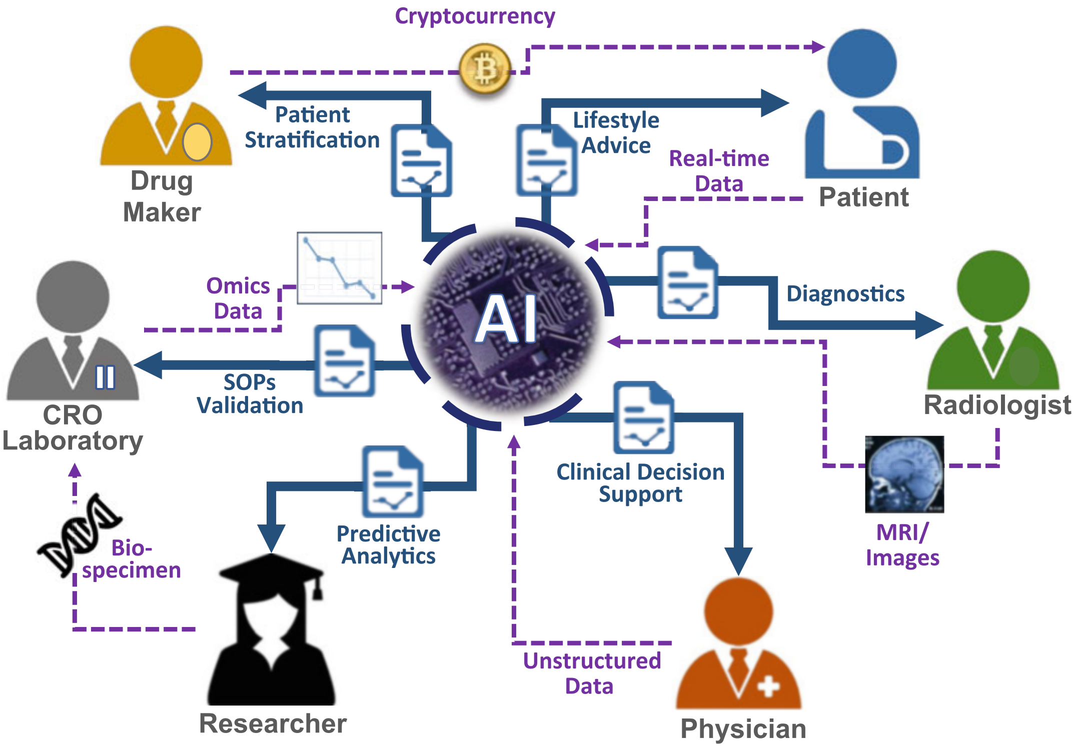


Fig. 20.5 An AI service provider represents the central hub of the blockchain. Radiologists and other clinical users run the provider's AI algorithms. The patient's data are published back to the blockchain. With these data the AI provider is able to refine the accuracy of its algorithms. In such a blockchain model intelligent algorithms are able to mine enormous amounts of structured and unstructured data from numerous sources, and provide scientific insight and business intelligence to the members of the blockchain (figure used with permission of A. Holzinger)

users because of their lack of basic knowledge in this technique. A basic knowledge of computer science, statistics, and deep learning should be regarded as a prerequisite to understand and foresee what can go wrong and how it can be avoided or improved [10]. In essence this can be compared with the basic medical physics knowledge that every radiologist working with MRI machines and interpreting MRI studies needs. As the physical principles of MRI are integrated in every radiologist's training, the basic principles of ML should also be part of every state-of-the-art training curriculum. The Radiological Society of North America for example organizes an online National Imaging Informatics Course (NIIC) in joint collaboration with the Society of Imaging Informatics in Medicine (SIIM). The European Society of Radiologists (ESR) recently incorporated a training in medical imaging informatics in the latest version of the European Training Curriculum (ETC), which can be found on the ESR website (https://www.myesr.org/education/training-curricula). The imaging informatics criteria were developed in collaboration with the European Society of Medical Imaging Informatics (EuSoMII), which is a subspecialty society affiliated to the ESR.

In the previous part of this chapter several technical bottlenecks were already addressed, such as the availability and exchangeability of data, and the need for both a solid technical infrastructure and an internationally accepted framework or ecosystem. Besides this technical "shadow side" of DL there are other aspects connected to usage and integration of DL-based software applications that need further attention, such as the quality and readiness of data, and the legal and ethical issues. These topics will be briefly discussed in the following text.

20.4.1 Quality and Validation of Data

For most readers of this book it will be obvious that there is still an ongoing quest for usable, i.e., curated or vetted data. The importance of such data was explained in more detail in several earlier chapters. When made available, the enormous amount of data that is needed for training and testing of DL algorithms has to be organized and prepared, which is a very labor-intensive and time-consuming task to begin with, requiring a lot of human labor and computer power. If the data are not managed properly it will be impossible to retrieve or discover the right data. Furthermore, not only the volume but also the technical quality of data is primordial. Motion artifacts, image noise, beam hardening, partial voluming, and other are all examples of imperfections that need to be avoided in training datasets. There is a huge technical diversity in datasets depending on the type and brand of modality, and scanning protocols that have been used. Therefore robust methods are required for controlling the quality and completeness of training data in order to feed algorithms with high-quality information instead of "garbage." For example when developing an algorithm for detecting hepatocellular carcinoma, the algorithm will underperform if not *all* contrast phases on CT or MRI are included [38]. The supervised learning process of algorithms is based upon the availability of a so-called *gold standard*. The utility of the algorithm is critically dependent on the quality of this gold standard, which is still questionable since there is no absolute standard for obtaining such gold standard (also called "ground truth" in the computer scientists' literature). For some it can be based upon the consensus of an expert panel of radiologist; for others it can be the interpretation of a single first-year resident. The cost of a physician annotation for a sufficient amount of high-quality gold standard data can be quite high for a start-up company and thus be a major impediment to the development of algorithms [38]. The importance of the availability of curated high-quality data is explained in more detail in several other chapters of this book.

20.4.2 Data Security and Privacy

The security and privacy issues related to the management and exchange of patient-related data certainly should not be discarded. Protection of patient data should be safeguarded at the highest level, since smart algorithms are able to mine databases and abuse them for malicious purposes or financial profit. Earlier in this chapter we already mentioned the *hacking of algorithms* by adding corrupted or misleading image datasets, which is a significant risk, mainly when using DL algorithms for trial purposes [7].

On May 25th, 2018, the new General Data Protection Regulation (GDPR) became in force throughout the European Union (EU). An informative paper explaining what the radiologist should know about this new regulation was published recently by the European Society of Radiology [39]. Following this regulation, consumers of all kinds, including patients, must give their explicit consent for use of their personal data—and can withdraw it at any time. The explicit consent given by the patient for participating in a single trial will now be insufficient for sharing the same data for analysis by others, thus necessitating a new consent [40]. This also implies that for European patients, the vetting of data has to proceed according to the GDPR. The key concept of the law is "privacy by design," which means that the consumers own their data and have the power to make corrections. Giving the patient ownership and control over his or her data also means that, for example, a hospital cannot automatically share data with an AI provider. As mentioned earlier, blockchain and cloud technology have made it possible to seriously think about a patient data record under personal control, where patients can see their health-related information, share it with the parties they approve, and keep track of how others have used their data (Figs. 20.4 and 20.5). Although with the GDPR it has become more complicated to construct, maintain, and manage large healthcare and image databases, the ongoing technological evolution and associated trends may lead to a shift of viewing patients as collaborators provided with the tools to manage their own health and data [40]. Radiological and other medical societies will have to take up their responsibility in drafting the codes of conduct in how to manage, share, and use patient data with the intention to promote the rapid AI developments without harming the patient and to improve the overall level of care.

In the United States the Food and Drug Administration (FDA) plays an important role in approval and integration of machine learning in the clinical setting. Before clinical use, ML applications have to submit specific information about the algorithm development and clinical validation. The acceptance by human experts needs to be demonstrated with clinical validation studies, for which standards still need to be established. It will be essential however for radiologists to actively participate in the creation and application of these standards. In anticipation of the new challenges involved in appropriately regulating this software, the FDA is developing several regulatory pathways for ML applications. However, several challenges are lying ahead in the prospect establishing rules for validation of ML algorithms by humans, certainly for those developed to find associations in data that are invisible to the human eye [16]. Not only the FDA but also the scientific and radiological societies should be proactive and provide guidelines for the creation, validation, and integration of AI applications, both for algorithms for image interpretation and non-image interpretation tasks. Other relevant FDA programs and applicable US regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, are addressed in more detail in the chapter about the role of an AI ecosystem in radiology.

