Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study

Jarrel C Y Seah, Cyril H M Tang, Quinlan D Buchlak, Xavier G Holt, Jeffrey B Wardman, Anuar Aimoldin, Nazanin Esmaili, Hassan Ahmad, Hung Pham, John F Lambert, Ben Hachey, Stephen J F Hogg, Benjamin P Johnston, Christine Bennett, Luke Oakden-Rayner, Peter Brotchie, Catherine M Jones

Summary

Background Chest x-rays are widely used in clinical practice; however, interpretation can be hindered by human error and a lack of experienced thoracic radiologists. Deep learning has the potential to improve the accuracy of chest x-ray interpretation. We therefore aimed to assess the accuracy of radiologists with and without the assistance of a deep-learning model.

Methods In this retrospective study, a deep-learning model was trained on 821 681 images (284 649 patients) from five data sets from Australia, Europe, and the USA. 2568 enriched chest x-ray cases from adult patients (≥16 years) who had at least one frontal chest x-ray were included in the test dataset; cases were representative of inpatient, outpatient, and emergency settings. 20 radiologists reviewed cases with and without the assistance of the deep-learning model with a 3-month washout period. We assessed the change in accuracy of chest x-ray interpretation across 127 clinical findings when the deep-learning model was used as a decision support by calculating area under the receiver operating characteristic curve (AUC) for each radiologist with and without the deep-learning model. We also compared AUCs for the model alone with those of unassisted radiologists. If the lower bound of the adjusted 95% CI of the difference in AUC between the model and the unassisted radiologists was more than –0·05, the model was considered to be non-inferior for that finding. If the lower bound exceeded 0, the model was considered to be superior.

Findings Unassisted radiologists had a macroaveraged AUC of 0·713 (95% CI 0·645–0·785) across the 127 clinical findings, compared with 0·808 (0·763–0·839) when assisted by the model. The deep-learning model statistically significantly improved the classification accuracy of radiologists for 102 (80%) of 127 clinical findings, was statistically non-inferior for 19 (15%) findings, and no findings showed a decrease in accuracy when radiologists used the deep-learning model. Unassisted radiologists had a macroaveraged mean AUC of 0·713 (0·645–0·785) across all findings, compared with 0·957 (0·954–0·959) for the model alone. Model classification alone was significantly more accurate than unassisted radiologists for 117 (94%) of 124 clinical findings predicted by the model and was non-inferior to unassisted radiologists for all other clinical findings.

Interpretation This study shows the potential of a comprehensive deep-learning model to improve chest x-ray interpretation across a large breadth of clinical practice.

Funding Annalise.ai.

Lancet Digit Health 2021; 3: e496–506

Published **Online**
July 1, 2021
https://doi.org/10.1016/S2589-7500(21)00106-0

Annalise.ai, Sydney, NSW, Australia (J C Y Seah MBBS, C H M Tang MD, Q D Buchlak, X G Holt MCS, J B Wardman GD, A Aimoldin MCS, H Ahmad MD, H Pham PhD, J F Lambert MBBS, B Hachey PhD, S J F Hogg MEC, B P Johnston BEng, P Brotchie PhD); **Department of Radiology, Alfred Health, Melbourne, VIC, Australia** (J C Y Seah); **School of Medicine, University of Notre Dame Australia, Sydney, NSW, Australia** (N Esmaili PhD, Prof C Bennett MBBS); **Faculty of Engineering and IT, University of Technology Sydney, Sydney, NSW, Australia** (N Esmaili); **Australian Institute for Machine Learning, The University of Adelaide, Adelaide, SA, Australia** (L Oakden-Rayner MBBS); **Department of Radiology, St Vincent's Health Australia, Melbourne, VIC, Australia** (P Brotchie); **I-MED Radiology Network, Brisbane, QLD, Australia** (C M Jones MBBS)

Correspondence to:
Dr Quinlan D Buchlak, Annalise.ai, Sydney 2000, NSW, Australia
quinlan.buchlak1@my.nd.edu.au

Introduction

Chest x-ray is the most frequently used medical imaging test worldwide.[3] This relatively simple method has allowed investigation of chest pathology, including infection, cardiac pathology, chest trauma, and malignancy, in almost every country worldwide. Advances in digital image acquisition and safe principles of ionising radiation use have led to improved image quality, reduced radiation burden, and wide availability.

However, diagnostic use of chest x-rays has some limitations. Assessment of soft tissue contrast is limited by two-dimensional projection of x-rays through multiple organs, with superimposed densities leading to reduced sensitivity for subtle findings.[4] 90% of cases in which a lung cancer diagnosis was missed were due to errors in the interpretation of chest x-rays.[5] Human error, due to factors such as fatigue or interruptions, and reader inexperience contribute to inaccuracy;[4,6] however, few experienced thoracic radiologists (radiologists who are fellowship trained and have >5 years of post-training experience) are available. For these reasons, several attempts have been made to create artificial intelligence (AI) systems to aid radiologists in the interpretation of chest x-rays.[7,8] Deep-learning diagnostic image processing algorithms based on convolutional neural networks have shown strong performance.[9]

However, although usually highly accurate, most deep-learning systems have a narrow scope and are often

Research in context

Evidence before this study

Deep learning has the potential to improve the accuracy and speed of chest x-ray diagnosis. We searched PubMed and Google Scholar from Jan 1, 1999, to Oct 1, 2020, using search terms “machine learning chest x-ray”, “deep learning”, “artificial intelligence”, “chest x-ray”, “chest radiography”, and “automat* detect*”. Our search identified 559 machine learning studies in chest x-ray diagnostics, which were mostly small proof-of-concept studies. Most previously developed deep-learning chest x-ray interpretation models did not compare clinician accuracy with and without the use of artificial intelligence, focused on only one or a few clinical findings, solely used public prelabelled datasets, did not appropriately address hidden stratification, or involved only a small group (less than ten) of clinical radiologists. We therefore evaluated the effects of a comprehensive deep-learning model on the interpretation of chest x-rays by radiologists.

Added value of this study

Our study evaluated the effects of a deep-learning model for 127 clinical findings on the accuracy of chest x-ray interpretation by radiologists. This model is, to our knowledge, the most comprehensive to date, and was trained on a labelled dataset larger than that used in previous studies (821 681 chest x-rays from 520 014 cases). Diagnostic accuracy was compared with a robust ground truth. When comparing the performance of radiologists with and without assistance from the deep-learning model, we found that the model improved performance across most chest x-ray clinical findings. We report the full underlying ontology tree, which represents the comprehensive chest x-ray clinical interpretation framework of a practising radiologist, to enable future research. Our model has been developed into a clinical decision support tool.

