MEDICAL IMAGING–REVIEW ARTICLE

# Chest radiographs and machine learning – Past, present and future

Catherine M Jones,[1,2] Quinlan D Buchlak,[2,3,4] Luke Oakden-Rayner,[5] Michael Milne,[1,2] Jarrel Seah,[2,4,6] Nazanin Esmaili[3,7] and Ben Hachey[2,4]

1 I-MED Radiology Network, Brisbane, Queensland, Australia
2 Annalise.ai, Sydney, New South Wales, Australia
3 School of Medicine, The University of Notre Dame Australia, Sydney, New South Wales, Australia
4 Harrison.ai, Sydney, New South Wales, Australia
5 Australian Institute for Machine Learning, The University of Adelaide, Adelaide, South Australia, Australia
6 Department of Radiology, Alfred Health, Melbourne, Victoria, Australia
7 Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, New South Wales, Australia

**CM Jones** MBBS, FRCR, FRANZCR; **QD Buchlak** MD, MPsych, MIS; **L Oakden-Rayner** MBBS, FRANZCR; **M Milne** MS; **J Seah** MBBS; **N Esmaili** PhD, MBA; **B Hachey** PhD.

**Correspondence**
Dr Quinlan D Buchlak, The University of Notre Dame Australia, 160 Oxford St, Darlinghurst, NSW 2010, Australia.
Email: quinlan.buchlak1@my.nd.edu.au

Conflict of interest: All authors have reviewed and approved this manuscript. CJ, QB, JS, MM and BH were employed by or seconded to Annalise.ai. The authors listed above report personal fees from Annalise.ai during the conduct of this study and personal fees from Annalise.ai outside the submitted work. The remaining authors (NE, LOR) have no competing interests.

Submitted 25 February 2021; accepted 8 June 2021.

doi:10.1111/1754-9485.13274

## Summary

Despite its simple acquisition technique, the chest X-ray remains the most common first-line imaging tool for chest assessment globally. Recent evidence for image analysis using modern machine learning points to possible improvements in both the efficiency and the accuracy of chest X-ray interpretation. While promising, these machine learning algorithms have not provided comprehensive assessment of findings in an image and do not account for clinical history or other relevant clinical information. However, the rapid evolution in technology and evidence base for its use suggests that the next generation of comprehensive, well-tested machine learning algorithms will be a revolution akin to early advances in X-ray technology. Current use cases, strengths, limitations and applications of chest X-ray machine learning systems are discussed.

**Key words:** chest X-ray; clinical decision support; deep learning; machine learning; radiomics.

## Introduction

The discovery of the X-ray by Wilhelm Rontgen in 1895 quickly led to the application of chest radiography and fluoroscopy to diagnose numerous chest diseases, including tuberculosis, pneumonia and pneumothorax.[1] This diagnostic leap forward quickly established the chest X-ray (CXR) as an essential component of the diagnostic pathway for chest disease.

For the next half-century, extensive work identified and validated anatomical and pathological signs on the chest radiograph, leading to the principles of CXR interpretation used today. X-ray imaging as a diagnostic tool advanced quickly from initial work to describe imaging appearances of different diseases, later validated in large trials, to the physics of X-ray production and radiation safety. CXR is now extensively used across medical practice, from the acute setting to disease surveillance and screening. It accounts for around 30–40% of all X-ray investigations conducted and compared to other imaging techniques, it is fast, widely available and inexpensive with a low radiation dose.[2,3] It is used globally as a first-line imaging tool for chest assessment.

Despite the ease with which chest radiographs can be obtained, their interpretation can be challenging. The CXR is fundamentally a 2-dimensional representation of a 3-dimensional anatomical structure. X-rays are absorbed by multiple structures as they pass through the thorax, with the overall attenuation of each ray producing the different pixel values in the X-ray image. The composite attenuation of each X-ray beam limits the assessment of the image; ribs and mediastinum obscure up to 40% of the lung parenchyma,[4] and depending on where a pathological lesion is sited, the differences in density between the pathology and adjacent normal structures may be subtle. This may be exacerbated when patient positioning or the degree of inspiration is suboptimal, medical devices or external objects are in the field of view, or in patients with a larger body habitus.