20.4.3 Ethics and AI

For the next generation of physicians and radiologists familiarity with ML tools will be a fundamental requirement, and will also require awareness of and attention to the ethical challenges inherent in implementing ML in healthcare [41]. AI is being developed to help augment doctors' decision-making, but obviously there is concern about who will take responsibility when those decisions are wrong. What if the machine misses a diagnosis, the doctor accepts the judgment and the patient dies? Will machine-heads roll similarly to those of medical doctors when they make life-threatening errors or injure the patient? The progressive transition to automation can have far-reaching consequences for both patients and physicians, certainly when autonomous AI-based systems are being embedded in the decision-making related to life and death, such as cancer diagnosis. The rules and principles followed by autonomous and intelligent systems regarding the appropriate treatment of patients should not be less than those applicable for radiologists. The ethical issues related to the usage of ML-based systems can be categorized in three main categories [42]:

- Data ethics
- Algorithms ethics
- Practice ethics

In the "safety and privacy" section we already addressed the importance of protecting the *patient data* according to the existing directives and legislation, applicable in the country of each patient. Privacy and confidentiality are closely related issues but the terms should not be used interchangeably. Whereas data privacy refers to the rights of individuals to maintain control over their health data and information, confidentiality refers to the responsibility of those entrusted with those data to maintain privacy [43]. There are also serious concerns that algorithms may mirror human biases in decision-making [41]. The patient data used for training algorithms may include a bias against group-level subsets of individuals, such as specific ethnic or economic groups [44]. Besides the existing geographic variations between the patient populations, there are also many variations related to race, gender, socioeconomic background, body habitus, and prevalence of disease. Certainly the higher level type of artificial intelligence algorithms, made for predicting outcomes of certain diseases or treatments by combining data from radiomics and genomics (radiogenomics), will be biased if there have been few genetic studies in certain populations [41]. Both developers and consumers of AI applications in healthcare, and diagnostic imaging in particular, should be aware of this diversity in patient populations and ensure that a representative variation is provided in their training database so that algorithms will be free of unintended bias. On the other hand, algorithms could be built to compensate for known biases or identify areas of needed research [41]. The risk for a potential bias related to the representativeness of data is actually also part of the quality aspect of data, but then rather from a patient-related instead of a technical point of view. Again, this underlines the importance of controlling the quality and completeness of training data in order to feed algorithms with high-quality information instead of "garbage." Ideally, efforts should be made to ensure that formatted datasets for purposes of training and testing the algorithms are made available through a centralized vendor neutral platform managed by authorized organizations, such as the national radiological societies, which could also guarantee the representativeness and quality of the data. It will be necessary to find a balance between protecting the patient's privacy and sustain the potential of developing intelligent machines. The American College of Radiology Data Science Institute (ACR DSI) is already taking initiatives in this direction by fostering a publicly accessible Data Sets Directory, which is part of the ACR DSI framework for an AI ecosystem [25].

The second ethical issue concerns *the transparency of algorithms*. For each algorithm it should be explainable in what direction it progresses and why it chooses that path. The physician will be ultimately making a decision based upon on two elements: the regulatory approval and the standards of care [45]. The regulatory approval needs to come from the FDA or analogue organization (e.g., the CE mark in Europe) that needs validated and up-to-date tools and procedures to verify the product's properties and intentions. The software also needs to be permitted or required by the latest consensus of the professional societies including their subspecialty sections, responsible for issuing guidelines on the practice of medicine and radiology. As soon as more general intelligence tools will become available, capable of making decisions on a fully autonomous level, based upon a judgment that might supersede human perception, the question should be asked if these robots should be able to obtain an "electronic personality." A debate on this subject was held in European Parliament in February 2017, where the Parliament's legal affairs committee proposed a resolution for granting self-learning robots a form of "electronic personality" [45]. Such a status could allow robots to be insured individually and held liable for damages, which—according to the proponents—would set them on par with corporations, which also have statuses as "legal persons." The opposition, supported by a letter from an international group of AI experts, arguments that seeking a legal status of personality for robots merely is a cunning way of manufacturers to get out of their responsibility for the actions of their machines [46]. Adapted legislation and guidelines will be necessary soon, and it's obvious that radiological societies should timely proclaim their professional opinion about these issues in order to ensure that the right decisions are made at both the political and legal level. More in-depth information on this subject can be found in the chapter on the legal identity of robots and AI.

The third category is the "practice ethics." As much as AI evolves in the coming years, it still takes people and organizations to research, design, implement, and maintain these advanced algorithms. Kohl and Geis explained that policies must be in place at a practice level that promote progress and protect the rights of every individual affected by AI. They also noted that imaging leaders should learn from the recent mistakes made by the social media platform Facebook [42, 47]. The intent behind the design of ML systems should always be transparent and verifiable.

In the creation of policies regarding the usage of AI applications, the effects of a so-called collective brain on human behavior should also be incorporated. Due to the gradual incorporation of ML-based algorithms in daily practice, a progressive shift towards a computer-based decision-making process will probably take place. A collective brain will be created, which may take on an authority that was perhaps never intended. The dominance of such collective brain could eventually cause physicians to become over-reliant on automated instructions, with a consequence that they abandon common sense. ML tools will become important actors in the diagnostic and therapeutic decision-making process, but it will be challenging to anticipate how these rapidly evolving systems may go wrong or could be abused. Nevertheless they need to be bound by the core ethical principles, and radiologists will play an essential role in determining the right thing to do now and in the future.

20.5 Take-Home Messages

- ML-based technology will have a significant impact on radiology in the automation of mundane tasks, and in providing decision support in more complicated and advanced medical image interpretations. The key concept is *decision support*, indicating that computers will augment human decision-making, making it more effective and efficient.
- Future stages of AI in radiology will be based on full integration of multiple data systems. This will give radiologists the tools to increase their value in a more holistic and personalized approach of the patient.
- AI will be able to provide highly effective and low-cost diagnostic services in underserved areas and developing countries, thus increasing the access to medical care for millions of people.
- Radiology has a leading role in the development and management of ML-based imaging technology in the definition of clinical use cases, in the provision and validation of relevant training data, in assessing the relevance of ML-based findings in clinical practice, and in investigating the meaning of newly generated image-based data.
- It is most likely that radiologists will be held accountable for the performance of the ML-based system and robustness of the diagnostic data, certainly when implementing narrow-AI types of applications.
- To cope with the changing medical imaging landscape, radiologists should include the basics of imaging informatics in the training of residents and in their educational programs.
- As we live in an age of an increasing availability, creation, and manipulation of personal data, it is likely that there will be a growing

financial incentive to provide such data to third parties. By its nature technological advancement will probably create new situations for which there are no existing laws or ethical standards. Being physicians, radiologists should never forget whom they should serve and therefore strictly adhere to their oath of "do no harm": "*Primum non nocere*."

