BMJ Open

Real-World Impact of a Comprehensive Deep Learning Model Designed to Assist Chest Radiograph Reporting

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-052902
Article Type:	Original research
Date Submitted by the Author:	04-May-2021
Complete List of Authors:	Jones, Catherine; Annalise-AI; I-Med Network Danaher, Luke; I-Med Network Milne, Michael; Annalise-AI; I-Med Network Tang, Cyril; Annalise-AI Seah, Jarrel; Alfred Health, Radiology; Annalise AI, Oakden-Rayner, Luke; The University of Adelaide, Australian Institute for Machine Learning Johnson, Andrew; Annalise-AI Buchlak, Quinlan; Annalise-AI; The University of Notre Dame Australia School of Medicine Sydney Campus Esmaili, Nazanin; The University of Notre Dame Australia School of Medicine Sydney Campus; University of Technology Sydney
Keywords:	Chest imaging < RADIOLOGY & IMAGING, RADIOLOGY & IMAGING, Diagnostic radiology < RADIOLOGY & IMAGING, Interventional radiology < RADIOLOGY & IMAGING

SCHOLARONE™
Manuscripts

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Real-world impact of a comprehensive deep learning model designed to assist chest radiograph reporting

Catherine M Jones[1,2], Luke Danaher[2], Michael R Milne[1,2*], Cyril Tang[1], Jarrel Seah[1,3], Luke Oakden-Rayner[4], Andrew Johnson[1], Quinlan D Buchlak[1,5], Nazanin Esmaili[5,6]

[1]Annalise-AI, Sydney, NSW, Australia
[2]I-MED Radiology Network, Sydney, NSW, Australia
[3]Department of Radiology, Alfred Health, Melbourne, VIC, Australia
[4]Australian Institute for Machine Learning, The University of Adelaide, Adelaide, SA, Australia
[5]School of Medicine, University of Notre Dame Australia, Sydney, NSW, Australia
[6]Faculty of Engineering and IT, University of Technology Sydney, Ultimo, NSW, Australia

*Correspondence to: michael.milne@annalise.ai

Corresponding author:
Name: Michael Milne
Annalise-AI
Sydney, Australia
E-mail: michael.milne@annalise.ai

Keywords: Machine learning; radiomics; chest X-ray, deep learning.

Word Count: 4,201

ABSTRACT

Objectives: AI algorithms have been developed to detect imaging features on chest X-ray (CXR), however most of these algorithms are limited to detecting a single finding or a small set of findings. Recently, a comprehensive AI model capable of detecting 124 CXR findings was developed and cleared for clinical use. The aim of this study was to evaluate the real-world performance of the model as a diagnostic assistance device for radiologists.

Design: This prospective real-world multicentre study involved a group of radiologists using the model in their daily reporting workflow to report consecutive chest X-rays and recording their case-by-case feedback on level of agreement with the model findings and whether this significantly affected their reporting.

Setting: The study took place at multiple radiology clinics and hospitals within a large radiology network in Australia between November and December, 2020.

Participants: Eleven consultant radiologists of general diagnostic and interventional backgrounds, and varying levels of experience participated in this study.

Primary outcome measures: Proportion of CXR cases that had significant material changes to the radiologist report, to patient management, or to imaging recommendations due to the model's recommendations. Level of agreement between the radiologist and the model findings.

Results: Of 2,972 cases reviewed with the model, 92 cases (3.1%) had significant report changes, 43 cases (1.4%) had changed patient management and 29 cases (1.0%) had further imaging recommendations. In terms of agreement with the model, 2,572 cases showed complete agreement (86.5%). 390 (13%) cases had one or more findings rejected by the radiologist. There were 16 findings across 13 cases (0.5%) that were deemed to be missed by the model.

Conclusions: Use of an AI model in a real-world reporting environment significantly improved radiologist reporting and showed good agreement with radiologists, highlighting the potential for AI decision support to improve clinical practice.

ARTICLE SUMMARY

Strengths and limitations of this study

- This is the first study to evaluate the real-world significance of integrating a comprehensive CXR AI model into a radiology workflow.
- This was a multicentre study conducted across a mix of public hospitals, private hospitals, and community clinic settings.
- Due to the design of the study, diagnostic accuracy of the decision support system was not a measurable outcome.
- Results of this study are self-reported and may therefore be prone to bias.
- Determination of the significance of report changes due to the model's recommendations was made at the discretion of each radiologist on a case-by-case basis.

INTRODUCTION

Radiology is a data-rich medical specialty and is well placed to embrace artificial intelligence [1] especially in high volume imaging tasks such as chest x-ray imaging. The rapid application of X-ray technology to diagnosing chest diseases at the end of the 19th century led to the chest X-ray (CXR) becoming a first-line diagnostic imaging tool [2] and it remains an essential component of the diagnostic pathway for chest disease. Due to advancements in digital image acquisition, low ionising radiation and low cost, the chest radiograph is more easily accessible worldwide than any other imaging modality [3].

The challenges of interpreting CXR, however, have not lessened over the last half-century. CXR images are 2D representations of complex 3D structures, relying on soft tissue contrast between structures of different densities. Multiple overlapping structures lead to reduced visibility of both normal and abnormal structures [4], with up to 40% of the lung parenchyma obscured by overlying ribs and the mediastinum [5]. This can be further exacerbated by other factors including the degree of inspiration, other devices in the field of view, and patient positioning. In addition, there is a wide range of pathology in the chest which is visible to varying degrees on the CXR. These factors combine to make CXRs difficult to accurately interpret, with an error rate of 20-50% for CXRs containing radiographic evidence of disease reported in the literature [6]. Notably, lung cancer is one of the most common cancers worldwide and is the most common cause of cancer death worldwide [7], and CXR interpretation error accounts for 90% of cases where lung cancer is missed [8]. Despite technological advancements in CXR over the past 50 years, this level of diagnostic error has remained constant [6].

A rapidly developing field attempting to assist radiologists in radiological interpretation involves the application of machine learning, in particular deep neural networks [9]. Deep neural networks learn patterns in large, complex datasets, enabling the detection of subtle features and outcome prediction [10,11]. The potential of these algorithms has grown rapidly in the past decade thanks to the development of more useful neural network models, the advancements in computational power, and the increase in the volume and availability of digital imaging datasets [11]. Of note is the rise of convolutional neural networks (CNNs), a type of deep neural network that excels at image feature extraction and classification, and demonstrate strong performance in medical image analysis, leading to the rapid advancement of computer vision in medical imaging [12,13]. CNNs have been used to develop models to successfully detect targeted clinical findings on CXR, including lung cancer [14,15], pneumonia [16,17], COVID-19 [18], pneumothorax [19–22], pneumoconiosis [23], cardiomegaly [24], pulmonary hypertension [25] and tuberculosis [26–30]. These studies highlight the effectiveness of applied machine learning in CXR interpretation, however most of these deep learning systems are limited in scope to a single finding or a small set of findings, therefore lacking the broad utility that would make them useful in clinical practice.

Recently, our group developed a comprehensive deep learning CXR decision support model, which was designed to assist clinicians in CXR interpretation and improve diagnostic accuracy, validated for 124 clinically relevant findings seen on frontal and lateral chest radiographs [31]. The primary objective of the current study was to evaluate the real-world performance of the model as a diagnostic assist device for radiologists in both hospital and community clinic settings. This involved examining the frequency at which the model's recommendations led to a ‘significant impact on the report’, defined as the inclusion of findings recommended by the model which altered the radiologists report in a meaningful way. The rate of change in patient management and recommendations for further imaging were also evaluated. A secondary endpoint was investigating the agreement between the radiologist and the findings detected by the model. The other secondary endpoint was the assessment of radiologist attitudes towards the tool and the AI models in general.

METHODS

Ethics Statement

This study was approved by the institutional human research ethics committee of the Wesley Hospital, Brisbane, Queensland Australia (2020.14.324). The requirement of patient consent was waived by the ethics committee due to the low-risk nature of the study.

Model development and validation

A modified version of a commercially available CNN-based decision support system (CXR viewer) (Annalise CXR ver 1.2, Annalise-AI, Sydney, Australia) was evaluated [32]. Details of model development and validation have been published in Seah et al [31]. Briefly, a deep learning model consisting of attribute and classification CNNs based on the EfficientNet architecture [33] and a segmentation CNN based on U-Net [34] with EfficientNet backbone was developed. The model was trained on a dataset consisting of 821,681 de-identified CXR images from 284,649 patients originating from inpatient, outpatient and emergency settings across Australia, Europe, and North America. Training dataset labelling involved independent triple labelling of all images by three radiologists selected from a wider pool of 120 consultant radiologists. The model was validated for 124 clinical findings in a multi-reader, multi-case (MRMC) study [31]. Thirty-four of these findings were deemed priority findings based on their clinical importance. The full list of 124 findings is available in Supplementary Table 1, and the 34 critical findings are listed in Table 1, the full list of findings were identical for this study. Ground truth labels for the validation study dataset were determined by a consensus of three independent radiologists drawn from a pool of seven fully credentialed subspecialty thoracic radiologists. The algorithm is publicly available at https://cxrdemo.annalise.ai. The AI model was used in line with pre-existing regulatory approval.

Table 1 - List of the 34 critical clinical findings that the model is validated to detect. ETT: endotracheal tube, NGT: nasogastric tube, PAC: pulmonary artery catheter.

Critical Clinical Findings

Acute humerus fracture	Loculated effusion	Subcutaneous emphysema
Acute rib fracture	Lung collapse	Subdiaphragmatic gas
Air Space Opacity - Multifocal	Multiple masses or nodules	Suboptimal central line
Cavitating mass with content	Perihilar airspace opacity	Suboptimal ETT
Cavitating mass(es)	Pneumomediastinum	Suboptimal NGT
Diffuse airspace opacity	Pulmonary congestion	Suboptimal PAC
Diffuse lower airspace opacity	Segmental collapse	Superior mediastinal mass
Diffuse upper airspace opacity	Shoulder dislocation	Tension pneumothorax
Focal airspace opacity	Simple effusion	Tracheal deviation
Hilar lymphadenopathy	Simple pneumothorax	Widened aortic contour
Inferior mediastinal mass	Solitary lung mass	Widened cardiac silhouette
	Solitary lung nodule	

Technical Integration

Prior to the start of the study, technical integration of the software into existing radiology practice systems and testing occurred over several weeks. First, an integration adapter was installed on the IT network of each radiology clinic and acted as a gateway between the internal IT infrastructure and the AI model. Auto-routing rules were established ensuring only CXR studies were forwarded to the Integration Adapter from the picture archiving and communication system (PACS). Following a successful testing period, the Annalise CXR viewer was installed and configured on workstations for the group of study radiologists.