Even with experienced radiologists and technological advancements in chest radiography, the reported error rates for CXR interpretation have remained constant for decades.[5–7] This may be at least partially due to the unchanging principles of CXR interpretation used by radiologists and other clinicians. The effect of non-image factors such as fatigue, interruptions to reporting, environmental factors (temperature, lighting and ergonomics), information system delays or failures, staffing issues and workload may also contribute to radiologist error.[8] Error reduction is one of the key driving factors behind the intense interest in machine learning-driven diagnostic tools to facilitate CXR interpretation.

Biases and non-image factors affecting radiologist performance, including satisfaction of report, satisfaction of search,[9] fatigue, interruptions and the work environment, do not influence machine learning models. A machine learning model may assess the image for all target findings, irrespective of the clinical presentation, underlying disease processes, complexity of the anatomy or study acquisition parameters. Similarly, while radiologists vary in experience and capability (residents, consultants and subspecialists), a model can perform with high accuracy consistently.

We summarise current standards in CXR machine learning, outline factors that lead to successful delivery of CXR machine learning and look to the future of technology and radiology practice.

## Machine learning applications in CXR interpretation

Over the last decade, advances in machine learning technology have led to the development of many new algorithms, including those intended to assist clinicians in interpreting CXR. Digitisation of radiology has allowed for the curation of large, data-rich image collections that are well suited for training deep convolutional neural networks (CNNs). CNNs are commonly used for image interpretation and are based loosely on the functioning of complex neural networks in the human brain.[10,11] CNNs are able to recognise salient clinical features in images once trained on a large data set. While the requirement for large volumes of data has previously been a barrier to effective training, recent years have seen compelling applications developed using well-curated, high-volume CXR data sets.[12]

Convolutional neural networks have now been applied to CXR analysis to successfully detect a wide range of clinical findings. Assessment of diagnostic performance is based on the calculation of many metrics, the most common of which is the area under the receiver operating characteristic curve (AUC). This is a summary statistic indicating diagnostic accuracy independent of disease prevalence in the testing data set. Sensitivity and specificity are well-recognised metrics in clinical practice and are also often reported. The Matthews correlation coefficient (MCC) has been highlighted as a preferred metric for binary classification.[13] Model performance may be compared to that of clinicians and clinician performance with and without model assistance can be assessed.

## Narrow vs comprehensive machine learning models

One useful way to characterise machine learning models is by distinguishing between those that are 'narrow' and those that are 'comprehensive'. Narrow models are trained to complete a single or small number of clinical diagnostic tasks for a given image modality. Whereas comprehensive models are trained to assess an image modality in its entirety for many clinical findings, completing most or all of the tasks that a human expert would be expected to perform in clinical practice.

Many narrow machine learning models have been developed to detect a single finding. The rationale for these has been based on clinical need for particularly salient findings. For example, lung cancer is the most common cancer worldwide and the most common cause of cancer death, with a poor prognosis overall.[14] While computed tomography has greater sensitivity for lung cancer detection in screening programs, the widespread use of CXR across medicine means that it often provides the first opportunity for early diagnosis. However, ninety per cent of missed lung cancers are due to CXR diagnostic errors.[15] Recent evidence suggests that machine learning models designed to identify lung cancer on CXR are highly sensitive.[16] Other studies have demonstrated strong performance of narrow models designed to detect pneumonia,[17] pneumothorax,[18] pneumoconiosis,[19] cardiomegaly,[20] pulmonary hypertension[21] and tuberculosis.[22] However, narrow models may be problematic in that they draw attention to the presence or absence of the finding they were trained to detect, which may distract the interpreting clinician from other subtle but salient clinical findings.