Implications of all the available evidence

Radiologist accuracy improved across a large number of clinical chest x-ray findings when assisted by the deep-learning model. Effective implementation of the model has the potential to augment clinicians and improve clinical practice. The labelled training dataset continues to grow, and research is being done to iteratively and progressively improve the model over time. Detailed subpopulation and error analyses are also being done to enable model development. Research is underway to assess the generalisability of results to various clinical environments and health systems.

limited to a single finding or a small number of findings,[7,10] restricting their use in clinical practice. For example, a decision support system that finds a pneumothorax but misses a pulmonary mass is of questionable clinical benefit. Furthermore, lateral radiographs are often not assessed, despite clear evidence that they contain clinically important information.[11]

Deep-learning chest x-ray analysis systems have been developed to automate lung segmentation and bone exclusion;[12] diagnose tuberculosis;[13] detect pneumonia,[14,15] COVID-19,[16] pneumothorax,[17] pneumoconiosis,[18] and lung cancer;[19] identify the position of feeding tubes;[20] and to predict temporal changes in imaging findings.[21] Deep-learning diagnostic tools have also been shown to improve the classification accuracy of radiologists in the detection of pulmonary nodules,[22] pneumoconiosis,[18] pneumonia,[14,15] emphysema,[7] and pleural effusion.[23] Tschandl and colleagues[24] showed that coupling AI models with clinicians can lead to higher diagnostic accuracy than either AI or physicians alone. Some evidence suggests that use of AI can reduce reporting time[14] and that adjunct diagnostic tools are especially useful for junior clinicians.[25] Several studies have compared the accuracy of AI and radiologists; however, the number of radiologists included in these studies was often less than ten.[14,26,27]

Here we evaluate a deep-learning model designed to assist clinicians in the interpretation of chest x-rays, encompassing the full range of clinically relevant findings on frontal and lateral chest x-rays.

Methods

Study design and participants

In this retrospective, multireader multicase study we evaluated the diagnostic accuracy of 20 radiologists with and without the aid of a deep-learning system. Radiologists interpreted cases without access to the deep-learning tool, then interpreted the same cases with the support of the deep-learning tool after a 3-month washout period. We assessed the change in diagnostic accuracy of radiologists when the deep-learning model was used as a decision support, and also compared the performance of the model alone with that of unassisted radiologists.

Model development and evaluation involved three groups of radiologists (147 fully accredited radiologists in total): 120 consultant radiologists from Vietnam labelled the training dataset, seven specialist thoracic radiologists from Australia did ground truth labelling for the test dataset, and a third group of 20 consultant radiologists from Vietnam interpreted cases in the test dataset. Training dataset labelling defined the radiological findings present on each case in the training dataset. Ground truth labelling defined the radiological findings present in the test dataset.

An overview of the study design is presented in figure 1. This study was reviewed and approved by the human research ethics committee at the University of Notre Dame Australia (Sydney, NSW, Australia; approval number 2020-127S).

Data sources and processing

Chest x-rays used for the training dataset were obtained from multiple datasets: I-MED Radiology Network (I-MED; Australia), MIMIC (Beth Israel Deaconess Medical Center, Boston, MA, USA), ChestX-ray14 (NIH Clinical Center, Bethesda, MD, USA), CheXpert (Stanford University Medical Center, CA, USA), and PadChest (Hospital San Juan, Spain; appendix p 4).[28–31] Images from patients (≥16 years) who had at least one frontal chest x-ray were included in the test dataset. Selected cases were from inpatient, outpatient, and emergency settings. Digital Imaging and Communications in Medicine tags were removed. Protected health information (excluding age and sex) was removed from reports through an automated deidentification process, and patient and case identification were anonymised to deidentify patients while retaining the temporal and logical association between cases and patients. Image data were preserved at the original resolution and bit-depth.

127 chest x-ray findings were identified prospectively by three clinical experts. A chest x-ray finding ontology tree was developed to evaluate and develop the test set (appendix p 5). Each of the findings was defined by consensus between three Australian radiologists, including one subspecialist thoracic radiologist. All participating radiologists in the labelling and evaluation phases were trained to identify chest x-ray findings according to these definitions.

Each case in the training set was independently labelled by three radiologists. Cases were randomly shuffled and placed in a queue. After the radiologist labelled a case, they were allocated the next case according to the random queue order. If a case had already been labelled by that radiologist, the next case in the queue was drawn instead. This ensured that each case was labelled by three different radiologists. Each case consisted of multiple images and the clinical report, which was consistent for each radiologist. Each radiologist was masked to the labels of the other two.

For the training dataset, clinical reports, age, and sex were provided, together with frontal and lateral chest x-rays. Each finding was assigned a present or absent label. Labels consisted of both classification labels on a case level, indicating whether each finding was present in the entire case (multiple images) and each segment for relevant findings. The consensus for each finding for each triple-read case was generated as a consensus score between 0 and 1 using the Dawid–Skene consensus algorithm,[32] which considers the relative accuracies of each labeller for each finding. Segmentation overlays were generated by a single radiologist to localise and depict pathology. This was used to train the model to produce overlay outputs.

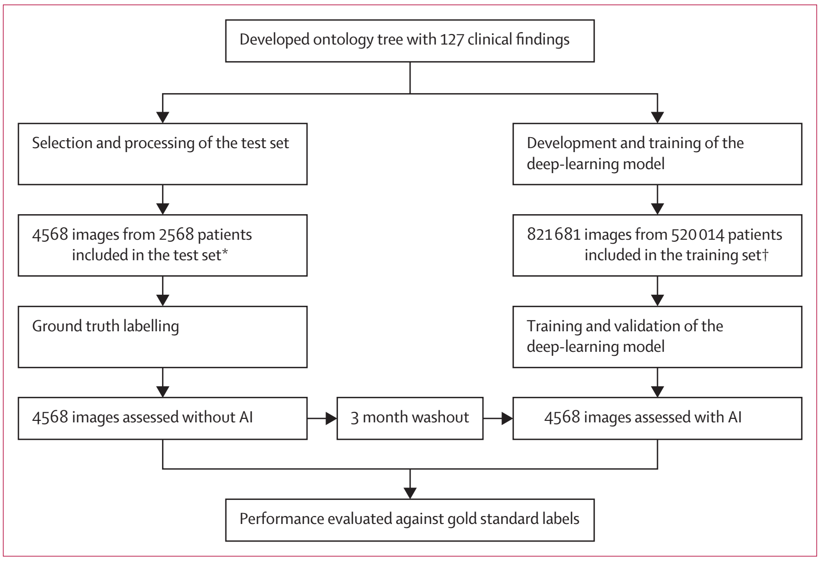
Developed ontology tree with 127 clinical findings
Selection and processing of the test set
4568 images from 2568 patients included in the test set*
Ground truth labelling
4568 images assessed without AI
3 month washout
4568 images assessed with AI
Development and training of the deep-learning model
821 681 images from 520 014 patients included in the training set†
Training and validation of the deep-learning model
Performance evaluated against gold standard labels

Figure 1: Study design

An ontology tree containing 127 clinical findings was developed, datasets were gathered and aggregated, and clinical findings were labelled by a large team of radiologists. The test set contained past and future images, together with clinical reports, which facilitated robust ground truth labelling by three thoracic subspecialist radiologists. The deep-learning model was trained with five-fold cross-validation. The test set was assessed by 20 radiologists both with and without deep learning assistance. *Data sources were I-MED and MIMIC-CXR. †Data sources were I-MED, MIMIC-CXR, NIH ChestX-ray14, CheXpert, and PadChest (appendix p 4).

See Online for appendix

In addition to training the model on the original labels, derived training labels were created based on the ontology tree. For example, focal airspace opacity and airspace opacity—multifocal each belonged to the same group of findings: airspace opacity. Because of this, any case that was labelled with either focal airspace opacity or airspace opacity—multifocal was also automatically labelled with airspace opacity. This meant that the model learnt not just from the original labels but from the structure of the ontology tree. It was penalised less if it classified the original label incorrectly but still correctly classified the parent label.