Study Participants

Eleven consultant radiologists working for a large Australian radiology network were invited to participate in the study through their local radiologist network. This group included a mix of general diagnostic and interventional radiologists who had completed specialist radiology training. The group included radiologists with a range of experience levels: five radiologists had 0–5 years post-training experience, three radiologists had 6–10 years of experience, and three radiologists had more than 10 years of experience. Radiologists were situated across four states in Australia and worked in public hospitals, private hospitals and community clinic settings. Written informed consent was obtained from each participating radiologist. Prior to study commencement, each radiologist attended a training seminar and a one-on-one training session to fully understand the CXR viewer and its features. In addition, the participating radiologists were able to familiarise themselves with the viewer prior to commencement of data collection.

CXR Case Selection

In this multicentre real-world prospective study, all consecutive chest radiographs reported by the radiologists originating from inpatient, outpatient, and emergency settings were included for a period covering nearly six weeks. The CXR cases were reported with the assistance of the AI tool in real-world clinical practice, using high resolution diagnostic radiology monitors within the radiologists' normal reporting environment.

At least one frontal chest radiograph was required for analysis by the model, and cases that did not include at least one were excluded. Chest radiographs from patients aged younger than 16 years were excluded, as the CXR viewer has not been validated in these patients. Data from all sources was de-identified for analysis.

AI-Assisted Reporting

For each CXR case, radiologists produced their clinical report with access to clinical information, the referral and available patient history, in line with the normal workflow. Model output was displayed to the radiologist in a customised image viewer, linked to the image in the PACS, automatically launching when a CXR case was opened (Figure 1)

The modified version of the commercially available AI software gathered feedback from radiologists during the reporting process. For each case, the model provided a list of suggested findings, listed as "priority" or "other", along with a confidence indicator and, in some cases, a region of interest localiser overlayed on the image. The CXR viewer was configured to display its findings after the radiologists initial read of the case. For each case, the radiologist was asked to review the CXR viewer's findings and provide feedback within the viewer. The options presented to the radiologists in the viewer are listed in Table 2.

Table 2 - List of review options presented to the radiologist with each case.

REVIEW OPTION	DESCRIPTION
Rejected clinical finding	A model-detected finding disputed by the radiologist
Missed clinical finding	A model-detected finding missed by the radiologist
Add additional findings	Finding(s) identified by the radiologist but not identified by the model
These findings significantly impacted my report	A yes/no binary question relating to the effect of the model output on the radiologist report
These findings may impact patient management	A yes/no binary question relating to the effect of the model output on patient management, as perceived by the reporting radiologist
These findings led to additional imaging recommendations	A binary yes/no question related to whether the radiologist recommended further imaging based on the model output

The outcome measure of 'significant impact on the report' was the primary outcome measure. A significant change was described as the inclusion of findings recommended by the model, which altered the radiologists report in a meaningful way. As this varied by patient and clinical setting, it was left to the discretion of the radiologist. For example, missing a pneumothorax in a ventilated ICU patient with known pneumothorax would not have the same significance as a previously unknown pneumothorax in an outpatient. During the analysis of radiologist feedback, it was assumed that a change in patient management or further imaging recommendation would not occur without radiologists indicating a material change in the CXR report, and thus management and imaging questions were dependent on a significant change in the report. Free text input describing missed findings or other relevant data were manually added after data collection was complete.

Post-Study Survey

Upon completion of data collection, a post-study survey was distributed to all participating radiologists to obtain feedback on the usefulness of the CXR viewer and how it affected their opinion of AI in radiology. A table of the survey questions is presented in Supplementary Table 2.

Statistics and Data Analysis

A 1% rate of significant changes in reports (the primary outcome measure) was deemed to be clinically significant prior to commencing the study. Based on estimations of the prevalence of missed critical findings on CXR, preliminary power calculations estimated that the number of cases required to detect at least a 1% rate of significant changes in reports was approximately 2000 cases in total, with alpha value 0.05 and desired power of 0.90. To account for any dropout in radiologists or cases, a target of 3000 cases was set for the study. Ten radiologists were recruited, with an eleventh included for any unexpected participant drop out and to achieve this target in a reasonable time period.

A two-tailed binomial test was used to test the hypothesis that the rate of significant report change, patient management change, or imaging recommendation change was 1%. To ensure that the sampling of CXRs reasonably approximates a random snapshot of the true population, radiologists in various states, experience levels as well as different conditions of practice (community clinic vs hospital based) were selected. Additionally, the study was conducted prospectively which further aligns the structure of the sampled data with the expected structure of the population, justifying the choice of analysing the sample using a binomial test without adjustment for each radiologist.

Multivariate logistic regression using generalised linear mixed effect analysis was used to assess the effect of several possible confounders on the measured outcomes, including the number of critical clinical findings per case identified by the model, the inpatient/outpatient status of the patients, the experience level of the radiologists, and the presence or absence of a lateral radiograph. The Wald test was applied to the derived regression coefficients to determine their significance.

Radiologists were grouped by experience level into 0-5 years post completion of radiology training, 6-10 years, and more than ten years. A likelihood ratio test comparing a binomial logistic regression with categorical radiologist experience against a null model was performed to assess the hypothesis that each of the outcomes (significant changes in reports, management, or imaging recommendation) were associated with experience.

A significance threshold of 0.05 was chosen, with the Benjamini-Hochberg procedure [35] applied to all reported outcomes to account for multiple hypothesis testing. Two clinically qualified researchers independently performed statistical analyses using different software. Calculations were performed in Excel 2016 with RealStatistics resource pack and cross-checked in Python 3.7 using the Pandas 1.0.5 [36], NumPy 1.18.5 [37], SciPy 1.4.1 [38], Scikit-Learn 0.24.0 [39], pymer4 0.7.1 (linked to R 3.4.1, Ime4 1.1.26) [40] and Statsmodels 0.12.1 [41] libraries.

RESULTS

A total of 2,972 cases were reported by 11 radiologists over a period of six weeks. These cases came from 2,665 unique patients (52.7% male), with a median age of 67 (IQR 50–77). Information on radiologist experience, diagnostic/interventional specialty, number of cases reported, source of cases and outcome measures for each radiologist are listed in Table 3.

Table 3 - Demographics and results for the eleven radiologists involved in this study. Percentages (%) represent the associated value as a proportion of the total case number for that radiologist.

Radiologist ID	Number of years post-training	Cases reported (% outpatient)	Interventional?	Report changes (%)	Patient management changes (%)	Imaging recommendations (%)
1	19	136 (21.3)	Yes	1 (0.7)	1 (0.7)	0 (0.0)
2	1	325 (46.2)	No	4 (1.2)	0 (0.0	1 (0.3)
3	4	230 (86.1)	Yes	20 (8.6)	14 (6.1)	10 (4.3)
4	6	375 (22.7)	No	3 (1.0)	0 (0.0)	1 (0.2)
5	4	186 (45.7)	No	22 (11.8)	9 (4.8)	8 (4.3)
6	20	333 (11.1)	No	3 (1.0)	2 (0.6)	1 (0.3)
7	3	312 (48.4)	Yes	15 (4.8)	8 (2.5)	1 (0.3)
8	26	408 (39.7)	No	10 (2.4)	5 (1.2)	4 (1.0)
9	9	214 (43.0)	No	6 (2.8)	2 (0.9)	2 (0.9)
10	6	159 (98.1)	No	1 (0.6)	1 (0.6)	1 (0.6)
11	5	294 (40.1)	No	7 (2.4)	1 (0.3)	0 (0.0)
Total		**2,972**		**92 (3.1)**	**43 (1.4)**	**29 (1.0)**

Of the 2,972 cases, 1,825 (61.4%) cases had lateral (as well as frontal) radiographs available for interpretation. 1,709 (57.5%) cases were from an inpatient setting, and 1,263 (42.5%) from an outpatient setting. The median number of findings per case was five (mean: 5.1, SD: 3.9), with a wide range in the number of findings per case (maximum=20). A total of 364 cases returned zero findings predicted by the model from the complete 124 findings list. 1,526 of the 2,972 cases had one or more critical findings detected by the CXR viewer, with the critical findings in 1,459 (96%) of these cases being confirmed by the radiologist. The number of critical findings per case is summarised in **Error! Reference source not found.**.

Influence of the AI model on radiologist reporting

Across all 2,972 cases, there were 92 cases identified by radiologists as having significant report changes (3.1%), 43 cases of changed patient management (1.4%) and 29 cases of additional imaging recommendations (1.0%) as a result of exposure to the AI model output. When compared to the hypothesised 1% rate of change, the findings were significantly higher for changed reports (p <0.01) and changed patient management (p<0.01), and not significantly different for rate of imaging recommendation (p=0.50).

Agreement with model findings

Of the 2,972 cases, 2,569 had no findings rejected or added by the radiologists, indicating agreement with the model over all 124 possible findings in 86.5% of cases. 306 (10.2%) cases had one finding rejected by the radiologist and 84 (2.8%) had two or more findings rejected by the radiologist. 13 cases (0.5%) had findings (16 in total) added by the radiologists which they deemed were missed by the model, of which 8 were critical findings. These are presented in **Error! Reference source not found.**.

Table 4 - Findings added by the radiologist, and their respective counts. Critical findings are highlighted.

Finding Added	***Count***
Atelectasis	4
Solitary Lung Nodule	3
Cardiac valve prosthesis	2
Solitary Lung Mass	1
Pneumomediastinum	1
Pneumothorax	1
Spinal Wedge Fracture	1
Pulmonary Congestion	1
Peribronchial Thickening	1
Subdiaphragmatic Gas	1

Factors influencing reporting, management, or imaging recommendation

The number of critical findings displayed by the model was significantly higher in cases where there was a change in report, patient management, or imaging recommendation ($p < 0.001$, $p = 0.001$, $p = 0.004$; Table 5). The presence of a lateral projection image in the CXR case interpreted by the model was associated with a significantly greater likelihood of changes to imaging recommendation ($p = 0.005$), but not to the report or patient management ($p = 0.105$ and $p = 0.061$, respectively).

Radiologists with fewer than 5 years consultant experience contributed 1,347 cases, and indicated a rate of 5.0% for significant report change, 2.4% patient management change, and 1.5% recommendations for further imaging. These numbers were higher than for the radiologists with 6-10 years of experience (1.3%, 0.4%, 0.5% respectively over 748 cases) and also for radiologists with greater than 10 years of experience (1.6%, 0.9%, 0.6% over 877 cases). However, a likelihood ratio test applied to binomial logistic regression analysis indicated that the level of radiologist experience did not significantly influence the rate of change in report, patient management, or imaging recommendation ($p = 0.120$, $p = 0.262$, and $p = 0.516$, respectively). Whether a patient was imaged as an inpatient or outpatient was not significantly associated with any change in report, patient management, or imaging recommendation ($p = 0.358$, $p = 0.572$, $p = 0.326$, respectively).