References

1. King BF. Artificial intelligence and radiology: what will the future hold? J Am Coll Radiol. 2018; 15(3):1–3.
2. Schier R. Artificial intelligence and the practice of radiology: an alternative view. J Am Coll Radiol. 2018;15(7):1004–7.
3. Washington RA. Why scan-reading artificial intelligence is bad news for radiologists. The Economist [Internet]. 2017. Available from: https://www.economist.com/free-exchange/2017/11/29/why-scan-reading-artificial-intelligence-is-bad-news-for-radiologists
4. Jha S, Topol EJ. Adapting to artificial intelligence: radiologists and pathologists as information specialists. JAMA. 2016;316(22):2353–4.
5. Obermeyer Z, Emanuel EJ. Predicting the future – big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216–9.
6. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.
7. Finlayson SG, Chung HW, Kohane IS, Beam AL. Adversarial attacks against medical deep learning systems. 2018. arXiv:1804.05296 [cs.CR].
8. Rajpurkar P, Irvin J, Zhu K, Yang B. arXiv HMAP, 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017. arXiv:1711.05225v3 [cs.CV].
9. Chen L, Carlton Jones AL, Mair G, Patel R, Gontsarova A, Ganesalingam J, et al. Rapid automated quantification of cerebral leukoaraiosis on CT images: a multicenter validation study. Radiology [Internet]. 2018. Available from: https://pubs.rsna.org/doi/abs/10.1148/radiol.2018171567
10. Tang A, Tam R, Cadrin-Chênevert A, Guest W, Chong J, Barfett J, et al. Canadian association of radiologists white paper on artificial intelligence in radiology. Can Assoc Radiol J. 2018;69(2):120–35.
11. Davenport T, Dreyer K. AI will change radiology, but it won't replace radiologists. Harvard Business Review [Internet]. 2018. Available from: https://hbr.org/2018/03/ai-will-change-radiology-but-it-wont-replace-radiologists
12. Langs G, Röhrich S, Hofmanninger J, Prayer F, Pan J, Herold C, et al. Machine learning: from radiomics to discovery and routine. Radiologe. 2018:1–6. https://doi.org/10.1007/s00117-018-0407-3.
13. Collado-Mesa F, Alvarez E, Arheart K. The role of artificial intelligence in diagnostic radiology: a survey at a single radiology residency training program. J Am Coll Radiol. 2018; https://doi.org/10.1016/j.jacr.2017.12.021.
14. Lee J-G, Jun S, Cho Y-W, Lee H, Kim GB, Seo JB, et al. Deep learning in medical imaging: general overview. Korean J Radiol. 2017;18(4):570–84.
15. Erickson BJ, Korfiatis P, Kline TL, Akkus Z, Philbrick K, Weston AD. Deep learning in radiology: does one size fit all? J Am Coll Radiol. 2018;15(3):521–6.
16. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS, et al. Current applications and future impact of machine learning in radiology. Radiology. 2018;288(2):318–28.
17. Paiva OA, Paiva OA, Prevedello LM. The potential impact of artificial intelligence in radiology. Radiol Bras. 2017;50(5):V–VI.
18. Taub LM. 2017 AI overview [Internet]. Medium.com. 2017. Available from: https://medium.com/@lolitataub/the-ai-2017-guide-de03ae82054. Accessed 15 June 2018.
19. Walter M. Are European radiologists skeptical about AI? A report from ECR 2018 [Internet]. Radiology Business. 2018. Available from: https://www.radiologybusiness.com/topics/artificial-intelligence/are-european-radiologists-skeptical-about-ai-report-ecr-2018. Accessed 15 June 2018.
20. Massat MB. Artificial intelligence in radiology: hype or hope? [Internet]. 2018. Available from: https://www.appliedradiology.com/articles/artificial-intelligence-in-radiology-hype-or-hope. Accessed 22 July 2018.
21. ACR, 2014. LungRADS version 1.0 assessment categories release date: April 28, 2014. Available from: https://www.emmc.org/emmc/files/c3/c337137e-f68d-403e-938a-fd3c045e35c2.pdf
22. Nagel S, Sinha D, Day D, Reith W, Chapot R, Papanagiotou P, et al. e-ASPECTS software is non-inferior to neuroradiologists in applying the ASPECT score to computed tomography scans of acute ischemic stroke patients. Int J Stroke. 2017;12(6):615–22.
23. European Stroke Organisation Conference: Abstracts. Eur Stroke J. 2018;3(1_suppl):3–204.
24. Fornell D. What is new in FFR technology [Internet]. 2017. Available from: https://www.dicardiology.com/article/what-new-ffr-technology. Accessed 22 July 2018.
25. Allen B, Dreyer K. The artificial intelligence ecosystem for the radiological sciences: ideas to clinical practice. J Am Coll Radiol. 2018; https://doi.org/10.1016/j.jacr.2018.02.032.

26. Melendez J, Hogeweg L, Sánchez CI, Philipsen RHHM, Aldridge RW, Hayward AC, et al. Accuracy of an automated system for tuberculosis detection on chest radiographs in high-risk screening. Int J Tuberc Lung Dis. 2018;22(5):567–71.
27. Bruynseels K, Santoni de Sio F, van den Hoven J. Digital twins in health care: ethical implications of an emerging engineering paradigm. Front Genet. 2018;9:31.
28. Holzinger A. Project iML – proof of concept interactive machine learning. Brain Inform. 2016;3(2):119–31.
29. Ranschaert E. MIDL 2018 panel discussion about radiology and artificial intelligence [Internet]. Medium.com. 2018. Available from: https://medium.com/@erik.ranschaert/midl-2018-panel-discussion-about-radiology-and-artificial-intelligence-2bec1d252adb. Accessed 7 July 2018.
30. Ruder S. An overview of multi-task learning in deep neural networks. 2017. arXiv:1706.05098v1 [cs.LG].
31. Moeskops P, Wolterink JM, van der Velden BHM, Gilhuijs KGA, Leiner T, Viergever MA, et al. Deep learning for multi-task medical image segmentation in multiple modalities. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Cham: Springer; 2016. p. 478–86. (Lecture Notes in Computer Science; vol. 9901).
32. Bassett M. The reality of deep learning/artificial intelligence in radiology: they will redefine the specialty [Internet]. rsna.org. 2018. Available from: http://www.rsna.org/RSNANewsDetailWireframe.aspx?pageid=15319&id=23533&ekfxmen_noscript=1&ekfxmensel=falsefalsetruetruetruefalsefalse10-18.0.0.0730truefalse. Accessed 17 June 2018.
33. Pearson D. Orchestrating interoperability: one size does not fit all [Internet]. Imaging Biz. 2015. Available from: https://www.imagingbiz.com/sponsored/1174/topics/technology-management/orchestrating-interoperability-one-size-does-not-fit. Accessed 17 June 2018.
34. Chang P. Preparing your radiology practice and IT department for big data. Maarssen, The Netherlands. 2018. Available from: https://phit.nl/nl/masterclass-meaningful-use-it-radiology-and-ai-imaging-paul-chang
35. Bassett M. At your service: will service oriented architecture add interoperability to imaging? [Internet]. HealthImaging.com. 2010. Available from: https://www.healthimaging.com/topics/practice-management/your-service-will-service-oriented-architecture-add-interoperability. Accessed 17 June 2018.
36. Patel V. A framework for secure and decentralized sharing of medical imaging data via blockchain consensus. Health Informatics J. 2018; https://doi.org/10.1177/1460458218769699.
37. Schumacher A. Blockchain and healthcare. 2017. 49 p. Available from: https://www.researchgate.net/publication/317936859_Blockchain_Healthcare_-_2017_Strategy_Guide
38. Sagreiya H. A realist's look at artificial intelligence in medicine [Internet]. Medium.com. 2018. Available from: https://opmed.doximity.com/artificial-intelligence-in-medicine-beyond-the-hype-3b8cc0c8b893. Accessed 1 July 2018.
39. ESR ESOR. The new EU general data protection regulation: what the radiologist should know. Insights Imaging. 2017;8(3):295–9.
40. Haug CJ. Turning the tables—the new European general data protection regulation. N Engl J Med. 2018;379(3):207–9. https://doi.org/10.1056/NEJMp1806637.
41. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med. 2018;378(11):981–3.
42. Kohli M, Geis R. Ethics, artificial intelligence, and radiology. J Am Coll Radiol. 2018;15(9):1317–9.
43. Balthazar P, Harri P, Prater A, Safdar NM. Protecting your patients' interests in the era of big data, artificial intelligence, and predictive analytics. J Am Coll Radiol. 2018;15(3 Pt B):580–6.
44. Kohli MD, Summers RM, Geis JR. Medical image data and datasets in the era of machine learning—whitepaper from the 2016 C-MIMI meeting dataset session. J Digit Imaging. 2017;30(4):392–9.
45. Mesko B. Could you sue diagnostic algorithms or medical robots in the future? The Medical Futurist [Internet]. 2018. Available from: https://medicalfuturist.com/could-you-sue-diagnostic-algorithms-or-medical-robots-in-the-future. Accessed 16 June 2018.
46. Delcker J. Europe divided over robot "personhood" [Internet]. Politico. 2018. Available from: https://www.politico.eu/article/europe-divided-over-robot-ai-artificial-intelligence-personhood/. Accessed 16 July 2018.
47. Dwoskin E, Romm T. Facebook makes its privacy controls simpler as company faces data reckoning [Internet]. The Washington Post. 2018. Available from: https://www.washingtonpost.com/news/the-switch/wp/2018/03/28/facebooks-makes-its-privacy-controls-simpler-as-company-faces-data-reckoning/. Accessed 23 July 2018.