Test dataset cases were excluded from the model training process such that no patient within the test dataset was present within the training dataset. The test dataset included cases from the I-MED and MIMIC datasets only, and was designed to contain approximately 50% of cases from I-MED and 50% from the MIMIC dataset. Cases were randomly drawn to achieve the target number of cases per finding, while keeping the total number of cases as low as possible. Commonly co-occurring findings were controlled so that episodes of co-occurrence comprised no more than 50% of all cases of that finding within this dataset or did not exceed the baseline co-occurrence rate in the training dataset by more than 10%.

Ground truth labels for the test dataset were determined by consensus between three specialist thoracic radiologists from Australia drawn from a pool of seven who did the ground truth labelling. Radiologists had access to anonymised clinical information, past and future chest x-ray images and reports, and, where available, relevant chest CT reports. They did not have access to the outputs of the deep-learning model. The ground truth labels were derived from a Dawid–Skene[33] consensus algorithm from independent labelling of the cases by the three radiologists.

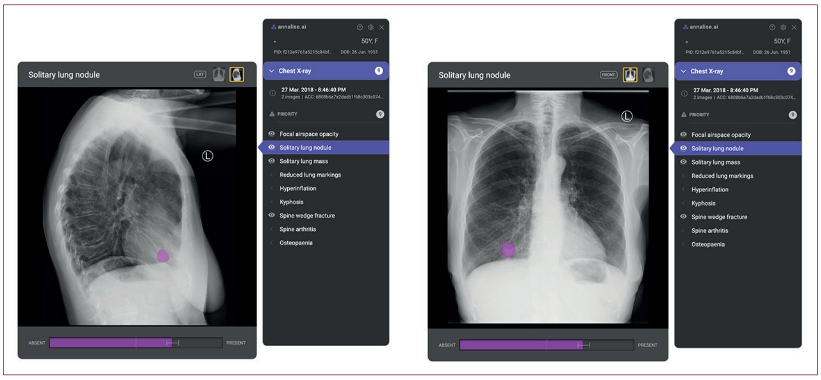

Figure 2: **Deep-learning tool interface**
The clinical findings detected by the deep-learning model are listed on the interface and an image segmentation overlay is presented. The finding likelihood score and CI are displayed as a bar graph under the x-ray. Patient details have been replaced with dummy data.

Before labelling or ground truth annotation, radiologists underwent rigorous training and a screening examination, which involved familiarisation with the annotation tool, reviewing definitions of each clinical finding, and training on a dataset of 113 chest x-rays from the I-MED database covering most findings on the ontology tree. Each labeller was then assessed with the F1 metric (the harmonic mean of precision [positive predictive value] and recall [sensitivity]).[33] Each ground truth labeller had an F1 score averaged across all findings exceeding 0·5, and each training data labeller an F1 score exceeding 0·45.

20 radiologists, each with 5–25 years of clinical experience (median 10·5 years [IQR 6·75–16·75]) after completing radiology specialist training, interpreted all cases in the test dataset. Patient age and sex were shown, but no radiological report or other comparison images were provided. As such, radiologists did a research read in both study groups rather than a clinical read (ie, radiologists identified findings in a research database, but did not write a detailed clinical report). Radiologists were asked to rate their confidence in the presence of each of the 127 findings using a five-point scale (from 0 [unlikely] to 4 [consistent]; appendix p 4) and did not have access to their labelling results when they reviewed cases for the second time after the 3-month washout period.

Labelling, ground truth annotation, and interpretation were done with the same custom-built, web-based digital imaging and communications in medicine viewer (Annalise Web Labelling Tool, Annalise.ai, Sydney, NSW, Australia). Radiologists viewed images and recorded responses on diagnostic quality monitors and hardware, and interpretation times were recorded by the software platform.

Deep learning model

The deep-learning tool consisted of three convolutional neural networks designed for clinical decision support: an image projection classification model (attributes model), a clinical finding classification model (classification model), and a clinical finding segmentation model (segmentation model). The attributes and classification models were based on the EfficientNet architecture,[34] whereas the segmentation model was based on the U-Net[35] architecture with an EfficientNet backbone. Focal loss[36] was minimised with respect to the Dawid–Skene consensus labels. Class-balanced loss weighting[37] accounted for class imbalance and models were trained using five-fold cross-validation.

We assessed the accuracy of the classification model (version 1.2.0; figure 2). Segmentation output was displayed with the model output, but segmentation was not directly evaluated. The attributes model was not directly evaluated, but it prevented the system from producing output for cases that did not have a recognisable frontal anterior–posterior or posterior–anterior image.

Statistical analysis

For the test dataset, we calculated that a minimum dataset of 2568 cases was required to detect a mean difference in area under the receiver operating characteristics (ROC) curve (AUC) of 0·02 in the diagnostic accuracy of at least 18 radiologists labelling all 127 findings (alpha=0·05, beta=0·8). 20 radiologists were recruited to mitigate the risk of dropouts; all of the radiologists completed the assessments.

We first assessed the change in AUC when radiologists were assisted by the model. Clinical findings for which model output was insufficiently powered were retained for this analysis, but discarded for the comparison between the model alone and unassisted radiologists.

The positive predictive value, sensitivity, and specificity for each finding were estimated to assess performance. AUC ROC curves were plotted. The generalised Roe and Metz model and US Food and Drug Administration iMRMC (version 4.0.1) software were used to analyse radiologist accuracy (measured as the AUCROC) with and without the assistance of the model.[38,39] The Matthews correlation coefficient represents the quality of a binary classifier, ranging from –1 (total disagreement) to +1 (total agreement).[40,41] An AUC difference more than 0·05 and a Matthews correlation coefficient difference more than 0·1 were considered to be superior.[42,43] Therefore, findings for which the lower bound of the 95% CI was less than –0·05 were considered inconclusive, findings for which the lower bound was between –0·05 and 0·0 were considered non-inferior, and findings for which the lower bound was higher than 0·0 were considered superior to unaided radiologists.

Positive predictive value, sensitivity, specificity, and Matthews correlation coefficients for each radiologist were calculated by binarising confidence scores for each finding. Any finding with a rating of one or more was considered positive. Bootstrapping was used to assess statistically significant differences in the average Matthews correlation coefficient across the 20 radiologists for each finding between assessment with and without the deep-learning tool. When bootstrapping was done, 10 000 bootstraps of all cases were drawn with resampling to estimate the empirical distribution of the parameter concerned. A two-way repeated measures ANOVA was used to compare mean interpretation time, accounting for between-labeller variation.

A model that has high performance for many findings, but fails to accurately identify crucial findings is not useful in a real-world setting. A subset of 34 crucial findings were identified before the start of the study by a subspecialist thoracic radiologist (appendix p 4). These crucial findings were determined before the analysis and were used to determine the utility of the model. These crucial findings represented the findings most likely to be clinically relevant.

The AUC of the deep-learning model was compared with the average radiologist AUC for each finding with a bootstrapping technique. The Benjamini–Hochberg procedure[39] was used to control the false positive rate, accounting for multiple comparisons.

Significance testing and data processing were done with Python (version 3.7.5), Pandas (version 1.1.0), and NumPy (version 1.19.1). Scikit-learn (version 0.22.2.post1), TensorFlow (version 2.3.0), and EfficientNet (version 1.0.0) were used for design, training, and validation of the deep learning model. Two researchers (JCYS and CHMT) independently did the analysis to verify results. The statistical analyses were verified by an independent biostatistician.