Comprehensive machine learning models are more clinically useful, removing the need for the application of multiple narrow models and providing valuable information about model performance in images that contain combinations of findings. Some recently developed comprehensive CNN models have demonstrated high performance in identifying a wide range of pathologies on CXR.[23–25] Comprehensive deep learning software can match and exceed the performance of human readers in a non-clinical environment. One CNN model achieved radiologist-level performance for 11 of 14 pathologies.[23] Assessment of a machine learning model capable of detecting 72 findings showed good overall performance compared to radiology residents.[24] The most comprehensive model validated to date outperformed radiologists in a non-clinical environment in the detection of 118 findings on chest radiographs and was non-inferior in a total of 124 findings.[25] Across all 124 findings, radiologist macro-averaged AUC was 0.713 and model macro-averaged AUC was 0.956 when compared to a gold standard of thoracic radiologist panel consensus. It is worth noting that the comprehensiveness of these models lies on a spectrum. A model that detects 14 pathologies simultaneously is likely to be more clinically useful than a narrow model, but is less useful than a model detecting 72 findings, which in turn may have less impact than a model that accurately recognises 124 findings. As machine learning models approach the level of comprehensiveness we expect from highly trained human experts, they are becoming more clinically useful.

After a standalone model performance assessment, the next logical step is to assess whether clinician performance is improved when machine learning software is used to assist interpretation. Narrow models have been shown to improve radiologist diagnostic performance for pneumonia, lung nodules and tuberculosis.[26–28] However, it is in multiple disease detection and comprehensive clinical interpretation where machine learning has the greatest potential to deliver substantial improvements to radiologist performance. Several recent studies have shown significant improvements in radiologist performance when assisted by comprehensive CXR machine learning models. One study assessed a machine learning model used to assist radiology residents, demonstrating improved performance in interpreting chest radiographs obtained in the emergency department setting across a range of findings.[29] Most recently, the model developed by Seah *et al*.[25] demonstrated improved radiologist accuracy in over 100 CXR findings comprising a range of acute and non-acute findings.

## Additional benefits to machine learning in CXR

In addition to assisting in the detection of pathology, machine learning has the potential to improve quantitative assessment, such as volumes and distances. These include estimates of lung nodule size or density, positioning of lines and tubes relative to anatomical landmarks, lung volume or cardiothoracic ratio estimates.[30] Machine learning systems are well suited to facilitating these tasks, providing accurate estimates quickly without requiring substantial human input. They contribute additional useful clinical information without the cost of workflow disruption. On a similar theme, bone and calcium suppression to improve lung and mediastinal visibility on CXR studies is typically performed by dual-energy X-ray machines, requiring specialised hardware.[31] Deep learning-based bone suppression systems may offer similar benefits, increasing soft tissue conspicuity and improving the diagnostic accuracy of clinicians,[32] but solely through software, which is much more accessible, particularly in low-resource settings.[33]

In addition to improving radiologist accuracy, machine learning models have the capacity to integrate with workflow systems to triage studies, identifying and serving high priority, time-sensitive findings for faster reporting. Studies suggest that these systems reduce reporting time and alleviate radiologist workloads.[28,34,35] Triage functionality is likely to become more effective as machine learning solutions become more comprehensive; a model that only looks for a single finding (such as pneumothorax) can up-triage cases where that pathology is identified, however, in doing so it will necessarily down-triage cases with other serious or urgent problems (such as free sub-diaphragmatic gas).

## Limitations

Machine learning systems, however, are susceptible to their own limitations.

The generalisability of trained machine learning algorithms is a salient issue. Models may perform well in one context and poorly in another because of variations in imaging infrastructure, patient population characteristics, disease distributions and overfitting.[36] This may lead to over- or under-estimation of clinical findings in some population subgroups or in different clinical environments. The data set used to train a model must be carefully selected and should reflect the patient population to which the model will be applied.