328 *Table 5 - Factors affecting AI model influence on report, patient management, or imaging recommendation. Significance*
329 *testing by the Benjamini-Hochberg algorithm to account for multiple hypotheses. Odds ratios derived from stepwise logistic*
330 *regression coefficients with confidence intervals calculated with Benjamini-adjusted thresholds. Radiologist experience*
331 *analysed as a categorical variable with odds ratios representing effect of changing experience levels from the baseline (0 to*
332 *5 years) to a different level.*

Predictor	Change	Odds Ratios (Adjusted CI)	P Value	Benjamini-Adjusted Threshold	Significance
Number of Critical Findings	Report	1.306 (1.132-1.507)	0	0.0042	YES
Number of Critical Findings	Patient Management	1.267 (1.056-1.521)	0.001	0.0083	YES
Number of Critical Findings	Imaging Recommendation	1.319 (1.035-1.681)	0.004	0.0125	YES
Lateral CXR	Imaging Recommendation	6.495 (1.297-32.530)	0.005	0.0167	YES
Lateral CXR	Patient Management	2.158 (0.837-5.565)	0.061	0.0208	NO
Lateral CXR	Report	1.542 (0.848-2.805)	0.105	0.025	NO
Radiologist Experience	Report	0 to 5 years: Baseline 6 to 10 years: 0.255 (0.043-1.521) > 10 years: 0.305 (0.065-1.439)	0.120	0.0292	NO
Radiologist Experience	Patient Management	0 to 5 years: Baseline 6 to 10 years: 0.165 (0.009-3.214) > 10 years: 0.378 (0.054-2.654)	0.262	0.0333	NO
Radiologist Experience	Imaging Recommendation	0 to 5 years: Baseline 6 to 10 years: 0.357 (0.034-3.783) > 10 years: 0.380 (0.044-3.287)	0.516	0.0458	NO
Inpatient/Outpatient	Imaging Recommendation	1.550 (0.613-3.919)	0.326	0.0375	NO
Inpatient/Outpatient	Report	0.794 (0.476-1.323)	0.358	0.0417	NO
Inpatient/Outpatient	Patient Management	0.818 (0.408-1.640)	0.572	0.0500	NO

333

334 **Survey Results**

335 The post-study survey was completed by 10 out of the 11 radiologists (Figure 3 and Figure 4).
336 Notably, 70% of participants felt that their reporting time was slightly worse, however when asked how
337 satisfied they were with their reporting time, 70% indicated that they were satisfied.

Ninety percent of radiologists responded that their reporting accuracy was improved while using the CXR viewer and 90% of participants were satisfied with accuracy of the CXR model's findings. Ninety percent of radiologists demonstrated an improved attitude towards the use of the AI diagnostic viewer by the end of the study and 90% demonstrated an improved attitude towards AI in general. No radiologists reported a more negative attitude towards the CXR viewer or towards AI in general.

DISCUSSION

We have previously shown that using the output of this comprehensive deep learning model improved radiologist diagnostic accuracy [31] in a non-clinical setting, but it is important to demonstrate that this improvement translates into meaningful change in a real-world environment. In this multicentre real-world prospective study, we determined how often the finding recommendations of the comprehensive deep learning model led to a material change in the radiologist's report, a change in the patient management recommendation, or a change in the subsequent imaging recommendation. To the authors' knowledge, this is the first time that the impact of a comprehensive deep learning model developed to detect radiological findings on CXR has been studied in a real-world reporting environment. Other commercially available deep learning models able to detect multiple findings on CXR have been studied in the non-clinical setting, yielding encouraging results and outperforming physicians in the detection of major thoracic findings [42] as well as improving resident diagnostic sensitivity [43]. Other models have demonstrated diagnostic accuracy that is comparable to that of test radiologists [44].

We showed that radiologists agreed with all findings identified by the AI model in 86.5% of cases on a per case basis. Notably, there was a significant change to the report in 3.1% of cases leading to changes in recommended patient management in 1.4% of cases, and changes to imaging recommendations in 1% of cases. Of note, two lung lesions that were flagged by the model, but missed by radiologists, led to additional imaging and changed management and were subsequently diagnosed as lung carcinoma, highlighting the real-world value of integrating this type of system into the radiology workflow.

The significant impact of the CXR viewer on radiologist reporting and recommendations did however come at the cost of false positives, with 13% of cases having one or more model findings rejected by the radiologist. When this false positive rate is compared against the false positive rates per case reported in other studies investigating CXR models, which range from 14 – 88% [14,45,46], it is considered an acceptable value. Furthermore, these studies report false-positive rates for CXR models which only detect lung nodules, while the current study this represents the false positive rate across 124 findings. In addition, this trade-off appears to be reasonable to the participating radiologists, who reported a high level of satisfaction with the model.

In this study, analysis of radiologists by experience level using logistic regression found no significant relationship between experience level and increased changes to reports, patient management changes, or imaging recommendations as a result of the model. Statistical analysis of the relationship between experience level and change in report was associated with a p value of 0.12, suggesting that, with further research, a significant relationship may be identified. It is expected that the inclusion of a larger group of radiologists may lead to a significant finding, as the association between experience and level of change has been noted in other studies. For example Jang et al., showed that less experienced radiologists benefited the most from the diagnostic assistance in detecting lung nodules on CXR [14]. The primary factor that influenced the likelihood of the model findings leading to a change in the report was the presence of critical findings in the model's recommendation. This is particularly notable because it indicates that the changes to the report are significant. They did not simply involve the inclusion of additional non-critical findings in the report, which may be interpreted as overestimating the impact of the model. The inpatient or outpatient status of a case was found not to significantly affect the likelihood of significant changes to the radiologists' report, to patient management, or to imaging recommendations.

The post-study survey provided further insight into the impact that the CXR viewer had on participant reporting, in addition to the level of agreement and changes to the radiology report and patient management recommendations outlined above. The first notable response was that the CXR viewer may have negatively affected reporting times (albeit only mildly) for the majority of radiologists. This outcome was expected in this study setting because the radiologists were taking additional time to provide feedback on the model's recommendations for each case. Previous studies that surveyed radiologists reported that 74.4% thought AI would lower the interpretation time [47]. It is notable that even with the negative impact the model had on reporting time, the majority of radiologists (70%) were still satisfied with reporting time while using the CXR viewer, suggesting that the diagnostic improvements offered by the model were enough to offset the additional perceived reporting time. Additional insight from the survey suggested that very little training was required before radiologists felt comfortable using the tool. This is useful as education on AI has been a primary concern amongst clinicians, as a large proportion of radiologists report having little knowledge of AI [48].

Limitations and future research

The results presented in this study are self-reported by participating radiologists and are likely an underestimation of the model's actual impact. It is expected that radiologists would not report every instance in which they made an interpretive error. Another limitation is that there was no objective gold standard against which the radiologist and model interpretation could be measured. This is a small-scale study involving a limited sample size, conducted over several weeks. As a result, it lacks the statistical power to examine the benefit of the model on a finding-by-finding basis. In future, it would be beneficial to conduct a similar study with a larger sample size to allow for more powerful statistical analysis and examination of specific finding changes. Another useful next step would be to include a gold standard to determine the ground truth for the CXR findings, as this would prevent any under reporting which may occur with self-reported results, as well as enable the detection of false negatives as a result of the CXR viewer.

Conclusion

The present study indicated that the integration of a comprehensive AI model capable of detecting 124 findings on CXR into a radiology workflow led to significant changes in reports and patient management, with an acceptable rate of additional imaging recommendations. These results were not affected by the inpatient status of the patient, and although approaching significance, the experience level of the radiologists did not significantly relate to the primary endpoint outcomes. In secondary endpoint outcomes, the model output showed good agreement with radiologists, and radiologists showed high rates of satisfaction with their reporting times and diagnostic accuracy when using the CXR viewer as a diagnostic assist device. Results highlight the usefulness of AI-driven decision support tools in improving clinical practice and patient outcomes.

AUTHOR STATEMENT

CJ contributed to conception and design of the work, acquisition of data, analysis and visualisation of data, interpretation of data, drafting of the work, and project management. LD contributed to design of the work and acquisition of data. MM contributed to conception and design of the work, interpretation and visualisation of data, development of diagrams, drafting of the work, and project management. CT and JS contributed to analysis and visualisation of data, interpretation of data, development of diagrams, and drafting of the work. LO, AJ, QB and NE contributed to interpretation of data. All authors revised the work critically for important intellectual content, gave final approval of the version to be published, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

ACKNOWLEDGEMENTS

The authors would like to thank Mark Wilson, Marc Northrop, Nicolaus Carr and Trina Shnier for their valuable contributions to designing and managing the study.

COMPETING INTERESTS

CJ is a radiologist employed by the radiology practice and a clinical consultant for Annalise-AI. LD, LO and NE are independent of Annalise-AI and have no interests to declare. MM, JS, CT, AJ and QB are employed by or seconded to Annalise-AI. Study conception, study design, ethics approval and data security were conducted independent of Annalise-AI.

FUNDING STATEMENT

This work was supported by Annalise-AI Pty Ltd grant number N/A. Annalise-AI supported this work through free provision of the model to participating radiologists for the duration of the study and financing of an external biostatistician.

PATIENT AND PUBLIC INVOLVEMENT

Patients and public were not involved in the design, conduct, or reporting of this study.

DATA AVAILABILITY STATEMENT

All data relevant to the study are included in the article or uploaded as online supplemental information. No additional data are available.