Role of the funding source

Employees of the funder (Annalise.ai) were involved in study design, data collection, data analysis, data interpretation, and writing of the report. The two co-chief executive officers of the funder were not involved in data analysis, data interpretation, or writing of the report, but one was involved in study design and one oversaw data collection.

Results

A total of 821 681 images from 520 014 cases were labelled and included in the training dataset (table 1). The median number of model training cases per clinical finding was 5427 (IQR 1515–18 804). 4568 images from 2568 cases were included in the test dataset (table 1). All 2568 cases were classified by the radiologists and 2551 cases were classified by the model. 17 (0·6%) cases were not interpreted by the attributes model: nine were rejected because no frontal image was recognised by the model, four were rejected because no chest x-ray image was found by the model, three raised a processing error, and one had missing data.

Initially, 127 clinical findings were identified on the ontology tree. However, review of the training and test datasets showed that suboptimal intercostal catheter position, pneumobilia, and portal venous gas were infrequently present in both datasets. Pneumobilia and portal venous gas were therefore not included in the secondary outcome analysis. For intercostal catheter position, the initially separate labels of suboptimal intercostal catheter and satisfactory intercostal catheter were merged to create a single label to identify the presence of an intercostal catheter, which was sufficiently prevalent in the test dataset for analysis. Four findings were dropped (pneumobilia, portal venous gas, in-position intercostal catheter and suboptimal intercostal catheter) and one parent finding was added (intercostal catheter), resulting in the change from 127 to 124 findings. To alleviate concerns regarding multiple comparisons, these three additional comparisons were adjusted using the Benjamini–Hochberg procedure for 127 comparisons for the assessment of change in AUC when radiologists were assisted by the model. 124 clinical findings were predicted by the model, which formed the basis of the comparison between the performance of the model alone and that of unassisted radiologists.

Unassisted radiologists had a macroaveraged AUC of 0·713 (95% CI 0·645–0·785) across the 127 clinical findings. The lowest AUC was obtained for peribronchial cuffing (0·562 [0·504–0·697]). The highest AUCs were obtained for electronic cardiac devices (0·979

	Training dataset	Test dataset
Datasets	MIMIC, I-MED, ChestX-ray14, CheXpert, and PadChest	MIMIC and I-MED
Patients	284 649	2286
Studies	520 014	2568
Images	821 681	4568
Sex		
Male	125 246 (44%)	663 (29%)
Female	125 245 (44%)	640 (28%)
Unknown	34 158 (12%)*	983 (43%)*
Mean age, years	65 (18)	74 (15)
View positions		
Posterior–anterior	91 088 (32%)	640 (28%)
Anterior–posterior	74 009 (26%)	754 (33%)
Lateral	68 316 (24%)	709 (31%)
Unknown or other	76 855 (27%)	183 (8%)
Median number of findings per study	5 (3–7)	7 (5–9)

Data are n (%), mean (SD), or median (IQR). The MIMIC and I-MED datasets contained more complete data than the other publicly available datasets, which meant they were more suitable for use in the test dataset. They enabled a high-quality ground truth labelling process. *MIMIC does not provide complete demographic data for all studies.

Table 1: **Dataset characteristics**

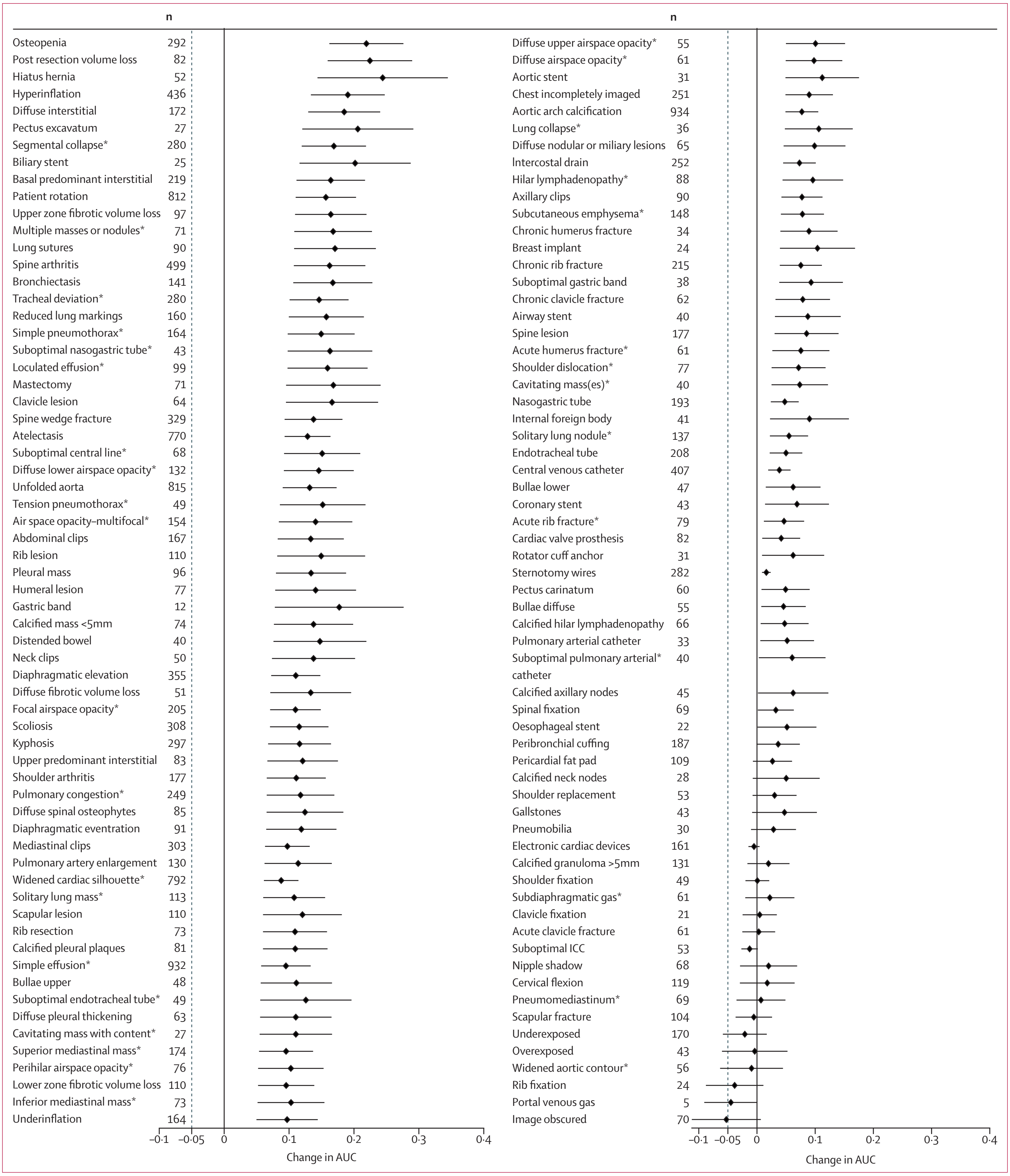
n
Osteopenia 292
Post resection volume loss 82
Hiatus hernia 52
Hyperinflation 436
Diffuse interstitial 172
Pectus excavatum 27
Segmental collapse* 280
Biliary stent 25
Basal predominant interstitial 219
Patient rotation 812
Upper zone fibrotic volume loss 97
Multiple masses or nodules* 71
Lung sutures 90
Spine arthritis 499
Bronchiectasis 141
Tracheal deviation* 280
Reduced lung markings 160
Simple pneumothorax* 164
Suboptimal nasogastric tube* 43
Loculated effusion* 99
Mastectomy 71
Clavicle lesion 64
Spine wedge fracture 329
Atelectasis 770
Suboptimal central line* 68
Diffuse lower airspace opacity* 132
Unfolded aorta 815
Tension pneumothorax* 49
Air space opacity–multifocal* 154
Abdominal clips 167
Rib lesion 110
Pleural mass 96
Humeral lesion 77
Gastric band 12
Calcified mass <5mm 74
Distended bowel 40
Neck clips 50
Diaphragmatic elevation 355
Diffuse fibrotic volume loss 51
Focal airspace opacity* 205
Scoliosis 308
Kyphosis 297
Upper predominant interstitial 83
Shoulder arthritis 177
Pulmonary congestion* 249
Diffuse spinal osteophytes 85
Diaphragmatic eventration 91
Mediastinal clips 303
Pulmonary artery enlargement 130
Widened cardiac silhouette* 792
Solitary lung mass* 113
Scapular lesion 110
Rib resection 73
Calcified pleural plaques 81
Simple effusion* 932
Bullae upper 48
Suboptimal endotracheal tube* 49
Diffuse pleural thickening 63
Cavitating mass with content* 27
Superior mediastinal mass* 174
Perihilar airspace opacity* 76
Lower zone fibrotic volume loss 110
Inferior mediastinal mass* 73
Underinflation 164
–0·1
–0·05
0
0·1
0·2
0·3
0·4
Change in AUC
n
Diffuse upper airspace opacity* 55
Diffuse airspace opacity* 61
Aortic stent 31
Chest incompletely imaged 251
Aortic arch calcification 934
Lung collapse* 36
Diffuse nodular or miliary lesions 65
Intercostal drain 252
Hilar lymphadenopathy* 88
Axillary clips 90
Subcutaneous emphysema* 148
Chronic humerus fracture 34
Breast implant 24
Chronic rib fracture 215
Suboptimal gastric band 38
Chronic clavicle fracture 62
Airway stent 40
Spine lesion 177
Acute humerus fracture* 61
Shoulder dislocation* 77
Cavitating mass(es)* 40
Nasogastric tube 193
Internal foreign body 41
Solitary lung nodule* 137
Endotracheal tube 208
Central venous catheter 407
Bullae lower 47
Coronary stent 43
Acute rib fracture* 79
Cardiac valve prosthesis 82
Rotator cuff anchor 31
Sternotomy wires 282
Pectus carinatum 60
Bullae diffuse 55
Calcified hilar lymphadenopathy 66
Pulmonary arterial catheter 33
Suboptimal pulmonary arterial* catheter 40
Calcified axillary nodes 45
Spinal fixation 69
Oesophageal stent 22
Peribronchial cuffing 187
Pericardial fat pad 109
Calcified neck nodes 28
Shoulder replacement 53
Gallstones 43
Pneumobilia 30
Electronic cardiac devices 161
Calcified granuloma &gt;5mm 131
Shoulder fixation 49
Subdiaphragmatic gas* 61
Clavicle fixation 21
Acute clavicle fracture 61
Suboptimal ICC 53
Nipple shadow 68
Cervical flexion 119
Pneumomediastinum* 69
Scapular fracture 104
Underexposed 170
Overexposed 43
Widened aortic contour* 56
Rib fixation 24
Portal venous gas 5
Image obscured 70
–0·1
–0·05
0
0·1
0·2
0·3
0·4
Change in AUC