Expanding the use of a machine learning model into populations of patients and disease spectrums different to those represented by the training data set is contingent on responsible consideration of the evidence underpinning the model. Broad generalisability needs to be tested in different local populations, and across a variety of different diseases and clinical contexts.[37,38] Clear and clinically relevant language within machine learning model accuracy validation studies, including describing the patient and disease distributions in the training and test data sets, as well as the potential limitations of model applicability and error rates, is likely to help set appropriate expectations for clinical users. Further research to support the application of machine learning models in wider patient and disease populations is ongoing.

Evidence demonstrating model accuracy and clinician performance improvement must be robust as the mechanisms of a trained algorithm's decision-making processes are often opaque.[39] In CXR interpretation, where some findings may be subjective, model opacity may lead to over- or under-confidence in generated results. This may reduce support for system implementation and degrade clinician adoption.[40] Research on improving algorithm interpretability is generating useful potential solutions.[41,42] Interrogating black box models to assess the reasons for their conclusions can be useful in minimising internal system bias. One method for assessing the areas in an image given most attention by a machine learning model is to visualise a heat map overlaid onto image pixels. This graphs the attention given to each region of the image over the multiple layers of the model network (Fig. 1). However, heatmaps are often difficult to interpret and may be misleading.[43]

One well-known consequence of low machine learning model transparency is hidden stratification. Hidden stratification is a phenomenon that can lead to poor model performance on clinically important patient subsets. While a system may demonstrate high performance overall on a broad disease category, it may perform poorly on clinically meaningful subtypes[44] and lead to failure to detect high-risk pathologies. The archetypal example is that of detecting pneumothoraces without chest drains. As chest drains are visually obvious and inserted to treat pneumothoraces, deep learning models trained to detect pneumothoraces often rely on the absence or presence of the chest drain to achieve high performance. However, when tested on the clinically relevant subset of pneumothoraces without chest drains, they often perform poorly.[45]

Analysing model performance in clinically relevant subsets of cases is therefore important. This may be difficult if labelling of the training and test data sets is not sufficiently detailed to distinguish between subsets of cases[46] and if there is inadequate disease variation in the training data set.[44] Recent evidence suggests that training a machine learning model to both classify the presence of a finding and to provide a localisation (or segmentation) overlay map can reduce the effect of hidden stratification.[45] High-quality CXR interpretation systems intended for implementation will need to appropriately address hidden stratification. The development of comprehensive models that record concurrent relevant findings and improved labelling of training data sets will help.

It has been difficult to obtain large data sets with high-quality, comprehensive labels for model training. While natural language processing techniques are often used to transform free-text reports into categorical labels, these may be inaccurate and subject to bias,[46] partly due to variations in language used in reports. The lack of standardisation in medical terminology and approach to reporting are significant challenges that have inhibited the widespread implementation of structured reporting.[47] RECIST is one example.[48] Structured reporting has been an aspiration for radiology professional bodies for many years.[49]

## Considerations in machine learning CXR implementation

The rapid evolution of CXR machine learning technology and associated evidence to support its implementation has not yet translated well into widespread adoption. High performing but narrow scope systems produce questionable value for clinical users. This has underpinned pessimistic implementation predictions. However, the development of comprehensive models and evidence demonstrating their real-world clinical effectiveness in both workflow triage and improved clinician reporting performance appears to be changing this landscape.

As clinical use becomes more widespread, the development of guidelines and professional standards to address acknowledged risks associated with machine learning systems in radiology safeguards patient safety.[40,50,51] Teams developing, testing and validating machine learning decision support tools should apply these frameworks. Clinicians implementing and using machine learning systems must carefully consider how they were developed and evaluated, including the nature of the training and testing data set populations, generalisability to their own clinical practice and clinical relevance of model outputs.