AI: A Glossary of Terms

Disclaimer

The glossary of terms contains entries that we think might come handy when studying *Artificial Intelligence in Medical Imaging*. Many of the terms can be found in the preceding chapters. Some descriptions were found on the Internet (see excellent websites such as Techopedia and Medium). In most cases no authors could be traced. If sources were identified, we obtained permission to reproduce. For legibility we avoided mentioning the sources at each entry. References will be given upon request. The authors appreciate feedback if sources are unrightfully omitted.

Thanks to (in alphabetical order): Ameen Abu-Hanna, Anjum Ahmed, Brad Genereaux, Peter van Ooijen, and Martijn Schut.

Jiapan Guo, Postdoc in Medical Imaging Informatics in University Medical Center Groningen, The Netherlands (j.guo@umcg.nl)

Violet Farhang-Razi, MD Northwest Hospital Alkmaar, The Netherlands (v.farhang-razi@nwz.nl)

Paul Algra MD PhD, neuroradiologist Northwest Hospital Alkmaar, The Netherlands (p.r.algra@nwz.nl)

Assembled by Jiapan Guo, Violet Farhang-Razi and Paul Algra (editor).

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2

Glossary

A

Algorithm A formula or set of rules (or procedure, processes, or instructions) for solving a problem or for performing a task. In Artificial Intelligence, the algorithm tells the machine how to find answers to a question or solutions to a problem. In machine learning, systems use many different types of algorithms. Common examples include decision trees, clustering algorithms, classification algorithms, or regression algorithms.

AlexNet The name of a neural network that won the ImageNet Large Scale Visual Recognition Challenge in 2012. It is named after Alex Krizhevsky, then a computer science PhD student at Stanford University. See *ImageNet.*

AlphaGo AlphaGo is the first computer program that defeated a professional player on the board game Go in October 2015. Later in October 2017, AlphaGo's team released its new version named AlphaGo Zero which is stronger than any previous human-champion-defeating versions. Go is played on 19 by 19 board which allows for 10^{171} possible layouts (chess 10^{50} configurations). It is estimated that there are 10^{80} atoms in the universe.

Analogical Reasoning Solving problems by using analogies, by comparing to past experiences.

Anonymization The process in which data is de-identified as part of a mechanism to submit data for machine learning.

Area under curve (AUC) The area under a curve between two points is calculated by performing the definite integral. In the context of a receiver operating characteristic for a binary classifier, the AUC represents the classifier's accuracy.

Artificial Intelligence (AI) Artificial intelligence (or machine intelligence) refers to systems that display intelligent behavior by analyzing their environment and taking actions—with some degree of autonomy—to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g., voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g., advanced robots, autonomous cars, drones, or Internet of Things applications). The term AI was first coined by John McCarthy in 1956.

Artificial Intelligence complete AI-complete, which is short for Artificial Intelligence complete or sometimes called AI-hard, describes the complexity of the computational problems is equal to that of the entire AI problem which aims at producing a general computerized system with the human-level intelligence. An AI-complete problem addresses the fact that the problem cannot be easily solved by a simple specific algorithm.

Artificial Intelligence Winters (AIWI) Artificial Intelligence Winters are periods of time during which artificial intelligence experienced reduced fundings for researches

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2

and low interest from the public. Two major winter periods were in 1974–1980 and 1987–1993. AIW are the result of inflated, unreal expectations.

Artificial General Intelligence (AGI) Artificial general intelligence as opposed to narrow intelligence, also known as complete, strong, super intelligence, Human Level Machine Intelligence, indicates the ability of a machine that can successfully perform any tasks in an intellectual way as the human being. Artificial superintelligence is a term referring to the time when the capability of computers will surpass humans.

Artificial Superintelligence (ASI) Artificial superintelligence is a term referring to the time when the capability of computers will surpass humans. "Artificial intelligence," which has been much used since the 1970s, refers to the ability of computers to mimic human thought. Artificial superintelligence goes a step beyond and posits a world in which a computer's cognitive ability is superior to a human's.

Artificial Narrow Intelligence (ANI) Artificial Narrow Intelligence, also known as weak or applied intelligence, represents most of the current artificial intelligent systems which usually focus on a specific task. Narrow AIs are mostly much better than humans at the task they were made for: for example, look at face recognition, chess computers, calculus, and translation. The definition of artificial narrow intelligence is in contrast to that of strong AI or artificial general intelligence, which aims at providing a system with consciousness or the ability to solve any problems. Virtual assistants and AlphaGo are examples of artificial narrow intelligence systems.

Artificial Neural Network (ANN) Artificial Neural Network (ANN) is a computational model in machine learning, which is inspired by the biological structures and functions of the mammalian brain. Such a model consists of multiple units called artificial neurons which build connections between each other to pass information. The advantage of such a model is that it progressively "learns" the tasks from the given data without specific programing for a single task.

Artificial Neuron An artificial neuron is a digital construct that seeks to simulate the behavior of a biological neuron in the brain. Artificial neurons are typically used to make up an artificial neural network—these technologies are modeled after human brain activity.

Asimov Isaac Asimov's Three Laws are as follows: (1) A robot may not injure a human being. (2) A robot must obey orders, unless they conflict with law number one. (3) A robot must protect its own existence, as long as those actions do not conflict with either the first or second law.

Association Subcategory of unsupervised learning. It can be best explained by market basket analysis (MBA). MBA attempts to identify association/relation between various items that have been chosen by a particular shopper and placed in their respective baskets (real or virtual). The output value from this lies in cross marketing of products and customer behavior analysis. Association is the generalization of m.b.a. Example: there is a good chance that a customer will buy bread if he has already bought milk and eggs.

Augmented Intelligence Augmented Intelligence is the intersection of machine learning and advanced applications, where clinical knowledge and medical data converge on a single platform. The potential benefits of Augmented Intelligence are realized when it is used in the context of workflows and systems that healthcare practitioners operate and interact with. Unlike Artificial Intelligence, which tries to replicate human intelligence, Augmented Intelligence works with and amplifies human intelligence.

Autoregressive Model An autoregressive model is a time series model that uses observations from previous time steps as input to a regression equation to predict the value at the next time step. In statistics and signal processing, an autoregressive model is a representation of a type of random process. It is used to describe certain time-varying processes in nature, economics, etc.

B

Backpropagation Backpropagation, also called "backward propagation of errors," is an approach that is commonly used in the training process of the deep neural network to reduce errors. It allows the machine learning algorithm to adjust itself according to looking at its past function. It involves the calculation of errors between prediction and the target values, the computation of the gradient of the error function, and then the update of the weights. Seen also *feedforward neural network.*

Backward Chaining Backward chaining, also called goal-driven inference technique, is an inference approach that reasons backward from the goal to the conditions used to get the goal. Backward chaining inference is applied in many different fields, including game theory, automated theorem proving, and artificial intelligence.

Batch Normalization A preprocessing step where the data are centered around zero, and often the standard deviation is set to unity.

Bayesian Filter A Bayesian filter is a program using Bayesian logic. It is used to evaluate the header and content of email messages and determine whether or not it constitutes spam—unsolicited email or the electronic equivalent of hard copy bulk mail or junk mail. A Bayesian filter works with probabilities of specific words appearing in the header or content of an email. Certain words indicate a high probability that the email is spam, such as Viagra and refinance.

Bayesian Network A Bayesian Network, also called Bayes Network, belief network, or probabilistic directed acyclic graphical model, is a probabilistic graphical model (a statistical model) that represents a set of variables and their conditional dependencies via a directed acyclic graph (see *DAG*).