References

1 Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 2016;**278**:563–77. doi:10.1148/radiol.2015151169

2 Greene R. Francis H. Williams, MD: father of chest radiology in North America. *RadioGraphics* 1991;**11**:325–32. doi:10.1148/radiographics.11.2.2028067

3 Schaefer-Prokop C, Neitzel U, Venema HW, *et al.* Digital chest radiography: an update on modern technology, dose containment and control of image quality. *Eur Radiol* 2008;**18**:1818–30. doi:10.1007/s00330-008-0948-3

4 Lee CS, Nagy PG, Weaver SJ, *et al.* Cognitive and System Factors Contributing to Diagnostic Errors in Radiology. *American Journal of Roentgenology* 2013;**201**:611–7. doi:10.2214/AJR.12.10375

5 Chotas HG, Ravin CE. Chest radiography: estimated lung volume and projected area obscured by the heart, mediastinum, and diaphragm. *Radiology* 1994;**193**:403–4. doi:10.1148/radiology.193.2.7972752

6 Berlin L. Accuracy of Diagnostic Procedures: Has It Improved Over the Past Five Decades? *American Journal of Roentgenology* 2007;**188**:1173–8. doi:10.2214/AJR.06.1270

7 Zaorsky NG, Churilla TM, Egleston BL, *et al.* Causes of death among cancer patients. *Annals of Oncology* 2017;**28**:400–7. doi:10.1093/annonc/mdw604

8 del Ciello A, Franchi P, Contegiacomo A, *et al.* Missed lung cancer: when, where, and why? *Diagn Interv Radiol* 2017;**23**:118–26. doi:10.5152/dir.2016.16187

9 Fazal MI, Patel ME, Tye J, *et al.* The past, present and future role of artificial intelligence in imaging. *European Journal of Radiology* 2018;**105**:246–50. doi:10.1016/j.ejrad.2018.06.020

10 Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science* 2015;**349**:255–60. doi:10.1126/science.aaa8415

11 Hosny A, Parmar C, Quackenbush J, *et al.* Artificial intelligence in radiology. *Nat Rev Cancer* 2018;**18**:500–10. doi:10.1038/s41568-018-0016-5

12 Erickson BJ, Korfiatis P, Akkus Z, *et al.* Machine Learning for Medical Imaging. *RadioGraphics* 2017;**37**:505–15. doi:10.1148/rg.2017160130

13 Esteva A, Chou K, Yeung S, *et al.* Deep learning-enabled medical computer vision. *npj Digital Medicine* 2021;**4**:1–9. doi:10.1038/s41746-020-00376-2

14 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. doi:10.1148/radiol.2020200165

15 Liang C-H, Liu Y-C, Wu M-T, *et al.* Identifying pulmonary nodules or masses on chest radiography using deep learning: external validation and strategies to improve clinical practice. *Clinical Radiology* 2020;**75**:38–45. doi:10.1016/j.crad.2019.08.005

16 Hurt B, Kligerman S, Hsiao A. Deep Learning Localization of Pneumonia: 2019 Coronavirus (COVID-19) Outbreak. *J Thorac Imaging* 2020;**35**:W87–9.

17 Kim JY, Choe PG, Oh Y, *et al.* The First Case of 2019 Novel Coronavirus Pneumonia Imported into Korea from Wuhan, China: Implication for Infection Prevention and Control Measures. *J Korean Med Sci* 2020;**35**. doi:10.3346/jkms.2020.35.e61

18 Bassi PRAS, Attux R. A Deep Convolutional Neural Network for COVID-19 Detection Using Chest X-Rays. *arXiv:200501578 [cs, eess]* Published Online First: 12 January 2021.http://arxiv.org/abs/2005.01578 (accessed 23 Mar 2021).

19 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. *Investigative Radiology* 2020;**55**:792–8. doi:10.1097/RLI.0000000000000707

20 Sze-To A, Wang Z. tCheXNet: Detecting Pneumothorax on Chest X-Ray Images Using Deep Transfer Learning. In: Karray F, Campilho A, Yu A, eds. *Image Analysis and Recognition*. Cham: : Springer International Publishing 2019. 325–32. doi:10.1007/978-3-030-27272-2_28

21 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. doi:10.1007/s00330-020-06771-3

22 Park S, Lee SM, Kim N, *et al.* Application of deep learning–based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy. *Eur Radiol* 2019;**29**:5341–8. doi:10.1007/s00330-019-06130-x

23 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. doi:10.1136/oemed-2019-106386

24 S Z, X Z, R Z. Identifying Cardiomegaly in ChestX-ray8 Using Transfer Learning. *Stud Health Technol Inform* 2019;**264**:482–6. doi:10.3233/shti190268

25 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. doi:10.1371/journal.pone.0236378

26 Pasa F, Golkov V, Pfeiffer F, *et al.* Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. *Scientific Reports* 2019;**9**:6268. doi:10.1038/s41598-019-42557-4

27 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. *Scientific Reports* 2020;**10**:210. doi:10.1038/s41598-019-56589-3

28 Heo S-J, Kim Y, Yun S, *et al.* Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers' Health Examination Data. *International Journal of Environmental Research and Public Health* 2019;**16**:250. doi:10.3390/ijerph16020250

29 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. *Scientific Reports* 2019;**9**:15000. doi:10.1038/s41598-019-51503-3

30 Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology* 2017;**284**:574–82. doi:10.1148/radiol.2017162326

31 Seah J, Tang C, Buchlak QD, *et al.* Radiologist chest X-ray diagnostic accuracy performance improvements when augmented by a comprehensive deep learning model. *In press* 2021.

32 Annalise.ai - Annalise CXR comprehensive medical imaging AI. Annalise.ai. https://annalise.ai/products/annalise-cxr/ (accessed 23 Mar 2021).

33 Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv:190511946 [cs, stat]* Published Online First: 11 September 2020.http://arxiv.org/abs/1905.11946 (accessed 30 Mar 2021).

34 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:150504597 [cs]* Published Online First: 18 May 2015.http://arxiv.org/abs/1505.04597 (accessed 30 Mar 2021).

35 Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)* 1995;**57**:289–300.

36 Mckinney W. pandas: a Foundational Python Library for Data Analysis and Statistics. *Python High Performance Science Computer* 2011.

37 Harris CR, Millman KJ, van der Walt SJ, *et al.* Array programming with NumPy. *Nature* 2020;**585**:357–62. doi:10.1038/s41586-020-2649-2

38 Jones E, Oliphant T, Peterson P. SciPy: Open Source Scientific Tools for Python. 2001.

39 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* Published Online First: 12 October 2011.https://hal.inria.fr/hal-00650905 (accessed 23 Mar 2021).

40 Jolly E. Pymer4: Connecting R and Python for linear mixed modeling. *Journal of Open Source Software* 2018;**3**:862.

41 Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. Austin, Texas: 2010. 92–6. doi:10.25080/Majora-92bf1922-011

42 Hwang EJ, Park S, Jin K-N, *et al.* Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. *JAMA Netw Open* 2019;**2**:e191095. doi:10.1001/jamanetworkopen.2019.1095

43 Hwang EJ, Nam JG, Lim WH, *et al.* Deep Learning for Chest Radiograph Diagnosis in the Emergency Department. *Radiology* 2019;**293**:573–80. doi:10.1148/radiol.2019191225

44 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. doi:10.1371/journal.pone.0204155

45 Dellios N, Teichgraeber U, Chelaru R, *et al.* Computer-aided Detection Fidelity of Pulmonary Nodules in Chest Radiograph. *J Clin Imaging Sci* 2017;**7**. doi:10.4103/jcis.JCIS_75_16

46 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* Published Online First: 12 November 2019. doi:10.1148/radiol.2019182465

47 Waymel Q, Badr S, Demondion X, *et al.* Impact of the rise of artificial intelligence in radiology: What do radiologists think? *Diagnostic and Interventional Imaging* 2019;**100**:327–36. doi:10.1016/j.diii.2019.03.015

48 Collado-Mesa F, Alvarez E, Arheart K. The Role of Artificial Intelligence in Diagnostic Radiology: A Survey at a Single Radiology Residency Training Program. *Journal of the American College of Radiology* 2018;**15**:1753–7. doi:10.1016/j.jacr.2017.12.021

FIGURE LEGENDS

Figure 1 - Flow diagram illustrating the AI-assisted reporting process described in this study. (RIS: Radiological information system)

Figure 2 - Counts of numbers of critical findings for the cases seen by the radiologist, defined as the number of critical findings agreed + the number of critical findings added. The number of cases which returned zero findings was 1,513.

Figure 3 – Diverging stacked bar chart depicting the first set of radiologist survey responses.

Figure 4 – Diverging stacked bar chart visualising the second set of survey responses of the radiologists.

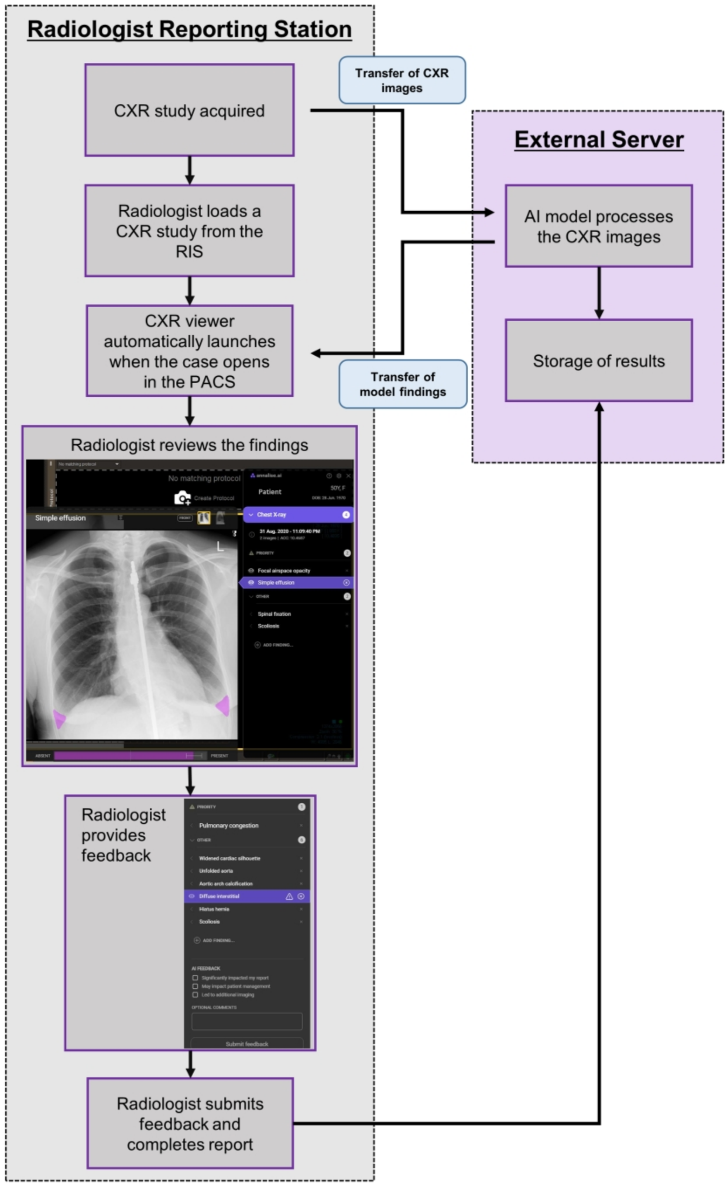

Figure 1 - Flow diagram illustrating the AI-assisted reporting process described in this study. RIS: Radiological information system.