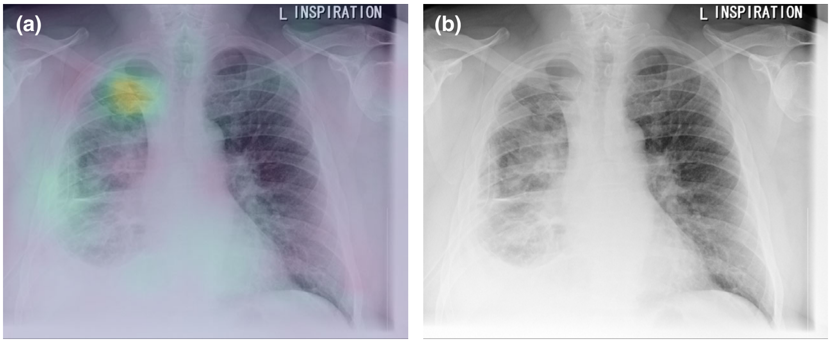

**Fig. 1.** (a) Heatmap investigating an exemplar algorithm[25] in classifying pneumothorax, demonstrating its focus on the right apical pneumothorax rather than the right-sided intercostal drain. (b) Original image demonstrating right apical pneumothorax.

Regulatory frameworks across global jurisdictions are variable. This may drive delays in achieving clearance for clinical use. Regulatory variations may lead to variations in tool accessibility, which in turn may lead to discrepancies in patient care and outcomes. This is especially relevant for comprehensive models that assess a large number of findings, as thorough testing across all findings is required. Small variations in regulatory requirements may require significant effort to satisfy. There is a need for global harmonisation of regulatory frameworks and the development of guidance and standards in relation to good machine learning practices as well as a common minimum standard of evidence to facilitate innovation and implementation while ensuring patient safety.

In addition to these considerations, other factors have, until now, acted as barriers to the widespread implementation of CXR decision support. These include variable awareness of machine learning models across the radiology community, apprehension regarding cost and implementation complexity, patient privacy protection concerns, and uncertain liability for suboptimal clinical outcomes.[52] However, as evidence accumulates suggesting performance improvements associated with machine learning-augmented reporting, the case for the implementation and use of CXR machine learning systems becomes more compelling.

## Looking to the future

Comprehensive image analysis and workflow augmentation comprise the new frontier in applied CXR machine learning practice. It includes systems trained to detect many clinical findings simultaneously, effective calibration to mitigate hidden stratification and generalisation issues, integration into clinical information systems with minimal workflow disruption and identification of time-sensitive findings for faster report generation. The best of these systems promise to augment radiologist clinical performance and increase efficiency. In future, models that effectively incorporate patient-specific information as input (e.g. clinical history and previous imaging) will provide more nuanced and tailored output, facilitating the advancement of precision medicine.

The quality of clinical machine learning decision support systems is dependent on the quality of the full product development lifecycle, from initial design to post-implementation monitoring. Careful data curation and processing are required to ensure that data is broadly representative of clinical populations, to manage label fidelity and to ensure quality model training and validation.[53] Robust clinical evidence is required to demonstrate reliability, validity, safety and beneficial clinical impact. Usability and interpretability for clinical end users are critical to adoption, and effective post-implementation performance and safety monitoring is key to quality management and ensuring patient care improvement.

Machine learning is undoubtedly part of the future of radiology, which may require a shift in mindset regarding achievable outcomes, and careful consideration of how to mitigate possible harms. Being a part of the machine learning development process and driving the implementation of high-quality machine learning systems will be a key responsibility and motivator for radiologists. Part of the role of clinicians is to demand quality machine learning systems and to hold these products to high clinical standards. Radiologists are the guardians of clinical excellence and will play a key role in quality control as powerful and mature machine learning systems begin to filter into clinical practice for the benefit of patients.

## Data availability statement

Data sharing not applicable – no new data generated.