***Figure 3*: Change in AUC when radiologists were aided by the deep-learning model**

Mean change in AUC and adjusted 95% CI is shown for each clinical finding. Findings for which the lower bound of the 95% CI crosses –0·05 are considered inconclusive, findings for which the lower bound is between –0·05 and 0·0 are considered non-inferior, and findings for which the lower bound is to the right of the 0·0 are superior to unaided radiologists. The numbers of positive cases in the testing dataset for each finding are presented. AUC=area under the receiver operator characteristic curve. ICC=intercostal catheter. *Crucial clinical finding.

[0·953–0·996]), sternotomy wires (0·967 [0·929–0·993]), and shoulder replacement (0·964 [0·627–1·000]). The accuracy of the unassisted radiologists across all clinical findings is reported in the appendix (pp 7–8).

When radiologists used the deep-learning tool, the macroaveraged AUC was 0·808 (95% CI 0·763–0·839) across the 127 clinical findings. The lowest AUC was obtained for portal venous gas (0·520 [0·499–0·816]). The highest AUCs were for shoulder replacement (0·995 [0·961–1·000]), sternotomy wires (0·983 [0·939–0·999]), and oesophageal stent (0·978 [0·900–1·000]).

Use of the deep-learning model significantly improved accuracy for 102 (80%) clinical findings (figure 3). AUC did not decrease significantly for any finding and was statistically non-inferior for 19 (15%) findings. The effect of the model for the remaining six findings (image obscured, portal venous gas, rib fixation, overexposed, widened aortic contour, and underexposed) was inconclusive because the lower bounds of the 95% CI were less than –0·05 and the upper bounds were more than 0·0. Changes in AUC with and without the use of the deep-learning model across all clinical findings are presented in the appendix (pp 7–8). The three findings that had the greatest AUC increase were hiatus hernia (0·633 to 0·877; difference 0·244 [95% CI 0·144–0·345]), post-resection volume loss (0·654 to 0·879; difference 0·225 [0·159–0·290]), and osteopenia (0·625 to 0·844; difference 0·219 [0·162–0·276]). Of note, rib lesion (0·741 to 0·890; difference 0·149 [0·082–0·217]) and simple pneumothorax (0·746 to 0·895; difference 0·149 [0·098–0·201]), two clinically important findings, also improved significantly.

100 findings had a statistically significant improvement in Matthews correlation coefficient when radiologists used the deep-learning model (appendix pp 7–8). 24 of the remaining findings were statistically non-inferior. Three findings (image obscured, portal venous gas, and overexposed) were inconclusive because the lower bounds of the 95% CI were less than –0·1 and the upper bounds were more than 0·0. Additionally, Matthews correlation coefficients for the detection of any crucial finding on a given case improved by 0·082 (95% CI 0·030–0·139), from 0·491 to 0·574, when radiologists used the deep learning model. Sensitivity for crucial findings also significantly improved from 0·890 to 0·956, and positive predictive value decreased slightly from 0·905 to 0·899 (figure 4). Most findings showed improved sensitivity,