Biased algorithm See *Inadvertent effects of AI.*

Big Data The term big data is used when traditional data mining and handling techniques cannot uncover the insights and meaning of the underlying data. Data that are unstructured or time sensitive or simply very large cannot be processed by relational database engines. This type of data requires a different processing approach which uses massive parallelism on readily available hardware.

Blockchain Blockchain is a distributed system that records transactions across all users in an expanding chain of encrypted blocks. Blockchain builds a decentralized ledger that indicates every user has the same copy of the record. The records cannot be easily altered unless all of them are altered. Blockchain was invented in 2008 for the use of cryptocurrency bitcoin as a public transaction ledger. Such a system also shows its potential applications in different fields regarding the recording of events, medical records, and other record management systems.

Boolean neural network Boolean neural network is an artificial neural network approach which only consists of Boolean neurons (and, or, not). Such an approach reduces the use of memory space and computation time. It can be implemented to the programmable circuits such as FPGA (Field-Programmable Gate Array or Integrated circuit).

C

Caffe Caffe is short for Convolutional Architecture for Fast Feature Embedding which is an open source deep learning framework developed in Berkeley AI Research. It supports many different deep learning architectures and GPU-based acceleration computation kernels.

Case-Based Reasoning (CBR) Case-Based Reasoning is a way to solve a new problem by using solutions to similar problems. It has been formalized to a process consisting of case retrieve, solution reuse, solution revise, and case retention.

CE Marking A certification marking indicating conformity with standards for products sold within the European Economic Area. In the context of medical devices, CE Marking is similar to US Food and Drug Administration approval.

Central processing unit (CPU) Central processing unit is the electronic circuit within that carries out the instructions of a computer program by performing the basic arithmetic, logical, control, and input/output operations specified by the instructions (see also *GPU*).

Chatbot Chatbot, also known as interactive agent, is an artificial intelligence system that uses natural language processing techniques to conduct a conversation via audio or texts. The most recognizable examples of chatbots are Apple's Siri, Microsoft's Cortana, and Amazon's Alexa.

Classification Classification is a general process for categorization which assigns a label to the samples. A classification system is an approach to accomplish categorization of samples.

Clinical Decision Support (CDS) A clinical decision support system is a health information technology system that is designed to provide physicians and other health professionals with clinical decision support, that is, assistance with clinical decision-making tasks.

Cloud The cloud is a general metaphor that is used to refer to the Internet. Initially, the Internet was seen as a distributed network and then with the invention of the World Wide Web as a tangle of interlinked media. As the Internet continued to grow in both size and the range of activities it encompassed, it came to be known as "the cloud." The use of the word cloud may be an attempt to capture both the size and nebulous nature of the Internet.

Cloud Computing Cloud Computing enables access to and usage of shared computer resources that can be provisioned with minimum management effort. The cloud is a general metaphor to refer to a group of networked computer resources that could provide computing services to avoid up-front IT infrastructures costs.

Clustering Clustering is a task to organize data into groups based on certain properties. Clustering analysis is widely used in data mining for pattern recognition, image analysis, and computer graphics, among others.

Cognitive computing Cognitive computing is used to refer to the systems that simulate the human brain to help with the decision-making. It uses self-learning algorithms that perform tasks such as natural language processing, image analysis, reasoning, and human–computer interaction. Examples of cognitive systems are IBM's Watson and Google DeepMind.

Cohort A sample in a clinical study (conducted to evaluate a machine learning algorithm, for example) where it is followed prospectively or retrospectively and subsequent status evaluations with respect to a disease or outcome are conducted to determine which initial participants' exposure characteristics (risk factors) are associated with it.

Computer-Aided Detection/Diagnosis (CAD) Computer-aided detection (CAD), or computer-aided diagnosis (CADx), uses computer programs to assist radiologists in the interpretation of medical images. CAD systems process digital images for typical appearances and highlight suspicious regions in order to support a decision taken by a professional.

Common Data Element (CDE) Common Data Element is a tool to support data management for clinical research.

Convolution The process of filtering. A filter (or equivalently: a kernel or a template) is shifted over an input image. The pixels of the output image are the summed product of the values in the filter pixels and the corresponding values in the underlying image.

Convolutional neural network (CNN) A convolutional neural network is a specific type of artificial neural network that uses *perceptrons*, a machine learning unit algorithm, for supervised learning, to analyze data. CNNs apply to image processing, natural language processing, and other kinds of cognitive tasks. A convolutional neural network is also known as a ConvNet. A CNN consists of an input and output layer as well as multiple hidden layers which are formed as mathematical operations. The hidden layers include convolutional layer, pooling layer, normalization, and fully connected layers. Since the success of AlexNet (see *Alexnet*) applied the ImageNet competition in 2013, there has been a rapid evolution of CNNs. VGGNet, GoogLeNet, ResNet, and DenseNet are some successful examples. See *Multilayer neural network*.

Computer Vision Computer Vision is an interdisciplinary field that uses computer science techniques to analyze and understand digital images and videos. Computer vision tasks include object recognition, event detection, motion detection, and object tracking, among others.

D

Data Data is a collection of qualitative and quantitative variables. It contains the information that is represented numerically and needs to be analyzed.

Data Cleaning Data Cleaning is the process of identifying, correcting, or removing inaccurate or corrupt data records.

Data Curation Data Curation includes the processes related to the organization and management of data which is collected from various sources.

Data-Driven Science Data-Driven Science, or Data Science, is an interdisciplinary field of employing computing algorithms to extract knowledge or insights from data acquired from different sources.

Data Extraction Data Extraction is the act or process of retrieving data out of data resources for further data processing or data storage.

Data Integration Data Integration involves the combination of data residing in different resources and then the supply in a unified view to the users. Data integration is in high demand for both commercial and scientific domains in which they need to merge the data and research results from different repositories.

Data Lake A type of data repository that stores data in its natural format and relies on various schemata and structure to index the data.

Data Mining Data Mining is the process of data analysis and information extraction from large amounts of datasets with machine learning, statistical approaches. and many others.

Deductive Reasoning Deductive Reasoning, also known as logical deduction, is a reasoning method that relies on premises to reach a logical conclusion. It works in a top-down manner, in which the final conclusion is obtained by reducing the general rules that hold the entire domain until only the conclusion is left.

Data Refinement Data refinement is used to convert an abstract data model in terms of sets for example into implementable data structures such as arrays.

Decision Tree A decision tree uses tree-like graph or model as a structure to perform decision analysis. It uses each node to represent a test on an attribute, each branch to represent the outcome of the test, and each leaf node to represent a class label.

Data Warehouse A data warehouse is typically an offline copy of production databases and copies of files in a non-production environment.

Deep Blue Deep Blue was a chess supercomputer developed by IBM. It was the first computer chess player that beat the world champion Garry Kasparov, after six-game match in 1997.

Deep Learning (DL) Deep Learning is a subfield of machine learning concerned with algorithms that are inspired by the human brain that works in a hierarchical way. Deep Learning models, which are mostly based on the (artificial) neural networks, have been applied to different fields, such as speech recognition, computer vision, and natural language processing.

DeepMind DeepMind is an artificial intelligence company founded in 2010 and later acquired by Google in 2014. DeepMind developed *Alphago* program that beat a human professional Go player for the first time.

Deep neural network A neural network architecture with many layers, typically 5–100. A network with only a few layers is called a shallow network.

Dice coefficient A measure to compare the similarity of two segmentations, e.g., by expert and by machine. It is the ratio of twice the number of common pixels to the sum of all pixels in both sets.

Directed Acyclic Graph (DAG) In computer science and mathematics, a directed acyclic graph is a finite directed graph with no directed cycles. It consists of finitely many vertices and edges, with each edge directed from one vertex to another, such that there is no way to start at any vertex and follow a consistently directed sequence of edges that eventually loops back to that starting vertex again.

E

Electronic Medical Record (EMR) An electronic medical record, or electronic health record, is the systematized collection of patient and population electronically stored health information in a digital format. These records can be shared across different healthcare settings. Records are shared through network-connected, enterprise-wide information systems or other information networks and exchanges.

ELIZA The ELIZA effect is a term used to discuss progressive artificial intelligence. It is the idea that people may falsely attach meanings of symbols or words that they ascribe to artificial intelligence in technologies.