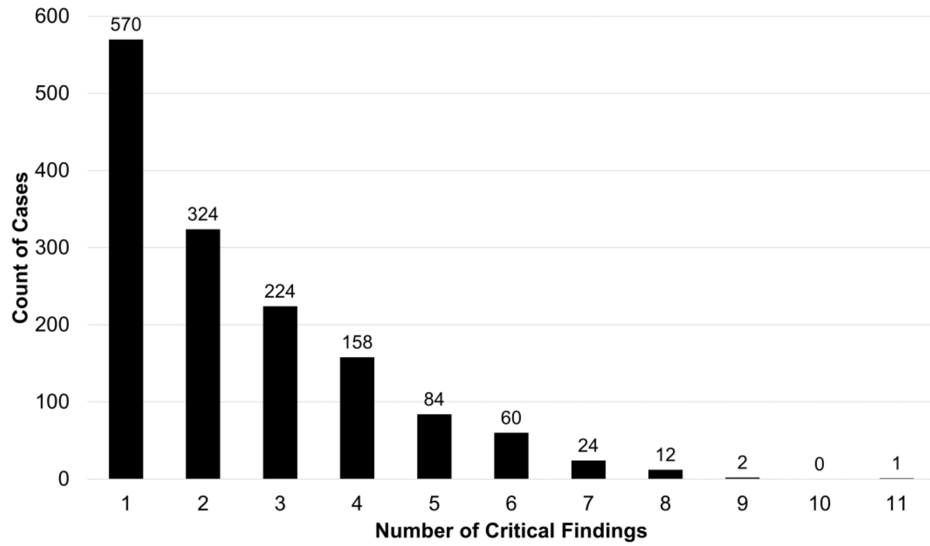

Figure 2 - Counts of numbers of critical findings for the cases seen by the radiologist, defined as the number of critical findings agreed + the number of critical findings added. The number of cases which returned zero findings was 1,513.

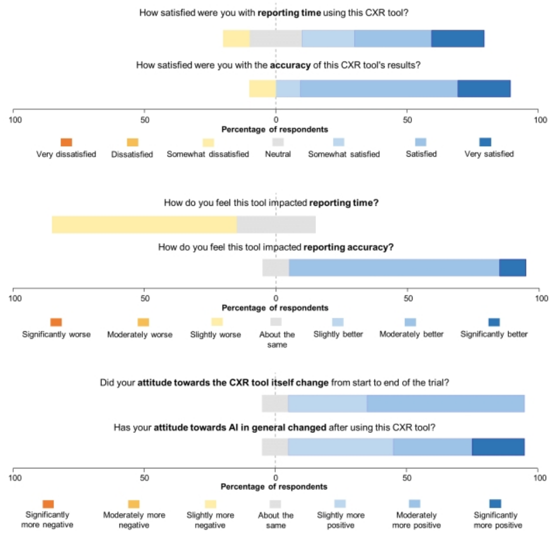

Figure 3 – Diverging stacked bar chart depicting the first set of radiologist survey responses.

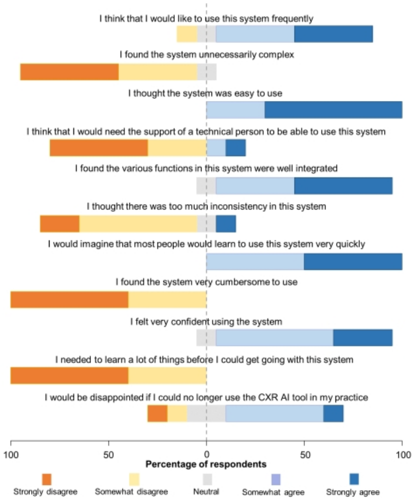

Figure 4 – Diverging stacked bar chart visualising the second set of survey responses of the radiologists.

Supplementary Table 1 - List of the 124 findings, including 34 critical findings which the model is validated to detect. ETT: endotracheal tube, NGT: nasogastric tube, PAC: pulmonary artery catheter.

	Critical Clinical Findings	
Acute humerus fracture	Loculated effusion	Subcutaneous emphysema
Acute rib fracture	Lung collapse	Subdiaphragmatic gas
Air Space Opacity - Multifocal	Multiple masses or nodules	Suboptimal central line
Cavitating mass with content	Perihilar airspace opacity	Suboptimal ETT
Cavitating mass(es)	Pneumomediastinum	Suboptimal NGT
Diffuse airspace opacity	Pulmonary congestion	Suboptimal PAC
Diffuse lower airspace opacity	Segmental collapse	Superior mediastinal mass
Diffuse upper airspace opacity	Shoulder dislocation	Tension pneumothorax
Focal airspace opacity	Simple effusion	Tracheal deviation
Hilar lymphadenopathy	Simple pneumothorax	Widened aortic contour
Inferior mediastinal mass	Solitary lung mass	Widened cardiac silhouette
	Solitary lung nodule	
	Non-Critical Clinical Findings	
Abdominal Clips	Coronary Stent	Pectus Excavatum
Acute Clavicle Fracture	Diaphragmatic Elevation	Peribronchial Cuffing
Airway Stent	Diaphragmatic Eventration	Pericardial Fat Pad
Aortic Arch Calcification	Diffuse Fibrotic Volume Loss	Pleural Mass
Aortic Stent	Diffuse Interstitial	Post Resection Volume Loss
Atelectasis	Diffuse Nodular / Miliary Lesions	Pulmonary Arterial Catheter
Axillary Clips	Diffuse Pleural Thickening	Pulmonary Artery Enlargement
Basal Predominant Interstitial	Diffuse Spinal Osteophytes	Reduced Lung Markings
Biliary Stent	Distended Bowel	Rib Fixation
Breast Implant	Electronic Cardiac Devices	Rib Lesion
Bronchiectasis	Endotracheal Tube	Rib Resection
Bullae Diffuse	Gallstones	Rotator Cuff Anchor
Bullae Lower	Gastric Band	Scapular Fracture
Bullae Upper	Hiatus Hernia	Scapular Lesion
Calcified Axillary Nodes	Humeral Lesion	Scoliosis
Calcified Granuloma (<5mm)	Intercostal Drain	Shoulder Arthritis
Calcified Hilar Lymphadenopathy	Internal Foreign Body	Shoulder Fixation
Calcified Mass (>5mm)	Kyphosis	Shoulder Replacement
Calcified Neck Nodes	Lower Zone Fibrotic Volume Loss	Spinal Fixation
Calcified Pleural Plaques	Lung Sutures	Spine Arthritis
Cardiac Valve Prosthesis	Mastectomy	Spine Lesion
Central Venous Catheter	Mediastinal Clips	Spine Wedge Fracture
Cervical Flexion	Nasogastric Tube	Sternotomy Wires
Chronic Clavicle Fracture	Neck Clips	Suboptimal Gastric Band
Chronic Humerus Fracture	Nipple Shadow	Unfolded Aorta
Chronic Rib Fracture	Oesophageal Stent	Upper Predominant Interstitial
Clavicle Fixation	Osteopaenia	Upper Zone Fibrotic Volume Loss
Clavicle Lesion	Pectus Carinatum	

Technical Findings		
Chest Incompletely Imaged	Image Obscured	Underexposed
Hyperinflation	Overexposed	Underinflation
	Patient Rotation	

Supplementary Table 2 – Example of the survey questions provided to the radiologists at the end of the study.

	Significantly worse	Moderately worse	Slightly worse	About the same	Slightly better	Moderately better	Significantly better
How do you feel this tool impacted **reporting time**?	O	O	O	O	O	O	O
How do you feel this tool impacted **reporting accuracy**?	O	O	O	O	O	O	O
	Very dissatisfied	Dissatisfied	Somewhat dissatisfied	Neutral	Somewhat satisfied	Satisfied	Very dissatisfied
How satisfied were you with **reporting time** using this CXR tool?	O	O	O	O	O	O	O
How satisfied were you with the **accuracy** of this CXR tool's results?	O	O	O	O	O	O	O
	Significantly more negative	Moderately more negative	Slightly more negative	About the same	Slightly more positive	Moderately more negative	Significantly more negative
Did your **attitude towards the CXR tool itself** change from start to end of the trial?	O	O	O	O	O	O	O
Has your **attitude towards AI in general changed** after using this CXR tool?	O	O	O	O	O	O	O
		Strongly disagree	Somewhat disagree	Neutral	Somewhat agree	Strongly agree	
I think that I would like to use this system frequently.		O	O	O	O	O	
I found the system unnecessarily complex.		O	O	O	O	O	
I thought the system was easy to use.		O	O	O	O	O	
I think that I would need the support of a technical person to be able to use this system.		O	O	O	O	O	
I found the various functions in this system were well integrated.		O	O	O	O	O	
I thought there was too much inconsistency in this system.		O	O	O	O	O	
I would imagine that most people would learn to use this system very quickly.		O	O	O	O	O	
I found the system very cumbersome to use.		O	O	O	O	O	
I felt very confident using the system.		O	O	O	O	O	
I needed to learn a lot of things before I could get going with this system.		O	O	O	O	O	

I would be disappointed if I could no longer use the CXR AI tool in my practice.	o	o	o	o	o

CLAIM: Checklist for Artificial Intelligence in Medical Imaging

Section / Topic	No.	Item	
TITLE / ABSTRACT			
	1	Identification as a study of AI methodology, specifying the category of technology used (e.g., deep learning)	**Yes**
	2	Structured summary of study design, methods, results, and conclusions	**Yes**
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the AI approach	**Yes – page 4/5**
	4	Study objectives and hypotheses	**Yes – page 5**
METHODS			
Study Design	5	Prospective or retrospective study	**Yes – page 8** (under: "CXR case section")
	6	Study goal, such as model creation, exploratory study, feasibility study, non-inferiority trial	**Yes – page 8** (under: "CXR case section")
Data	7	Data sources	**Yes – page 8** (under: "CXR case section")
	8	Eligibility criteria: how, where, and when potentially eligible participants or studies were identified (e.g., symptoms, results from previous tests, inclusion in registry, patient-care setting, location, dates)	**Yes – page 8** (under: "CXR case section")
	9	Data pre-processing steps	**N/A**
	10	Selection of data subsets, if applicable	**N/A**
	11	Definitions of data elements, with references to Common Data Elements	**Yes – page 8/9** (under: "AI-assisted reporting)
	12	De-identification methods	**Yes – page 8** (under: "CXR case section")
	13	How missing data were handled	**N/A**
Ground Truth	14	Definition of ground truth reference standard, in sufficient detail to allow replication	**Yes – page 6** (under: "model development and validation")
	15	Rationale for choosing the reference standard (if alternatives exist)	**N/A**
	16	Source of ground-truth annotations; qualifications and preparation of annotators	**N/A –** Described in reference 31
	17	Annotation tools	**N/A –** Described in reference 31
	18	Measurement of inter- and intrarater variability; methods to mitigate variability and/or resolve discrepancies	**N/A –** Described in reference 31