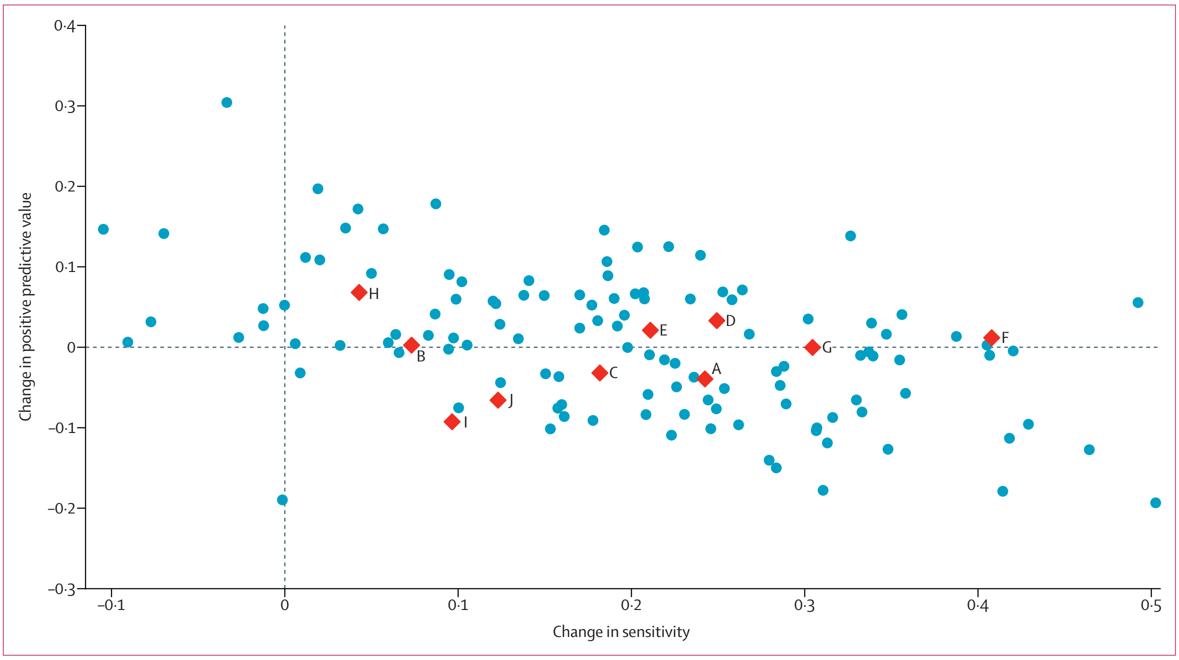

***Figure 4*: Change in positive predictive value and sensitivity when radiologists were aided by the deep-learning model**

Each point represents a single finding. The ten most clinically salient findings of the 34 crucial findings are explicitly labelled (A–J); these labels highlight that the model helps radiologists to more accurately detect findings that are clinically important. A=simple effusion. B=central venous catheter–in position. C=cardiomegaly. D=air space opacity–focal. E=air space opacity–diffuse (central or perihilar). F=lobar or segmental collapse. G=simple pneumothorax. H=free abdominal gas. I=acute rib fracture. J=solitary nodule (<3 cm).

	Model AUC	Unassisted radiologist AUC	AUC difference (adjusted 95% CI)
Acute humerus fracture	0·980	0·765	0·215 (0·009–0·399)
Acute rib fracture	0·948	0·808	0·141 (0·021–0·319)
Air space opacity–multifocal	0·892	0·590	0·302 (0·188–0·392)
Cavitating mass with content	0·979	0·652	0·326 (0·082–0·450)
Cavitating mass(es)	0·929	0·642	0·288 (0·124–0·424)
Diffuse airspace opacity	0·979	0·707	0·272 (0·137–0·405)
Diffuse lower airspace opacity	0·929	0·628	0·299 (0·156–0·414)
Diffuse upper airspace opacity	0·978	0·615	0·364 (0·174–0·478)
Focal airspace opacity	0·842	0·618	0·223 (0·141–0·327)
Hilar lymphadenopathy	0·939	0·617	0·320 (0·151–0·453)
Inferior mediastinal mass	0·963	0·645	0·318 (0·109–0·455)
Loculated effusion	0·945	0·649	0·296 (0·083–0·449)
Lung collapse	0·997	0·806	0·191 (0·022–0·401)
Multiple masses or nodules	0·954	0·679	0·275 (0·104–0·422)
Perihilar airspace opacity	0·934	0·641	0·293 (0·150–0·427)
Pneumomediastinum	0·962	0·677	0·285 (0·158–0·433)
Pulmonary congestion	0·910	0·586	0·324 (0·172–0·416)
Segmental collapse	0·908	0·624	0·283 (0·179–0·403)
Shoulder dislocation	0·977	0·772	0·204 (0·053–0·372)
Simple effusion	0·950	0·784	0·166 (0·086–0·443)
Simple pneumothorax	0·980	0·746	0·234 (0·113–0·391)
Solitary lung mass	0·935	0·727	0·206 (0·092–0·302)
Solitary lung nodule	0·876	0·662	0·214 (0·110–0·349)
Subcutaneous emphysema	0·992	0·871	0·121 (0·043–0·345)
Subdiaphragmatic gas	0·996	0·774	0·225 (0·077–0·407)
Suboptimal central line	0·969	0·668	0·300 (0·081–0·436)
Suboptimal endotracheal tube	0·995	0·746	0·247 (0·061–0·495)
Suboptimal nasogastric tube	0·984	0·631	0·355 (0·137–0·479)
Suboptimal pulmonary arterial catheter	0·992	0·594	0·397 (0·200–0·495)
Superior mediastinal mass	0·950	0·658	0·292 (0·199–0·380)
Tension pneumothorax	0·997	0·739	0·258 (0·037–0·437)
Tracheal deviation	0·948	0·709	0·240 (0·094–0·368)
Widened aortic contour	0·982	0·700	0·282 (0·105–0·491)
Widened cardiac silhouette	0·947	0·779	0·167 (0·103–0·303)

All differences were statistically significant. AUC=area under the receiver operator characteristic curve.

Table 2: **AUC for unassisted radiologists versus the deep-learning model across 34 crucial clinical findings**

with no overall decrease in positive predictive value (appendix pp 16–17).

Mean interpretation time per case when radiologists used the deep-learning model (107 s; SD 35·6) was significantly lower than when they did not use the model (122 s; 37·4; p=0·0045).

Unassisted radiologists had a macroaveraged AUC of 0·717 (95% CI 0·648–0·790) across all 124 clinical findings, compared with a macroaveraged AUC of 0·957 (0·954–0·959) for the deep learning model alone. The lowest AUCs were for peribronchial cuffing (0·829; appendix p 10) and focal airspace opacity (0·842; table 2). The highest AUC of 1·000 was obtained for shoulder replacement, electronic cardiac devices, and sternotomy wires (appendix p 10).

AUCs for the deep-learning model were statistically superior to those for unassisted radiologists for 117 (94%) of 124 clinical findings and statistically non-inferior for all other clinical findings (appendix pp 9–10). The seven remaining clinical findings (shoulder fixation, rib fixation, oesophageal stent, gastric band, pulmonary arterial catheter, clavicle fixation, and shoulder replacement) were non-inferior because the lower bounds of the change in AUC lay between –0·05 and 0·0. ROC curves comparing the model alone with unassisted radiologists are reported in the appendix (pp 18–28). AUCs for unassisted radiologists and the model alone are reported for the subset of 34 crucial findings in table 2.

Discussion

In this retrospective, multireader multicase study, the accuracy of radiologists assisted by a deep learning model was superior to that of unassisted radiologists for 80% of chest x-ray findings and non-inferior for 95% of findings. For the remaining 5% of findings, results were inconclusive. Model-assisted radiologists did not have an inferior performance on any findings compared with unassisted radiologists. Of note, human performance was markedly increased with model assistance.