### References

1. Mould RF. The early history of X-ray diagnosis with emphasis on the contributions of physics 1895–1915. *Phys Med Biol* 1995; **40**: 1741.
2. Schaefer-Prokop C, Neitzel U, Venema HW, Uffmann M, Prokop M. Digital chest radiography: an update on modern technology, dose containment and control of image quality. *Eur Radiol* 2008; **18**: 1818–30.
3. United Nations Scientific Committee on the Effects of Atomic Radiation. Sources and effects of ionizing radiation, 2008.
4. Chotas HG, Ravin CE. Chest radiography: estimated lung volume and projected area obscured by the heart, mediastinum, and diaphragm. *Radiology* 1994; **193**: 403–4.
5. Berlin L. Accuracy of diagnostic procedures: has it improved over the past five decades? *Am J Roentgenol* 2007; **188**: 1173–8.
6. Garland LH. Studies on accuracy of diagnostic procedures. *AJR* 1959; **82**: 25–38.
7. White CS, Salis AI, Meyer CA. Missed lung cancer on chest radiography and computed tomography: imaging and medicolegal issues. *J Thorac Imaging* 1999; **14**: 63–8.
8. Itri JN, Tappouni RR, McEachern RO, Pesch AJ, Patel SH. Fundamentals of diagnostic error in imaging. *Radiographics* 2018; **38**: 1845–65.
9. Busby LP, Courtier JL, Glastonbury CM. Bias in radiology: the how and why of misses and misinterpretations. *Radiographics* 2018; **38**: 236–47.
10. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; **521**: 436.
11. Szegedy C, Liu W, Jia Y *et al*. Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015; 1–9.
12. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018; **319**: 1317–8.

13. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom* 2020; **21**: 6.
14. Polanski J, Jankowska-Polanska B, Rosinczuk J, Chabowski M, Szymanska-Chabowska A. Quality of life of patients with lung cancer. *Onco Targets Ther* 2016; **9**: 1023.
15. del Ciello A, Franchi P, Contegiacomo A, Cicchetti G, Bonomo L, Larici AR. Missed lung cancer: when, where, and why? *Diagnostic Interv Radiol* 2017; **23**: 118.
16. Jang S, Song H, Shin YJ *et al*. Deep learning–based automatic detection algorithm for reducing overlooked lung cancers on chest radiographs. *Radiology* 2020; **296**: 652–61.
17. Hurt B, Yen A, Kligerman S, Hsiao A. Augmenting interpretation of chest radiographs with deep learning probability maps. *J Thorac Imaging* 2020; **35**: 285–93.
18. Hwang EJ, Hong JH, Lee KH *et al*. Deep learning algorithm for surveillance of pneumothorax after lung biopsy: a multicenter diagnostic cohort study. *Eur Radiol* 2020; **30**: 3660–71.
19. Wang X, Yu J, Zhu Q *et al*. Potential of deep learning in assessing pneumoconiosis depicted on digital chest radiography. *Occup Environ Med* 2020; **77**: 597–602.
20. Zhou S, Zhang X, Zhang R. Identifying cardiomegaly in chest X-ray8 using transfer learning. *Stud Health Technol Inform* 2019; **264**: 482–6.
21. Zou X-L, Ren Y, Feng D-Y *et al*. A promising approach for screening pulmonary hypertension based on frontal chest radiographs using deep learning: a retrospective study. *PLoS One* 2020; **15**: e0236378.
22. Pasa F, Golkov V, Pfeiffer F, Cremers D, Pfeiffer D. Efficient deep network architectures for fast chest X-ray tuberculosis screening and visualization. *Sci Rep* 2019; **9**: 1–9.
23. Rajpurkar P, Irvin J, Ball RL *et al*. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med* 2018; **15**: e1002686.
24. Wu JT, Wong KCL, Gur Y *et al*. Comparison of chest radiograph interpretations by artificial intelligence algorithm vs radiology residents. *JAMA Netw Open* 2020; **3**: e2022779.
25. Seah J, Tang C, Buchlak QD *et al*. Radiologist chest X-ray diagnostic accuracy performance when augmented by a comprehensive deep learning model. *Lancet Digit Health* 2021.
26. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 2017; **284**: 574–82.
27. Sim Y, Chung MJ, Kotter E *et al*. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. *Radiology* 2020; **294**: 199–209.
28. Kim JH, Kim JY, Kim GH *et al*. Clinical validation of a deep learning algorithm for detection of pneumonia on chest radiographs in emergency department patients with acute febrile respiratory illness. *J Clin Med* 2020; **9**: 1981.
29. Hwang EJ, Nam JG, Lim WH *et al*. Deep learning for chest radiograph diagnosis in the emergency department. *Radiology* 2019; **293**: 573–80.
30. Souza JC, Diniz JOB, Ferreira JL, da Silva GLF, Silva AC, de Paiva AC. An automatic method for lung segmentation and reconstruction in chest X-ray using deep neural networks. *Comput Methods Programs Biomed* 2019; **177**: 285–96.
31. Szucs-Farkas Z, Schick A, Cullmann JL *et al*. Comparison of dual-energy subtraction and electronic bone suppression combined with computer-aided detection on chest radiographs: effect on human observers' performance in nodule detection. *Am J Roentgenol* 2013; **200**: 1006–13.
32. Freedman MT, Lo S-CB, Seibel JC, Bromley CM. Lung nodules: improved detection with software that suppresses the rib and clavicle on chest radiographs. *Radiology* 2011; **260**: 265–73.
33. Gordienko Y, Gang P, Hui J *et al*. Deep learning with lung segmentation and bone shadow exclusion techniques for chest X-ray analysis of lung cancer. In: *International Conference on Computer Science, Engineering and Education Applications*. Springer, 2018; 638–47.
34. Khan FA, Majidulla A, Tavaziva G *et al*. Chest x-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: a prospective study of diagnostic accuracy for culture-confirmed disease. *Lancet Digit Health* 2020; **2**: e573–81.
35. Baltruschat I, Steinmeister L, Nickisch H *et al*. Smart chest X-ray worklist prioritization using artificial intelligence: a clinical workflow simulation. *Eur Radiol* 2021; **31**: 3837–45.
36. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med* 2018; **15**: e1002683.
37. Buchlak QD, Yanamadala V, Leveque J-C, Edwards A, Nold K, Sethi R. The Seattle spine score: predicting 30-day complication risk in adult spinal deformity surgery. *J Clin Neurosci* 2017; **43**: 247–55.
38. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* 2020; **2**: e489–92.
39. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip Rev Data Min Knowl Discov* 2019; **9**: e1312.
40. Buchlak QD, Esmaili N, Leveque J-C, Bennett C, Piccardi M, Farrokhi F. Ethical thinking machines in surgery and the requirement for clinical leadership. *Am J Surg* 2020; **220**: 1372–4.
41. Zhang Q, Zhu S-C. Visual interpretability for deep learning: a survey. *Front Inf Technol Electron Eng* 2018; **19**: 27–39.