Enterprise Imaging Enterprise Imaging has been defined as "a set of strategies, initiatives and workflows implemented across a healthcare enterprise to consistently and optimally capture, index, manage, store, distribute, view, exchange, and analyze all clinical imaging and multimedia content to enhance the electronic health record" by members of the HIMSS-SIIM Enterprise Imaging Workgroup.

Error backpropagation The process of adjusting the weights in a neural network by minimizing the error at the output. It involves a large number of iteration cycles with the training data.

Ethics of Artificial Intelligence The ethics of artificial intelligence is the ethics of technology specific to robots and other artificial intelligence beings, which is divided into robot ethics and machine ethics. The former one is about the concern with the moral behavior of humans as they design, construct, use, and treat artificially intelligent beings, and the latter one is about the moral behavior of artificial moral agents (see also *inadvertent effects*).

Expert System Expert system is a computer system that simulates the ability or behavior of a human expert on performing a task. An expert system incorporates the knowledge base that represents facts and rules, and the inference engine that uses the knowledge base to deduce new conclusions.

Explainable artificial intelligence (XAI) Explainable artificial intelligence is a key term in AI design and in the tech community as a whole. It refers to efforts to make sure that artificial intelligence programs are transparent in their purposes and how they work. Explainable AI is a common goal and objective for engineers and others trying to move forward with artificial intelligence progress.

F

Fast Healthcare Interoperability Resources (FHIR) Fast Healthcare Interoperability Resources is a draft standard describing data formats and elements (known as "resources") and an application programming interface for exchanging electronic health records. The standard was created by the Health Level Seven International healthcare standards organization.

Forward Chaining Forward Chaining, also called forward reasoning, is a reasoning approach that searches inference rules from available data and then makes deduction and decision based on the rule. Forward Chaining works in the opposite as the backward chaining.

Feedforward Neural Network A feedforward neural network is an artificial neural network in which the connections between units do not form a cycle. The feedforward neural network has an input layer, hidden layers, and an output layer. Information always travels in one direction—from the input layer to the output layer—and never goes backward. See also *backpropagation*.

Fully Convolutional Network (FCN) Fully Convolutional Network is the first convolutional neural network for semantic segmentation. It is trained end-to-end, pixel-to-pixel from arbitrary-sized inputs. Both learning and inference are performed whole image at a time by dense feedforward computation and backpropagation.

G

Generative Adversarial Network (GAN) A class or artificial intelligence algorithms used in unsupervised machine learning, where two neural networks (a generative network and a discriminative one) are pitted against one another—one network generates candidates, and the other evaluates them in a zero-sum game framework.

Genetic Algorithm Genetic algorithms are heuristic search and optimization algorithms inspired by the natural selection theory. A genetic algorithm requires a genetic representation of the solution and a fitness function to evaluate the solution.

Genomic data Genomic data refer to the genome and DNA data of an organism. They are used in bioinformatics for collecting, storing, and processing the genomes of living things. Genomic data generally require a large amount of storage and purpose-built software to analyze.

Gradient boost machine A type of machine learning technique that uses an ensemble of weak prediction models to perform regression and classification tasks.

Gradient descent A fast optimization method to find a minimum (e.g., error). The gradient is computed at the local position, and walking is done only a step in the downward direction. Repeating this process gives the fastest and most efficient way to the minimum.

Graphical Processing Unit (GPU) A graphical processing unit is a single chip processor designed for efficient manipulation of computer graphics and image processing, especially for computations that can be processed parallely. GPUs are widely used in embedded systems, mobile phones, personal computers, workstations, and many others. The rapid development of GPUs contributes to the rise of deep learning systems. The first GPU was developed by NVidia in 1999 and called the GeForce 256.

H

Heuristics A heuristic is a technique to provide fast or approximate solutions when the traditional methods are too slow or fail to give an accurate solution. A heuristic is commonly called a rule of thumb. While faster, it is typically less optimal than the classic methods it replaces.

Heuristic search techniques Support that narrows down the search for optimal solutions for a problem by eliminating options that are incorrect.

Human Level Machine Intelligence See: *Artificial General Intelligence.*

I

ImageNet ImageNet is a large image database with more than 14 million images over 20,000 categories. Since 2010, the ImageNet project runs annually a contest called the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) for object and scene recognition. The winner algorithm of the contest in the year 2012 is considered as the beginning of the deep learning revolution. See *AlexNet.*

Inadvertent effects of AI If training sets are poised with faulty data, then the algorithm will render faulty outcomes. A system is only as good as the data it learns from and databases must increase in order to let AI grow. See in the literature for racist, sexist algorithms. See also *Ethics of Artificial Intelligence.*

Inductive reasoning Inductive reasoning is a reasoning method which uses premises to supply evidence in order to support the conclusion. Opposed to deductive reasoning, inductive reasoning works as a down-top logic which provides the conclusion by generalizing or extrapolating from special cases to general rules.

Interactive Machine Learning Interactive Machine Learning are approaches based on a coupling of human input and machines during the learning process.

Internet of Things (IoT) The Internet of Things (IoT) is the network of electronic devices embedded with softwares and sensors that enable the interaction between machines. The connectivity between devices helps the collection of huge data which can be analyzed by computer-based systems.

Internet of Medical Things (IoMT) Internet of Medical Things (IoMT) specifies the network of devices that are used to monitor the health status during the daily life.

Interoperability Interoperability is the property that allows for the unrestricted sharing of resources between different systems. This can refer to the ability to share data between different components or machines, or it can be defined as the exchange of information and resources between different computers through local area networks (LANs) or wide area networks (WANs). Broadly speaking, interoperability is the ability of two or more components or systems to exchange information and to use the information that has been exchanged.

Isaac Asimov Isaac Asimov (1920–1992) was a science fiction author and formulated the Three Laws of Robotics in the latter, which continues to influence researchers in robotics and artificial intelligence (AI).

K

Kaggle Kaggle is a data science platform to host data analysis competitions launched by companies and users.

Knowledge-Based Systems It is a computer system that uses knowledge to solve a problem or support a decision. A knowledge-based system has three types of subsystems: a knowledge base, a user interface, and an inference engine.

L

Label Also known as annotation. In supervised learning, the answer or result portion of an example. Each example in a labeled dataset consists of one or more features and a label. For instance, in a housing dataset, the features might include the number of bedrooms, the number of bathrooms, and the age of the house, while the label might be the house's price. In a spam detection dataset, the features might include the subject line, the sender, and the email message itself, while the label would probably be either spam or not spam.

Layer A layer, as in convolutional layer, is a set of neurons in a neural network that process a set of input features, or the output of those neurons. Deep learning networks get their name because they have many layers; most systems now have 30–150 layers, compared with traditional ANNs that would fail if they had more than about three layers.

Learning Learning is the process of acquiring new or modifying existing knowledge, behaviors, skills, values, or preferences. The ability to learn is possessed by humans, animals, and some machines, and there is also evidence for some kind of learning in some plants. Some learning is immediate, induced by a single event but much skill and knowledge accumulates from repeated experiences. See also *deep learning*, *machine learning, unsupervised* and *reinforcement learning.*

Learning algorithm A learning algorithm is an algorithm used in machine learning to help the technology to imitate the human learning process. Combined with technologies like neural networks, learning algorithms create involved, sophisticated learning programs.

Learning algorithm, examples Logic regression, linear regression, decision trees, and random forests are all examples of learning algorithms. Algorithms like "nearest neighbor" also involve the ways that these algorithms are used to affect decision-making and learning in machine learning. In general, what all of these algorithms have in common is their ability to extrapolate from test or training data to make projections or build models in the real world. Think of these algorithms as tools for "pulling data points together" from a raw data mass or a relatively unlabeled background. Where learning algorithms are useful in both supervised and unsupervised machine learning, they are used in different ways in each type of discipline. Supervised machine learning benefits from having already labeled and isolated data, so the learning algorithms that are used will be different in some ways.

Learning rate A scalar used to train a model via gradient descent. During each iteration, the gradient descent algorithm multiplies the learning rate by the gradient. The resulting product is called the gradient step. Learning rate is a key hyperparameter.