The deep-learning model alone was either superior or non-inferior to unassisted radiologists for 124 clinical findings. The diagnostic accuracy of the model also compared favourably with that of previously published models (eg, mean ChestNet AUC 0·78).[7,8,10,22,23,44]

The accuracy of the model can be at least partly attributed to the large number of cases labelled by radiologists for model training. The evaluated chest x-ray model was trained on more than 800 000 images, each labelled by radiologists using a prospectively defined ontology tree of chest x-ray findings. Many other large-scale attempts to train deep-learning models on chest x-ray data have relied on text mining from the original radiology reports,[8,45] a process that has been criticised for inconsistency and inaccuracy.[46] Furthermore, the model uses all common chest x-ray projections (anterior–posterior, posterior–anterior, and lateral), which represents the standard of care in real-world settings.

The mechanism underlying improved accuracy for model-assisted radiologists might be complex. When multiple findings are present, radiologists are less likely to perceive them all.[47] In general, missed findings on radiology reports have been attributed to satisfaction of search, difficulties in interpreting technically suboptimal imaging, and human error.[4,6] Overall, the model provided additional information to radiologists, facilitating improved decision making and making interpretation more efficient.

The validity of any diagnostic assessment is dependent on the quality of the ground truth. The ideal ground truth would include cross-sectional imaging, correlation with clinical notes, and follow-up imaging. This is not practical across a large dataset, and many chest x-ray findings are never correlated or followed up. In this study, three subspecialist thoracic radiologists reviewed each case, with access to previous and follow-up chest x-rays and CT imaging, where available, with the corresponding reports. The consensus between the three ground truth radiologists was calculated for each clinical finding using the Dawid–Skene algorithm, which reflects a clinical gold standard for ground truth labels.

The number of radiologists included in the study in both the comparison between the deep-learning model alone and unassisted radiologists and assisted versus unassisted radiologists, was carefully determined, considering the possible effects of heterogeneous interpretation of x-rays. Other chest x-ray AI decision support studies have generally included a much smaller number of radiologists (less than ten radiologists), resulting in a reduction in power that is often unaccounted for in statistical analyses.[13,14] Given the substantial number of findings assessed in our study, we included 20 experienced radiologists.

Hidden stratification is a well known risk of deep-learning models applied to medical imaging,[48] in which visually distinct groups of imaging findings within the broad label categories of a dataset can produce unexpectedly poor accuracy for clinically relevant subgroups. The archetypal example is that of detecting pneumothoraces without chest drains. Because chest drains are visually obvious and inserted to treat pneumothoraces, deep-learning models trained to detect pneumothoraces often rely on the presence or absence of the chest drain for high accuracy. However, when tested on images of pneumothoraces without chest drains, models often perform poorly. Meaningful disease variation must be appropriately described in the dataset used to train the algorithm,[48] as faulty AI models can mislead clinicians.[24] This key issue was addressed by our comprehensive labelling process. Public datasets, such as the National Institutes of Health ChestXray14 dataset,[29] are labelled with a small number of broad finding classes, which do not account for important subsets; therefore, models trained on these datasets cannot effectively evaluate this issue. By comprehensively labelling the training and test datasets, we showed that the high accuracy of our model is maintained across various clinically important subclasses.[48]

Our study has several limitations. For some clinical findings, the AUC for unassisted radiologists was lower than in previous studies (eg, lung nodule detection AUC 0·742,[49] compared with 0·662 in this study). This might relate to differences in study design. Previous studies on diagnostic accuracy for specific chest x-ray findings generally include a higher proportion of positive cases than seen in this study or in clinical practice, with fewer other distracting findings present. This could be explained by the choice of labelling and participant radiologists. However, given that radiologists were required to pass a screening examination showing their understanding of definitions of clinical findings and their ability to detect them, we believe that the relatively lower initial radiologist accuracy is best explained by the difficult task of labelling all 127 findings concurrently without the benefit of other clinical information. Radiologists ground-truth labelling the test set had this additional information, which would usually be available in clinical practice. The ontology tree was highly specific and contained many findings. Radiologists were required not only to detect findings, but also to characterise them.

There was a risk of spectrum bias from the 20 study radiologists, who were from Vietnam and interpreted x-rays in a research setting. Although fully qualified specialist radiologists, their findings might not be representative of radiologists elsewhere. More research is underway to test the generalisability of all of our findings.

Although a 3-month washout period and randomised ordering of cases were used, recall bias cannot be completely eliminated. Washout periods in similar studies range from a minimum of 3 h to a maximum of 2 months, with a median of 1 month.[50] The washout period implemented in our study exceeded these benchmarks, minimising the risk of recall bias.

Our study did not include data on ethnicity and patient demographics beyond age and sex. We recognise this as an important bias mitigation issue and work is underway to explore the generalisability of the model in different geographical settings and people of different ethnicities.

Furthermore, the non-clinical, retrospective design of this study might have influenced the interpretation of chest x-rays. The dataset was enriched with a higher prevalence of rare findings than in normal clinical practice. These factors might restrict direct applicability to the clinical setting.

This study has shown the potential of a deep-learning model to improve the accuracy of chest x-ray interpretation. However, AI will have little effect on practice unless it is validated and implemented in usable tools.[51] A strength of our model is that it has been developed into a ready-to-implement tool that can determine that the input data are appropriate, analyse the images, and present the findings to reporting radiologists. Research investigating the clinical applicability of the model in real-world settings, including effects on patient outcomes, is required. Subset analyses of findings in clinically relevant situations are underway. In future research we intend to explore whether results would differ if radiologists were from the same settings as the data sources. This issue was mitigated in our study by implementing a rigorous training procedure for each of the reading and labelling radiologists, including an assessment of accuracy on a separate set of cases ground truth labelled in a similar manner. Radiologists were required to show competency in this assessment before labelling or reading. Research is also underway to investigate the effect on radiologist workflow and their attitudes towards an AI diagnostic adjunct. Additionally, analysis of the effect of the model on the interpretation of chest x-rays by non-radiologist clinicians will be required, as this system has the potential to improve chest x-ray interpretation in settings where radiologists are scarce.[52]

This diagnostic accuracy study showed that radiologist performance improved when assisted by a comprehensive chest x-ray deep-learning model. The model had a similar or better accuracy than the radiologists for most findings when compared with high-quality, gold standard assessment techniques. Research is underway to confirm the applicability of this model as a diagnostic adjunct in the clinical setting.

Contributors

JCYS, CHMT, BH, SJFH, BPJ, PB, and CMJ conceptualised and designed the study. JCYS, CHMT, and QDB searched the literature. JCYS, CHMT, HP, and PB acquisitioned the data. All authors analysed, visualised, and interpreted the data. JCYS, CHMT, QDB, XGH, PB, and CMJ drafted the manuscript. All authors critically revised, reviewed, and approved the manuscript. All authors had access to the datasets. JCYS, CHMT, JBW, BH, and XGH verified the data. An external biostatistician verified all data and statistical analyses. JCYS and CHMT independently did the statistical analyses. JCYS, CHMT, and CMJ verified the statistical analyses.

Declaration of interests

This study was funded by Annalise.ai. JCYS, CHMT, QDB, XGH, JBW, AA, HA, HP, JFL, BH, SJFH, BPJ, LO-R, PB, and CMJ were employed by or seconded to Annalise.ai and report personal fees from Annalise.ai during the study and outside the submitted work. CMJ was employed by I-MED. All other authors declare no competing interests.

Data sharing

The research team will make the model and radiologist performance datasets and the test dataset available to interested research partners with the goals of supporting the research community and making collaborative contributions to the literature. Requests for access can be made through the Annalise.ai website. The model is publicly available as a commercial software product. Free web-based demonstration can be accessed online.