42. Seah JCY, Tang JSN, Kitchen A, Gaillard F, Dixon AF. Chest radiographs in congestive heart failure: visualizing neural network learning. *Radiology* 2019; **290**: 514–22.
43. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019; **1**: 206–15.
44. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: *Proceedings of the ACM Conference on Health, Inference, and Learning*, 2020; 151–9.
45. Rueckel J, Trappmann L, Schachtner B *et al*. Impact of confounding thoracic tubes and pleural dehiscence extent on artificial intelligence pneumothorax detection in chest radiographs. *Invest Radiol* 2020; **55**: 792–8.
46. Oakden-Rayner L. Exploring large-scale public medical image datasets. *Acad Radiol* 2020; **27**: 106–12.
47. Nobel JM, Kok EM, Robben SGF. Redefining the structure of structured reporting in radiology. *Insights Imaging* 2020; **11**: 10.
48. Eisenhauer EA, Therasse P, Bogaerts J *et al*. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 2009; **45**: 228–47.
49. European Society of Radiology. ESR paper on structured reporting in radiology. *Insights Imaging* 2018; **9**: 1–7.
50. The Royal Australian and New Zealand College of Radiologists. *Standards of Practice for Artificial Intelligence*. Sydney, Australia, 2020.
51. Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *BMJ* 2020; **370**: m3210.
52. United States Government Accountability Office. AI in Healthcare, 2020.
53. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med* 2016; **375**: 1216.