Data Partitions	**19**	Intended sample size and how it was determined	**Yes – page 10** (under: "statistics and data analysis")
	20	How data were assigned to partitions; specify proportions	**N/A**
	21	Level at which partitions are disjoint (e.g., image, study, patient, institution)	**N/A**
Model	**22**	Detailed description of model, including inputs, outputs, all intermediate layers and connections	**Yes – page 6** (under: "model development and validation") **and described in reference 31**
	23	Software libraries, frameworks, and packages	**Yes – page 6** (under: "model development and validation") **and described in reference 31**
	24	Initialization of model parameters (e.g., randomization, transfer learning)	**Yes – page 6** (under: "model development and validation") **and described in reference 31**
Training	**25**	Details of training approach, including data augmentation, hyperparameters, number of models trained	**Yes – page 6** (under: "model development and validation") **and described in reference 31**
	26	Method of selecting the final model	**N/A**
	27	Ensembling techniques, if applicable	**N/A**
Evaluation	**28**	Metrics of model performance	**Yes – page 6** (under: "model development and validation") **and described in reference 31**
	29	Statistical measures of significance and uncertainty (e.g., confidence intervals)	**Yes – page 6** (under: "model development and validation") **and described in reference 31**
	30	Robustness or sensitivity analysis	**N/A**
	31	Methods for explainability or interpretability (e.g., saliency maps), and how they were validated	**N/A**
	32	Validation or testing on external data	**N/A**
RESULTS			
Data	**33**	Flow of participants or cases, using a diagram to indicate inclusion and exclusion	**Yes – Figure 1**
	34	Demographic and clinical characteristics of cases in each partition	**N/A**
Model performance	**35**	Performance metrics for optimal model(s) on all data partitions	**N/A**
	36	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	**N/A**
	37	Failure analysis of incorrectly classified cases	**N/A**
DISCUSSION			
	38	Study limitations, including potential bias, statistical uncertainty, and generalizability	**Yes – page 13** (under: " limitations and future research")

	39	Implications for practice, including the intended use and/or clinical role	**Yes – page 13** (under: "conclusion")
OTHER INFORMATION			
	40	Registration number and name of registry	**N/A**
	41	Where the full study protocol can be accessed	**N/A**
	42	Sources of funding and other support; role of funders	**Yes – page 21**

Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020; 2(2):e200029. https://doi.org/10.1148/ryai.2020200029

BMJ Open

Assessment of the effect of a comprehensive chest radiograph deep learning model on radiologist reports and patient outcomes: a real-world observational study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-052902.R1
Article Type:	Original research
Date Submitted by the Author:	08-Sep-2021
Complete List of Authors:	Jones, Catherine; Annalise-AI; I-Med Network Danaher, Luke; I-Med Network Milne, Michael; Annalise-AI; I-Med Network Tang, Cyril; Annalise-AI Seah, Jarrel; Alfred Health, Radiology; Annalise AI, Oakden-Rayner, Luke; The University of Adelaide, Australian Institute for Machine Learning Johnson, Andrew; Annalise-AI Buchlak, Quinlan; Annalise-AI; The University of Notre Dame Australia School of Medicine Sydney Campus Esmaili, Nazanin; The University of Notre Dame Australia School of Medicine Sydney Campus; University of Technology Sydney
Primary Subject Heading:	Radiology and imaging
Secondary Subject Heading:	Emergency medicine, Radiology and imaging
Keywords:	Chest imaging < RADIOLOGY & IMAGING, RADIOLOGY & IMAGING, Diagnostic radiology < RADIOLOGY & IMAGING

SCHOLARONE™ Manuscripts

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Assessment of the effect of a comprehensive chest radiograph deep learning model on radiologist reports and patient outcomes: a real-world observational study

Catherine M Jones[1,2], Luke Danaher[2], Michael R Milne[1,2*], Cyril Tang[1], Jarrel Seah[1,3], Luke Oakden-Rayner[4], Andrew Johnson[1], Quinlan D Buchlak[1,5], Nazanin Esmaili[5,6]

[1]Annalise-AI, Sydney, NSW, Australia
[2]I-MED Radiology Network, Sydney, NSW, Australia
[3]Department of Radiology, Alfred Health, Melbourne, VIC, Australia
[4]Australian Institute for Machine Learning, The University of Adelaide, Adelaide, SA, Australia
[5]School of Medicine, University of Notre Dame Australia, Sydney, NSW, Australia
[6]Faculty of Engineering and IT, University of Technology Sydney, Ultimo, NSW, Australia

*Correspondence to: michael.milne@annalise.ai

Corresponding author:
Name: Michael Milne
Annalise-AI
Sydney, Australia
E-mail: michael.milne@annalise.ai

Keywords: Machine learning; chest X-ray, deep learning.

Word Count: 4,426

ABSTRACT

Objectives: AI algorithms have been developed to detect imaging features on chest X-ray (CXR) with a comprehensive AI model capable of detecting 124 CXR findings being recently developed. The aim of this study was to evaluate the real-world usefulness of the model as a diagnostic assistance device for radiologists.

Design: This prospective real-world multicentre study involved a group of radiologists using the model in their daily reporting workflow to report consecutive chest X-rays and recording their feedback on level of agreement with the model findings and whether this significantly affected their reporting.

Setting: The study took place at radiology clinics and hospitals within a large radiology network in Australia between November and December 2020.

Participants: Eleven consultant diagnostic radiologists of varying levels of experience participated in this study.

Primary and secondary outcome measures: Proportion of CXR cases where use of the AI model led to significant material changes to the radiologist report, to patient management, or to imaging recommendations. Additionally, level of agreement between radiologists and the model findings, and radiologist attitudes towards the model were assessed.

Results: Of 2,972 cases reviewed with the model, 92 cases (3.1%) had significant report changes, 43 cases (1.4%) had changed patient management and 29 cases (1.0%) had further imaging recommendations. In terms of agreement with the model, 2,572 cases showed complete agreement (86.5%). 390 (13%) cases had one or more findings rejected by the radiologist. There were 16 findings across 13 cases (0.5%) deemed to be missed by the model. Nine out of 10 radiologists felt their accuracy was improved with the model and were more positive towards AI post-study.

Conclusions: Use of an AI model in a real-world reporting environment significantly improved radiologist reporting and showed good agreement with radiologists, highlighting the potential for AI diagnostic support to improve clinical practice.

ARTICLE SUMMARY

Strengths and limitations of this study

- This study substantially adds to the limited literature on real-world evaluation of comprehensive CXR AI models in radiology workflow.
- This was a multicentre study conducted across a mix of public hospitals, private hospitals, and community clinic settings.
- Due to the design of the study, diagnostic accuracy of the decision support system was not a measurable outcome.
- Results of this study are self-reported and may therefore be prone to bias.
- Determination of the significance of report changes due to the model's recommendations was made at the discretion of each radiologist on a case-by-case basis.

INTRODUCTION

Radiology is a data-rich medical specialty and is well placed to embrace artificial intelligence [1].This is especially true in high volume imaging tasks such as chest X-ray imaging. The rapid application of X-ray technology to diagnosing chest diseases at the end of the 19th century led to the chest X-ray (CXR) becoming a first-line diagnostic imaging tool [2] and it remains an essential component of the diagnostic pathway for chest disease. Due to advancements in digital image acquisition, low ionising radiation dose and low cost, the chest radiograph is more easily accessible worldwide than any other imaging modality [3].

The challenges of interpreting CXR, however, have not lessened over the last half-century. CXR images are 2D representations of complex 3D structures, relying on soft tissue contrast between structures of different densities. Multiple overlapping structures lead to reduced visibility of both normal and abnormal structures [4], with up to 40% of the lung parenchyma obscured by overlying ribs and the mediastinum [5]. This can be further exacerbated by other factors including the degree of inspiration, other devices in the field of view, and patient positioning. In addition, there is a wide range of pathology in the chest which is visible to varying degrees on the CXR. These factors combine to make CXRs difficult to accurately interpret, with an error rate of 20-50% for CXRs containing radiographic evidence of disease reported in the literature [6]. Notably, lung cancer is one of the most common cancers worldwide and is the most common cause of cancer death [7], and CXR interpretation error accounts for 90% of cases where lung cancer is missed [8]. Despite technological advancements in CXR over the past 50 years, this level of diagnostic error has remained constant [6].

A rapidly developing field attempting to assist radiologists in radiological interpretation involves the application of machine learning, in particular deep neural networks [9]. Deep neural networks learn patterns in large, complex datasets, enabling the detection of subtle features and outcome prediction [10,11]. The potential of these algorithms has grown rapidly in the past decade thanks to the development of more useful neural network models, advancements in computational power, and an increase in the volume and availability of digital imaging datasets [11]. Of note is the rise of convolutional neural networks (CNNs), a type of deep neural network that excels at image feature extraction and classification, and demonstrates strong performance in medical image analysis, leading to the rapid advancement of computer vision in medical imaging [12,13]. CNNs have been used to develop models to successfully detect targeted clinical findings on CXR, including lung cancer [14,15], pneumonia [16,17], COVID-19 [18], pneumothorax [19–22], pneumoconiosis [23], cardiomegaly [24], pulmonary hypertension [25] and tuberculosis [26–30]. These studies highlight the effectiveness of applied machine learning in CXR interpretation, however most of these deep learning systems are limited in scope to a single finding or a small set of findings, therefore lacking the broad utility that would make them useful in clinical practice.

Recently, our group developed a comprehensive deep learning CXR diagnostic assist device, which was designed to assist clinicians in CXR interpretation and improve diagnostic accuracy, validated for 124 clinically relevant findings seen on frontal and lateral chest radiographs [31]. The primary objective of the current study was to evaluate the real-world usefulness of the model as a diagnostic assist device for radiologists in both hospital and community clinic settings. This involved examining the frequency at which the model's recommendations led to a ‘significant impact on the report’, defined as the inclusion of findings recommended by the model which altered the radiologists report in a meaningful way. The frequency of change in patient management and recommendations for further imaging were also evaluated. Secondary endpoints included: (1) investigating agreement between radiologists and the findings detected by the model; and (2) assessing radiologist attitudes towards the tool and AI models in general.

METHODS

Ethics Statement

This study was approved by the institutional human research ethics committee of the Wesley Hospital, Brisbane, Queensland Australia (2020.14.324). Written informed consent was obtained from each participating radiologist. The requirement of patient consent was waived by the ethics committee due to the low-risk nature of the study.