Linear regression Linear regression is a kind of statistical analysis that attempts to show a relationship between two variables. Linear regression looks at various data points and plots a trend line. Linear regression can create a predictive model on apparently random data, showing trends in data. See *Learning algorithm, examples.*

Logistic regression Logistic regression is a kind of statistical analysis that is used to predict the outcome of a dependent variable based on prior observations. For example, an algorithm could determine the winner of a presidential election based on past election results and economic data. Logistic regression algorithms are popular in machine learning. See *Learning algorithm, examples.*

M

Machine intelligence See *Artificial Intelligence.*

Machine Learning Machine Learning is a field in computer science that builds computational models that have the ability of "learning" from the data and then provide predictions. Depending on whether there is a supervisory signal, machine learning can be divided into three categories: the *supervised learning*, *unsupervised learning*, and *reinforcement learning.*

Machine Vision Machine Vision is the technology used to provide image-based automatic analysis for applications in industry such as automatic inspection, process control, and robot guidance.

Markov Chain Any multivariate probability density whose independence diagram is a chain. In other words, the variables are ordered, and each variable "depends" only on its neighbors in the sense of being conditionally independent of the others. An equivalent definition is that you sample the variables left-to-right, conditional only on the last outcome.

Mask R-CNN Mask R-CNN is a general deep learning-based framework for object instance segmentation. It consists of two stages, in which the first stage performs a region proposal network that proposes candidate object bounding box, while the second stage provides a class prediction to the instances in the bounding box as well as a binary mask for instance segmentation.

Medical Imaging Informatics MII is the development, application, and assessment of information technology (IT) for clinical medical imaging. It includes the interfaces of IT and people. In practical terms, MII already occurs at a basic level throughout radiology practice, from the moment a clinician considers ordering an imaging study until images and interpretation are used to plan the patient's treatment.

Monte Carlo Methods Monte Carlo Methods, or Monte Carlo Simulation, are computational algorithms that rely on random sampling to obtain numerical results based on probability distributions. One example of using Monte Carlo Method is to approximate the value of π. It is done by uniformly scattering random points inside a square and then computing the ratio between the number of points falling in the circle and that of the total number of points within the square, which is equal to $\pi/4$.

Moore's Law Named after the cofounder of Intel, Moore predicted in 1965 that the number of transistors that can be placed on an integrated circuit doubles every 2 years. This trend has been continuing since 1965 with no signs of any slowdown yet. It can be applied in general to a range of technology areas that are growing at an accelerating rate.

Multilayer neural network A multilayer neural network contains more than one layer of artificial neurons or nodes. They differ widely in design. It is important to note that while single-layer neural networks were useful early in the evolution of AI, the vast majority of networks used today have a multilayer model. Multilayer neural networks can be set up in numerous ways. Typically, they have at least one input layer, which sends weighted inputs to a series of hidden layers, and an output layer at the end. These more sophisticated setups are also associated with nonlinear builds using sigmoids and other functions to direct the firing or activation of artificial neurons. While some of these systems may be built physically, with physical materials, most are created with software functions that model neural activity. Convolutional neural networks (CNNs), used for image processing and computer vision, as well as recurrent neural networks, deep networks, and deep belief systems are all examples of multilayer neural networks. CNNs, for example, can have dozens of layers that work sequentially on an image. All of this is central to understanding how modern neural networks function.

Multilayer Perceptrons (MLP) A multilayer perceptron is a class of feedforward artificial neural network. An MLP consists of at least three layers of nodes. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised backpropagation technique for training. Its multiple layers and nonlinear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

Multi-task Learning Multi-task learning (MTL) is a subfield of machine learning in which multiple learning tasks are solved at the same time while exploiting commonalities and differences across tasks. This can result in improved learning efficiency and prediction accuracy for the task-specific models, when compared to training the models separately.

N

Narrow artificial intelligence (narrow AI) Narrow artificial intelligence (narrow AI) is a specific type of artificial intelligence in which a technology outperforms humans in some very narrowly defined task. Unlike general artificial intelligence, narrow artificial intelligence focuses on a single subset of cognitive abilities and advances in that spectrum.

Natural Language Processing Natural language processing (NLP) is a method to translate between computer and human languages. Traditionally, feeding statistics and models have been the method of choice for interpreting phrases. Recent advances in this area include voice recognition software, human language translation, information retrieval, and artificial intelligence. There is difficulty in developing human language translation software because language is constantly changing. Natural language processing is also being developed to create human readable text and to translate between one human language and another. Already existing reports associated with radiology images can be used to learn about disease and conditions and the ultimate goal of NLP is to build software that will analyze, understand, and generate human languages naturally, enabling communication with a computer as if it were a human.

Neural networks Also known as artificial neural network, neural net, deep neural net; a computer system inspired by living brains. Neural networks found to perform best in ImageNet data challenges were *convolutional* neural networks (CNNs). This name comes from the mathematical concept of convolution, which is similar to the CNN convolutional operation wherein filters are applied to an image in fixed spatial regions and are swept across, or integrated, over the entire image. The resulting activations can then be aggregated in pooling operations, subjected to repeated convolutions, and eventually mapped to a vector of probabilities corresponding to likelihoods that the image belongs to a certain class.

O

Omics The word omics indicates the study of a body of information and refers to the fields of biology ending in -omics such as genome, proteome, microbiome, and exposome. Many of the emerging fields of large-scale data-rich biology are designated by adding the suffix -omics onto previously used terms.

OpenAI OpenAI is a nonprofit artificial intelligence research company (founded in December 2015 by partners including Elon Musk) that aims to promote and develop friendly AI in such a way as to benefit humanity as a whole. The organization aims to "freely collaborate" with other institutions and researchers by making its patents and research open to the public.

Overfitting In statistics and machine learning, overfitting occurs when a model tries to predict a trend in data that is too noisy. Overfitting is the result of an overly complex model with too many parameters. A model that is overfitted is inaccurate because the trend does not reflect the reality of the data. An overfitted model is a model with a trend line that reflects the errors in the data that it is trained with, instead of accurately predicting unseen data. This is better seen visually with a graph of data points and a trend line. An overfitted model shows a curve with higher and lower points, while a properly fitted model shows a smooth curve or a linear regression.

Overfitting, compensation of Overfitting typically results from an excessive number of training points. There are a number of techniques that machine learning researchers can use to mitigate overfitting, including cross-validation, regularization, early stopping, pruning, Bayesian priors, dropout, and model comparison.

P

Pattern matching Pattern recognition and pattern matching are sometimes confused as the same thing when, in fact, they are not. Whereas pattern recognition looks for a similar or most likely pattern in a given data, pattern matching looks for exactly the same pattern. Pattern matching is not considered part of *machine learning*, although in some cases it leads to similar results as pattern recognition. Pattern recognition has its origins in engineering, whereas machine learning grew out of computer science. Both can be viewed as two facets of the same field.

Pattern recognition In IT pattern recognition is a branch of *machine learning* that emphasizes the recognition of data patterns or data regularities in a given scenario. It is a subdivision of *machine learning* and it should not be confused with actual machine learning study. Pattern recognition can be either supervised, where previously known patterns can be found in a given data, or unsupervised, where entirely new patterns are discovered. The objective behind pattern recognition algorithms is to provide a reasonable answer for all possible data and to classify input data into objects or classes based on certain features. A most likely matching is performed between various data samples and their key features are matched and recognized.

Perceptron The perceptron computes a single output from multiple real-valued inputs by forming a linear combination according to its input weights and then possibly putting the output through some nonlinear activation function. A multilayered perceptron is a network of simple neurons called perceptrons. The basic concept of a single perceptron was introduced by Rosenblatt in 1958.

Perceptron algorithm Perceptron algorithm is a machine learning algorithm that helps provide classified outcomes for computing. Perceptron algorithm is called supervised classification because the computer is aided by the human classification of data points. Perceptron is also related to the development of "artificial neural networks," where computing structures are based on the design of the human brain.

Planning A branch of AI dealing with planned sequences or strategies to be performed by an AI-powered machine. Things such as actions to take, variable to account for, and duration of performance are accounted for.

Principal component analysis (PCA) Constructing new features which are the principal components of a dataset. The principal components are random variables of maximal variance constructed from linear combinations of the input features. Equivalently, they are the projections onto the principal component axes, which are lines that minimize the average squared distance to each point in the dataset. To ensure uniqueness, all of the principal component axes must be orthogonal. PCA is a maximum-likelihood technique for linear regression in the presence of Gaussian noise on both inputs and outputs. In some cases, PCA corresponds to a Fourier transform, such as the DCT used in JPEG image compression.