For the **Annalise.ai website** see https://annalise.ai/contact/

For more on the **model** see https://annalise.ai/products/annalise-cxr/

For more on the **web-based demonstration** see https://cxrdemo.annalise.ai/

Acknowledgments

We thank Graham Hepworth, the independent biostatistician, for his feedback on the statistical methods and data analysis.

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:** 1741–87.

2 Garland LH. Studies on the accuracy of diagnostic procedures. *Am J Roentgenol Radium Ther Nucl Med* 1959; **82:** 25–38.

3 Radiation UNSC on the E of A. Sources and effects of ionizing radiation. 2008. Report United Nations, New York, 2009.

4 Lee CS, Nagy PG, Weaver SJ, Newman-Toker DE. Cognitive and system factors contributing to diagnostic errors in radiology. *AJR Am J Roentgenol* 2013; **201:** 611–17.

5 Del Ciello A, Franchi P, Contegiacomo A, Cicchetti G, Bonomo L, Larici AR. Missed lung cancer: when, where, and why? *Diagn Interv Radiol* 2017; **23:** 118–26.

6 Brady AP. Error and discrepancy in radiology: inevitable or avoidable? *Insights Imaging* 2017; **8:** 171–82.

7 Rajpurkar P, Irvin J, Zhu K, et al. Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv* 2017; published online Nov 14. https://arxiv.org/abs/171105225 (preprint).

8 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.

9 Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. *Radiographics* 2017; **37:** 505–15.

10 Wang H, Xia Y. Chestnet: a deep neural network for classification of thoracic diseases on chest radiography. *arXiv* 2018; published online July 9. https://arxiv.org/abs/180703058 (preprint).

11 Moffett BK, Panchabhai TS, Nakamatsu R, et al. Comparing posteroanterior with lateral and anteroposterior chest radiography in the initial detection of parapneumonic effusions. *Am J Emerg Med* 2016; **34:** 2402–07.

12 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. International Conference on Computer Science, Engineering and Education Applications; Kiev, Ukraine; Jan 18–20, 2018 (638–47).

13 Qin ZZ, Sander MS, Rai B, et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: a multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Sci Rep* 2019; **9:** 15000.

14 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.

15 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.

16 Bassi PRAS, Attux R. A deep convolutional neural network for COVID-19 detection using chest x-rays. *arXiv* 2020; published online April 30. https://arxiv.org/abs/200501578 (preprint).

17 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.

18 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.

19 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.

20 Singh V, Danda V, Gorniak R, Flanders A, Lakhani P. Assessment of critical feeding tube malpositions on radiographs using deep learning. *J Digit Imaging* 2019; **32:** 651–55.

21 Singh R, Kalra MK, Nitiwarangkul C, et al. Deep learning in chest radiography: detection of findings and presence of change. *PLoS One* 2018; **13:** e0204155.

22 Nam JG, Park S, Hwang EJ, et al. Development and validation of deep learning-based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology* 2019; **290:** 218–28.

23 Cicero M, Bilbily A, Colak E, et al. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Invest Radiol* 2017; **52:** 281–87.

24 Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. *Nat Med* 2020; **26:** 1229–34.

25 Hwang EJ, Nam JG, Lim WH, et al. Deep learning for chest radiograph diagnosis in the emergency department. *Radiology* 2019; **293:** 573–80.

26 Nash M, Kadavigere R, Andrade J, et al. Deep learning, computer-aided radiography reading for tuberculosis: a diagnostic accuracy study from a tertiary hospital in India. *Sci Rep* 2020; **10:** 210.

27 Park S, Lee SM, Lee KH, et al. Deep learning-based detection system for multiclass lesions on chest radiographs: comparison with observer readings. *Eur Radiol* 2020; **30:** 1359–68.

28 Johnson AEW, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data* 2019; **6:** 317.

29 Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Conference on Computer Vision and Pattern Recognition; Honolulu, HI, USA; July 21–26, 2017 (2097–106).

30 Irvin J, Rajpurkar P, Ko M, et al. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence; Honolulu, HI, USA; Jan 27 to Feb 1, 2019 (590–97).

31 Bustos A, Pertusa A, Salinas J-M, de la Iglesia-Vayá M. PadChest: a large chest x-ray image dataset with multi-label annotated reports. *Med Image Anal* 2020; **66:** 101797.

32 Dawid AP, Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. *J R Stat Soc Ser C Appl Stat* 1979; **28:** 20–28.

33 Chinchor N, Sundheim BM. MUC-5 evaluation metrics. In: Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, MD, USA; Aug 25–27, 1993.

34 Tan M, Le Q V. Efficientnet: rethinking model scaling for convolutional neural networks. *arXiv* 2019; published online May 28. https://arxiv.org/abs/190511946 (preprint).

35 Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention; Munich, Germany; Oct 5–9, 2015 (234–41).

36 Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. International Conference on Computer Vision; Venice, Italy; Oct 22–29, 2017 (2980–88).

37 Cui Y, Jia M, Lin T-Y, Song Y, Belongie S. Class-balanced loss based on effective number of samples. Conference on Computer Vision and Pattern Recognition; Long Beach, CA, USA; June 16–20, 2019 (9268–77).

38 Gallas BD, Hillis SL. Generalized Roe and Metz receiver operating characteristic model: analytic link between simulated decision scores and empirical AUC variances and covariances. *J Med Imaging* 2014; **1:** 031006.

39 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* 1995; **57:** 289–300.

40 Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 1975; **405:** 442–51.

41 Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 2000; **16:** 412–24.

42 Obuchowski, NA, Bullen JA. Statistical considerations for testing an AI algorithm used for prescreening lung CT images. *Contemp Clin Trials Commun* 2019; **16:** 100434.

43 Gennaro G. The “perfect” reader study. *Eur J Radiol* 2018; **103:** 139–46.

44 Yates EJ, Yates LC, Harvey H. Machine learning “red dot”: open-source, cloud, deep convolutional neural networks in chest radiograph binary normality classification. *Clin Radiol* 2018; **73:** 827–31.

45 Elkin PL, Froehling D, Wahner-Roedler D, et al. NLP-based identification of pneumonia cases from free-text radiological reports. AMIA Annu Symp Proc; 2008 **2008:** 172–76.

46 Oakden-Rayner L. Exploring large-scale public medical image datasets. *Acad Radiol* 2020; **27:** 106–12.

47 Berbaum KS, Krupinski EA, Schartz KM, et al. Satisfaction of search in chest radiography 2015. *Acad Radiol* 2015; **22:** 1457–65.

48 Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. Conference on Health, Inference, and Learning; Toronto, ON, Canada; April 2–4, 2020 (151–59).

49 Shiraishi J, Abe H, Li F, Engelmann R, MacMahon H, Doi K. Computer-aided diagnosis for the detection and classification of lung cancers on chest radiographs: ROC analysis of radiologists’ performance. *Academic Radiol* **13:** 995–1003.

50 Dendumrongsup T, Plumb AA, Halligan S, Fanshawe TR, Altman DG, Mallett S. Multi-reader multi-case studies using the area under the receiver operator characteristic curve as a measure of diagnostic accuracy: systematic review with a focus on quality of data reporting. *PLoS One* 2014; **9:** e116018.

51 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:** 1372–74.

52 Adil SM, Elahi C, Gramer R, et al. Predicting the individual treatment effect of neurosurgery for TBI patients in the low resource setting: a machine learning approach in Uganda. *J Neurotrauma* 2021; **38:** 928–39.