Model development and validation

A modified version of a commercially available AI tool for use as a diagnostic assist device displaying results within a viewer (CXR viewer; Annalise CXR ver 1.2, Annalise-AI, Sydney, Australia) was evaluated [32]. The AI tool deploys an underlying machine learning model, developed and validated by Seah et al [31], which consists of attribute and classification CNNs based on the EfficientNet architecture [33] and a segmentation CNN based on U-Net [34] with EfficientNet backbone. The model was trained on 821,681 de-identified CXR images from 284,649 patients originating from inpatient, outpatient and emergency settings across Australia, Europe, and North America. Training dataset labelling involved independent triple labelling of all images by three radiologists selected from a wider pool of 120 consultant radiologists (none of whom were employed by the radiology network involved in this current study). The model was validated for 124 clinical findings in a multi-reader, multi-case (MRMC) study [31]. Thirty-four of these findings were deemed priority findings based on their clinical importance. The full list of 124 findings is available in Supplementary Table 1. Ground truth labels for the validation study dataset were determined by a consensus of three independent radiologists drawn from a pool of seven fully credentialed subspecialty thoracic radiologists. The algorithm is publicly available at https://cxrdemo.annalise.ai. The AI model was used in line with pre-existing regulatory approval [35].

Technical Integration

Prior to the start of the study, technical integration of the software into existing radiology practice systems and testing occurred over several weeks. First, an integration adapter was installed on the IT network of each radiology clinic and acted as a gateway between the internal IT infrastructure and the AI model. Auto-routing rules were established ensuring only CXR studies were forwarded to the integration adapter from the picture archiving and communication system (PACS). Following a successful testing period, the Annalise CXR viewer was installed and configured on workstations for the group of study radiologists.

Study Participants

Eleven consultant radiologists working for a large Australian radiology network were invited to participate in the study through their local radiologist network. This group included general diagnostic radiologists who had completed specialist radiology training and passed all diagnostic radiology college examinations required for consultant accreditation in Australia. All radiologists reported the minimum of 2000 chest radiographs per year (either within the radiology network or through other institutions) suggested to maintain competency [36]. No subspecialist chest radiologists were included.

The group included radiologists with a range of experience levels: five radiologists had 0–5 years post-training experience, three radiologists had 6–10 years of experience, and three radiologists had more than 10 years of experience. Radiologists were situated across four states in Australia and worked in public hospitals, private hospitals and community clinic settings. Both on site and remote reporting was included, in line with regular workflow. Prior to study commencement, each radiologist attended a training seminar and a one-on-one training session to fully understand the CXR viewer and its features. In addition, the participating radiologists were able to familiarise themselves with the viewer prior to commencement of data collection.

CXR Case Selection

In this multicentre real-world prospective study, all consecutive chest radiographs reported by the radiologists originating from inpatient, outpatient, and emergency settings were included for a period covering nearly six weeks. The CXR cases were reported with the assistance of the AI tool in real-world clinical practice, using high resolution diagnostic radiology monitors within the radiologists' normal reporting environment. As per usual workflow across a large radiology network spanning a geographically large area with many regional and remote clinics, both on-site and remote reporting of CXR cases was undertaken. A total of 106 sites contributed cases with case numbers varying from one case up to a maximum of 271 cases at the busiest site.

At least one frontal chest radiograph was required for analysis by the model, and cases that did not include at least one were excluded. Chest radiographs from patients aged younger than 16 years were excluded. Data from all sources was de-identified for analysis.

AI-Assisted Reporting

For each CXR case, radiologists produced their clinical report with access to clinical information, the referral and available patient history, in line with the normal workflow. The AI model analyses the CXR image(s) for each case but does not incorporate clinical inputs (such as previous imaging, referral information or patient demographic data) into the analysis. Model output was displayed to the radiologist in a user interface, linked to the image in the PACS, automatically launching when a CXR case was opened (Figure 1).

A modified version of the commercially available AI software was employed for this study, which incorporated changes into the user interface to allow radiologists to provide feedback on model recommendations. No changes were made to the underlying model. An example of the modified model user interface is presented in figure 2. For each case, the model provided a list of suggested findings, listed as “priority” or “other”, along with a confidence indicator. For a subset of findings, a region of interest localiser was overlayed on the image and the model indicated whether the finding was on the left or the right side, or both (see Supplementary Table 1). The CXR viewer was configured to display its findings after the radiologists’ initial read of the case. For each case, radiologists were asked to review the CXR viewer’s findings and provide feedback within the viewer. The options presented to the radiologists in the viewer are listed in Table 1.

Table 1 - List of review options presented to the radiologist with each case.

REVIEW OPTION	DESCRIPTION
Rejected clinical finding	A model-detected finding disputed by the radiologist
Missed clinical finding	A model-detected finding missed by the radiologist
Add additional findings	Finding(s) identified by the radiologist but not identified by the model
These findings significantly impacted my report	A yes/no binary question relating to the effect of the model output on the radiologist report
These findings may impact patient management	A yes/no binary question relating to the effect of the model output on patient management, as perceived by the reporting radiologist
These findings led to additional imaging recommendations	A binary yes/no question related to whether the radiologist recommended further imaging based on the model output

The outcome measure of ‘significant impact on the report’ was the primary outcome measure. A significant change was described as the inclusion of findings recommended by the model, which altered the radiologists report in a meaningful way. As this varied by patient and clinical setting, it was left to the discretion of the radiologist. During the analysis of radiologist feedback, it was assumed that a change in patient management or further imaging recommendation would not occur without radiologists indicating a material change in the CXR report, and thus management and imaging questions were dependent on a significant change in the report. This was also patient-specific; for example, missing a pneumothorax in a ventilated patient with known pneumothorax would not have the same impact on patient management as a previously unknown pneumothorax in an outpatient. Free text input describing missed findings or other relevant data were manually added after data collection was complete.

No formal adjudication of cases showing discrepancy between radiologist and model interpretation was performed. The study was not designed as a diagnostic accuracy validation. No review or ground truthing process was performed. Radiologists remained responsible for image interpretation and formulation of the report.

Post-Study Survey

Upon completion of data collection, a post-study survey was distributed to all participating radiologists to obtain feedback on the usefulness of the CXR viewer and how it affected their opinion of AI in radiology. A table of the survey questions is presented in Supplementary Table 2.

Statistics and Data Analysis

A 1% rate of significant changes in reports (the primary outcome measure) was deemed to be clinically significant prior to commencing the study. Based on estimations of the prevalence of missed critical findings on CXR, preliminary power calculations estimated that the number of cases required to detect at least a 1% rate of significant changes in reports was approximately 2000 cases in total, with alpha value 0.05 and desired power of 0.90. To account for any dropout in radiologists or cases, a target of 3000 cases was set for the study. Ten radiologists were recruited, with an eleventh included for any unexpected participant drop out and to achieve this target in a reasonable time period.

A two-tailed binomial test was used to test the hypothesis that the rate of significant report change, patient management change, or imaging recommendation change was at least 1%. To ensure that the sampling of CXRs reasonably approximated a random snapshot of the true population, radiologists in various states, experience levels as well as different conditions of practice (community clinic vs hospital based) were selected. Additionally, the study was conducted prospectively which further aligned the structure of the sampled data with the expected structure of the population, justifying the choice of analysing the sample using a binomial test without adjustment for each radiologist.

Multivariate logistic regression using generalised linear mixed effect analysis was used to assess the effect of several possible confounders on the measured outcomes, including the number of critical clinical findings per case identified by the model, the inpatient/outpatient status of the patients, the experience level of the radiologists, and the presence or absence of a lateral radiograph. The Wald test was applied to the derived regression coefficients to determine their significance.

Radiologists were grouped by experience level into 0-5 years post completion of radiology training, 6-10 years, and more than ten years. A likelihood ratio test comparing a binomial logistic regression with categorical radiologist experience against a null model was performed to assess the hypothesis that the outcomes (significant changes in reports, management, or imaging recommendation) were associated with experience.

A significance threshold of 0.05 was chosen, with the Benjamini-Hochberg procedure [37] applied to all reported outcomes to account for multiple hypothesis testing. Two clinically qualified researchers independently performed statistical analyses using different software. Calculations were performed in Excel 2016 with RealStatistics resource pack and cross-checked in Python 3.7 using the Pandas 1.0.5 [38], NumPy 1.18.5 [39], SciPy 1.4.1 [40], Scikit-Learn 0.24.0 [41], pymer4 0.7.1 (linked to R 3.4.1, Ime4 1.1.26) [42] and Statsmodels 0.12.1 [43] libraries.

RESULTS

A total of 2,972 cases were reported by 11 radiologists over a period of six weeks. These cases came from 2,665 unique patients (52.7% male), with a median age of 67 (IQR 50–77). Information on radiologist experience, number of cases reported, source of cases and outcome measures for each radiologist are listed in Table 2.

Table 2 - Demographics and results for the eleven radiologists involved in this study. Percentages (%) represent the associated value as a proportion of the total case number for that radiologist.

Radiologist ID	Number of years post-training	Cases reported (% outpatient)	Significant report impact (%)	Patient management changes (%)	Imaging reco mmendations (%)
1	19	136 (21.3)	1 (0.7)	1 (0.7)	0 (0.0)
2	1	325 (46.2)	4 (1.2)	0 (0.0	1 (0.3)
3	4	230 (86.1)	20 (8.6)	14 (6.1)	10 (4.3)
4	6	375 (22.7)	3 (1.0)	0 (0.0)	1 (0.2)
5	4	186 (45.7)	22 (11.8)	9 (4.8)	8 (4.3)
6	20	333 (11.1)	3 (1.0)	2 (0.6)	1 (0.3)
7	3	312 (48.4)	15 (4.8)	8 (2.5)	1 (0.3)
8	26	408 (39.7)	10 (2.4)	5 (1.2)	4 (1.0)
9	9	214 (43.0)	6 (2.8)	2 (0.9)	2 (0.9)
10	6	159 (98.1)	1 (0.6)	1 (0.6)	1 (0.6)
11	5	294 (40.1)	7 (2.4)	1 (0.3)	0 (0.0)
Total		**2,972**	**92 (3.1)**	**43 (1.4)**	**29 (1.0)**

Of the 2,972 cases, 1,825 (61.4%) cases had lateral (as well as frontal) radiographs available for interpretation. 1,709 (57.5%) cases were from an inpatient setting, and 1,263 (42.5%) from an outpatient setting. The median number of findings per case was five (mean: 5.1, SD: 3.9), with a wide range in the number of findings per case (maximum=20). A total of 364 cases returned zero findings predicted by the model from the complete 124 findings list. 1,526 of the 2,972 cases had one or more critical findings detected by the CXR viewer, with the critical findings in 1,459 (96%) of these cases being confirmed by the radiologist. The number of critical findings per case is summarised in Figure 3.