Pruning The use of a search algorithm to cut off undesirable solutions to a problem in an AI system. It reduces the number of decisions that can be made by the AI system.

Python Programming language that runs on most platforms and is often used for data science, machine learning, and deep learning.

R

Radiomics The -omics of images is an expansion of CADx. Radiomics refers to the extraction and analysis of large amounts of advanced quantitative image features with the intent of creating mineable databases from radiological images. From which prognostic associations can be made between images and outcomes.

Radiogenomics This term is used in two contexts. Either to refer to the study of genetic variation associated with response to radiation or to refer to the correlation between cancer imaging features and gene expression. It is the combination of radiomics and genomics, the gene profile of, for example, a tumor. Combining both radiomics and radiogenomics will lead to AI predicting which kind of gene profile defect there is based on its features seen on scans.

Random Forests (or Random Decision Forests) Random Forests or Random Decision Forests are ensembling learning methods for data classification and regression. They construct a multitude of *decision trees* during the training and output the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Receptive field (RF) The sensitivity pattern of a neuron. For example, the receptive field of a simple cell in the primary visual cortex V1 is determined by measuring its firing rate as a function of the pointwise scanning light stimulation of its receptive field area on the retina. A receptive field is the biological implementation of a filter.

Recurrent neural network (RNN) A type of neural network that makes sense of sequential information and recognizes patterns, and creates outputs based on those calculations. Remembers a previous state in its memory, and feeds this back as one of the inputs. It is characterized by a recurrent loop in the architecture. This type of neural network is used for sequential data, e.g., text and video.

Regression Regression is a process of predicting the value to a yes or no label provided it falls on a continuous spectrum of input values, subcategory of *supervised learning*.

Reinforcement Learning Reinforcement learning is a type of dynamic programming that trains algorithms using a system of reward and punishment. The algorithm is exposed to a total random and new dataset and it automatically finds patterns and relationships inside of that dataset. The system is rewarded when it finds a desired relationship inside of that dataset but it is also punished when finds an undesired relation. The algorithm learns from awards and punishments and updates itself continuously. This type of algorithm is always in production mode. It requires real-time data to be able to update and present actions. The agent learns without intervention from a human by maximizing its reward and minimizing its penalty.

Residual neural network (RNN) This network skips connections over network layers, by making shortcuts or jump-overs. A ResNet skips over a single layer.

S

Scikit-learn Scikit-learn (formerly scikits-learn) is a free software machine learning library for the *Python* programming language. It features various classification, regression, and clustering algorithms, including support vector machines, random forests, gradient boosting, k-means, and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Singularity Singularity refers to the emergence of superintelligent machines with capabilities that cannot be predicted by humans. The word singularity comes from astrophysics where it is used to refer to a point in space time where the rules of ordinary physics do not apply. This idea is parallel to the way the term is used in a technological context, because if a technological singularity were to occur, humans would be become unable to predict events beyond that point. See *Superintelligence.*

Strong AI An area of AI development that is working toward the goal of making AI systems that are as useful and skilled as the human mind.

Stride The step size in the shift of convolution filters. It is normally set to 1, but can be 2–10 or even higher, to increase the computational efficiency.

Supervised Learning Training a model from input data and its corresponding labels. Supervised machine learning is analogous to a student learning a subject by studying a set of questions and their corresponding answers. After mastering the mapping between questions and answers, the student can then provide answers to new questions on the same topic. See also *unsupervised machine learning*.

Support Vector Machine (SVM) Support Vector Machine, or in short SVM, is a supervised machine learning model for data classification and regression analysis. One of the most used classifiers in machine learning. It optimizes the width of the gap between the points of separate categories in feature space.

Superintelligence A superintelligence is an intelligence system that rapidly increases its intelligence in a short time, specifically, to surpass the cognitive capability of the average human being. Part of the idea of superintelligence is that certain kinds of artificial intelligence work are theoretically capable of triggering a "runaway reaction" where an artificial intelligence far exceeds human capacity for thought and starts to manipulate or control humans in specific ways. Superintelligence is tied to the idea of a "singularity," which is based on the idea that a catalyst or trigger would cause rapid change beyond what humans can anticipate. See *Singularity.*

T

TensorFlow TensorFlow is an open source software library for numerical computation using data flow graphs. Nodes in the graph represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) communicated between them. The flexible architecture allows you to deploy computation to one or more CPUs or GPUs in a desktop, server, or mobile device with a single API. TensorFlow was originally developed by researchers and engineers working on the Google Brain Team within Google's Machine Intelligence research organization for the purposes of conducting machine learning and deep neural networks research, but the system is general enough to be applicable in a wide variety of other domains as well.

Tensors Multidimensional arrays of primitive data values that are used in TensorFlow. A tensor consists of a set of primitive values shaped into an array of any number of dimensions. These massive numbers of large arrays are the reason that GPUs and other processors designed to do floating point mathematics excel at speeding up these algorithms.

Tensor Processing Unit A unit similar to a Graphic Processing Unit, it is a measure of tensor processing power.

Turing test A test developed by Alan Turing in the 1950s that tests the ability of a machine to mimic human behavior (see terms "*Computing Machinery and Intelligence*"). The test involves a human evaluator who undertakes natural language conversations with another human and a machine and rates the conversations. It is designed to determine whether or not a computer could be classed as intelligent. The test (also referred to as the imitation game) is conducted by having human judges chat to several people via a computer. Most of the people the judges will be speaking to are humans, but one will actually be a chatbot. The chatbot's objective will be to convince the human judges that they are speaking to a real person. If it does this, it has passed the Turing test.

U

Uncanny valley The uncanny valley is a phenomenon that occurs in the human psyche and perception with regard to objects that are human-like, usually robots and images, and determines our reaction toward that object. It is still just a hypothesis, and it is stated to the effect of "as an object such as a robot gets more human-like, the response of some observers will become increasingly positive and emphatic, until a point is reached in the robot's human-likeness beyond which the reactions turn to strong revulsion."

U-net A network with a U-shape, where connections exist between the horizontally corresponding layers of the contracting input branch and the expanding output branch. It was designed to work with fewer training images and to yield more precise segmentations.

Unsupervised learning Unsupervised learning is a type of machine learning algorithm used to draw inferences from sets of data consisting of input data without labeled responses, e.g., cluster analysis. This means that the system is exposed to a total random and new dataset and it automatically finds patterns and relationships inside of that dataset. Unsupervised learning is used in email clustering in order to distinguish between spam emails and useful emails. It can also be seen as Learning by Example. Another example of unsupervised machine learning is *principal component analysis* (PCA). For example, applying PCA on a dataset containing the contents of millions of shopping carts might reveal that shopping carts containing lemons frequently also contain antacids.

Underfitting Underfitting occurs when a statistical model cannot adequately capture the underlying structure of the data.

V

Variational Autoencoder Variational autoencoder (VAE) models inherit autoencoder architecture, but make strong assumptions concerning the distribution of latent variables. They use variational approach for latent representation learning, which results in an additional loss component and specific training algorithm called Stochastic Gradient Variational Bayes. It assumes that the data is generated by a directed graphical model and that the encoder is learning an approximation to the posterior distribution and denote the parameters of the encoder (recognition model) and decoder (generative model), respectively.

W

Watson Watson is named after Dr. Watson, a former IBM CEO. It is a question-answering supercomputer that uses AI to perform cognitive computing and data analysis. In the year 2011, Watson competed on the *Jeopardy!*

television show against human contestants and won the first place prize. Since then, Watson has been used for utilization management in medical centers.

Weak AI See: *Artificial Narrow Intelligence.*

Weights The connection strength (coefficients) between units or nodes in a neural network. These weights can be adjusted in a process called *learning*. The goal of training a linear model is to determine the ideal weight for each feature. If a weight is 0, then its corresponding feature does not contribute to the model.

Winters See *Artificial winters.*

Index

E. R. Ranschaert et al. (eds.), *Artificial Intelligence in Medical Imaging*,
https://doi.org/10.1007/978-3-319-94878-2