Influence of the AI model on radiologist reporting

Across all 2,972 cases, there were 92 cases identified by radiologists as having significant report changes (3.1%), 43 cases of changed patient management (1.4%) and 29 cases of additional imaging recommendations (1.0%) as a result of exposure to the AI model output. When compared to the hypothesised 1% rate of change, the findings were significantly higher for changed reports (p <0.01) and changed patient management (p<0.01), and not significantly different for rate of imaging recommendation (p=0.50).

Agreement with model findings

Of the 2,972 cases, 2,569 had no findings rejected or added by the radiologists, indicating agreement with the model over all 124 possible findings in 86.5% of cases. 306 (10.2%) cases had one finding rejected by the radiologist and 84 (2.8%) had two or more findings rejected by the radiologist. 202 (5.3%) critical findings detected by the model were rejected by radiologists. The missed and rejected critical findings are detailed in Table 3.

13 cases (0.5%) had findings (16 in total) added by the radiologists which they deemed were missed by the model, of which 8 were critical findings (see Table 3). The remaining 8 non-critical missed findings were atelectasis (4 findings), cardiac valve prosthesis (2 findings), spinal wedge fracture (1 finding) and peribronchial thickening (1 finding).

Table 3 – Breakdown of the critical findings detected by the model and the level of radiologist agreement with each, including the number of findings reportedly missed by the model (and added by the radiologist) or missed by the radiologist. Percentages (%) represent the associated value as a proportion of the total number of findings displayed by the model.

Critical Finding	Displayed by model	Radiologist agreed with finding (%)	Radiologist rejected finding (%)	Added in by radiologist	Missed by radiologist
Acute aortic syndrome	2	2.0 (100.0)	0 (0.0)	0	0
Acute humerus fracture	5	5 (100.0)	0 (0.0)	0	0
Acute rib fracture	54	39 (72.2)	15 (27.8)	0	5
Cardiomegaly	1,008	979 (97.1)	29 (2.9)	0	0
Cavitating mass	14	13 (92.9)	1 (7.1)	0	0
Cavitating mass internal content	6	5 (83.3)	1 (16.7)	0	0
Diffuse airspace opacity	13	13 (100.0)	0 (0.0)	0	0
Diffuse lower airspace opacity	153	148 (96.7)	5 (3.3)	0	0
Diffuse perihilar airspace opacity	45	45 (100.0)	0 (0.0)	0	0
Diffuse upper airspace opacity	2	2 (100.0)	0 (0.0)	0	0
Focal airspace opacity	341	321 (94.1)	20 (5.9)	0	2
Hilar lymphadenopathy	8	6 (75.0)	2 (25.0)	0	0
Inferior mediastinal mass	8	7 (87.5)	1 (12.5)	0	0
Loculated effusion	87	80 (92.0)	7 (8.0)	0	1
Lung collapse	11	10 (90.9)	1 (9.1)	0	0
Malpositioned CVC	85	78 (91.8)	7 (8.2)	0	1
Malpositioned ETT	52	43 (82.7)	9 (17.3)	0	0
Malpositioned NGT	39	31 (79.5)	8 (20.5)	0	0
Malpositioned PAC	13	9 (69.2)	4 (30.8)	0	0
Multifocal airspace opacity	125	120 (96.0)	5 (4.0)	0	1
Multiple pulmonary masses	43	38 (88.4)	5 (11.6)	0	0
Pneumomediastinum	5	5 (100.0)	0 (0.0)	1	0
Pulmonary congestion	220	215 (97.7)	5 (2.3)	1	0
Segmental collapse	292	290 (99.3)	2 (0.7)	0	1
Shoulder dislocation	1	0 (0.0)	1 (100.0)	0	0
Simple effusion	687	650 (94.6)	37 (5.4)	0	1
Simple pneumothorax	90	77 (85.6)	13 (14.4)	1	1
Single pulmonary mass	41	38 (92.7)	3 (7.3)	1	1
Single pulmonary nodule	105	95 (90.5)	10 (9.5)	3	5
Subcutaneous emphysema	53	51 (96.2)	2 (3.8)	0	1
Subdiaphragmatic gas	7	7 (100.0)	0 (0.0)	1	0
Superior mediastinal mass	37	32 (86.5)	5 (13.5)	0	0
Tension pneumothorax	11	7 (63.6)	4 (36.4)	0	0
Tracheal deviation	133	133 (100.0)	0 (0.0)	0	0
Total	3,796	3,594 (94.7)	202 (5.3)	8	20

Factors influencing reporting, management, or imaging recommendation

The number of critical findings displayed by the model was significantly higher in cases where there was a change in report, patient management, or imaging recommendation ($p < 0.001$, $p = 0.001$, $p = 0.004$; Table 4). The presence of a lateral projection image in the CXR case interpreted by the model was associated with a significantly greater likelihood of changes to imaging recommendation ($p = 0.005$), but not to the report or patient management ($p = 0.105$ and $p = 0.061$, respectively).

Radiologists with fewer than 5 years consultant experience contributed 1,347 cases, and indicated a rate of 5.0% for significant report change, 2.4% patient management change, and 1.5% recommendations for further imaging. These numbers were higher than for the radiologists with 6-10 years of experience (1.3%, 0.4%, 0.5% respectively over 748 cases) and also for radiologists with greater than 10 years of experience (1.6%, 0.9%, 0.6% over 877 cases). However, a likelihood ratio test applied to binomial logistic regression analysis indicated that the level of radiologist experience did not significantly influence the rate of change in report, patient management, or imaging recommendation ($p = 0.120$, $p = 0.262$, and $p = 0.516$, respectively). Whether a patient was imaged as an inpatient or outpatient was not significantly associated with any change in report, patient management, or imaging recommendation ($p = 0.358$, $p = 0.572$, $p = 0.326$, respectively).

Table 4 - Factors affecting AI model influence on report, patient management, or imaging recommendation. Significance testing by the Benjamini-Hochberg algorithm to account for multiple hypotheses. Odds ratios derived from stepwise logistic regression coefficients with confidence intervals calculated with Benjamini-adjusted thresholds. Radiologist experience analysed as a categorical variable with odds ratios representing effect of changing experience levels from the baseline (0 to 5 years) to a different level.

Predictor	Change	Odds Ratios (Adjusted CI)	P Value	Benjamini-Adjusted Threshold	Significance
Number of Critical Findings	Report	1.306 (1.132-1.507)	0	0.0042	YES
Number of Critical Findings	Patient Management	1.267 (1.056-1.521)	0.001	0.0083	YES
Number of Critical Findings	Imaging Recommendation	1.319 (1.035-1.681)	0.004	0.0125	YES
Lateral CXR	Imaging Recommendation	6.495 (1.297-32.530)	0.005	0.0167	YES
Lateral CXR	Patient Management	2.158 (0.837-5.565)	0.061	0.0208	NO
Lateral CXR	Report	1.542 (0.848-2.805)	0.105	0.025	NO
Radiologist Experience	Report	0 to 5 years: Baseline 6 to 10 years: 0.255 (0.043-1.521) > 10 years: 0.305 (0.065-1.439)	0.120	0.0292	NO
Radiologist Experience	Patient Management	0 to 5 years: Baseline 6 to 10 years: 0.165 (0.009-3.214) > 10 years: 0.378 (0.054-2.654)	0.262	0.0333	NO
Radiologist Experience	Imaging Recommendation	0 to 5 years: Baseline 6 to 10 years: 0.357 (0.034-3.783) > 10 years: 0.380 (0.044-3.287)	0.516	0.0458	NO
Inpatient/Outpatient	Imaging Recommendation	1.550 (0.613-3.919)	0.326	0.0375	NO
Inpatient/Outpatient	Report	0.794 (0.476-1.323)	0.358	0.0417	NO
Inpatient/Outpatient	Patient Management	0.818 (0.408-1.640)	0.572	0.0500	NO

Survey Results

The post-study survey was completed by ten out of the eleven radiologists (Figure 4 and Figure 5). Notably, 7 (70%) participants felt that their reporting time was slightly worse, however when asked how satisfied they were with their reporting time, 7 (70%) indicated that they were satisfied.

Nine out of ten radiologists responded that their reporting accuracy was improved while using the CXR viewer, with nine out of ten (90%) participants being satisfied with accuracy of the CXR model's findings. Nine radiologists (90%) demonstrated an improved attitude towards the use of the AI diagnostic viewer by the end of the study and 9 (90%) demonstrated an improved attitude towards AI in general. No radiologists reported a more negative attitude towards the CXR viewer or towards AI in general.

DISCUSSION

We have previously shown that using the output of this comprehensive deep learning model improved radiologist diagnostic accuracy [44] in a non-clinical setting, but it is important to demonstrate that this improvement translates into meaningful change in a real-world environment. In this multicentre real-world prospective study, we determined how often the finding recommendations of the comprehensive deep learning model led to a material change in the radiologist's report, a change in the patient management recommendation, or a change in subsequent imaging recommendation. To the authors' knowledge, this is the first time that the impact of a comprehensive deep learning model developed to detect radiological findings on CXR has been studied in a real-world reporting environment. Other commercially available deep learning models able to detect multiple findings on CXR have been studied in the non-clinical setting, yielding encouraging results and outperforming physicians in the detection of major thoracic findings [45] as well as improving resident diagnostic sensitivity [46]. Other models have demonstrated diagnostic accuracy that is comparable to that of test radiologists [47]. Additionally, studies have yielded promising results for the use of models in population screening, particularly for tuberculosis, where several models have met the minimum WHO recommendations for tuberculosis triage tests [29,48].

We showed that radiologists agreed with all findings identified by the AI model in 86.5% of cases on a per case basis, while on a per finding basis, agreed with the critical findings identified by the model on 94.7% of findings. Notably, there was a significant change to the report in 3.1% of cases leading to changes in recommended patient management in 1.4% of cases, and changes to imaging recommendations in 1% of cases. Of note, 146 lung lesions (solitary lung nodule and solitary lung mass) were present in the dataset according to the model. Two lung lesions flagged by the model but missed by radiologists were recommended for additional imaging and changed management, subsequently diagnosed as lung carcinoma, highlighting the real-world value of integrating this type of system into the radiology workflow. However, four findings of lung nodule were flagged by the radiologists as missed by the model, indicating that the model alone is not intended to replace radiologist interpretation.

The significant impact of the CXR viewer on radiologist reporting and recommendations did however come at the cost of false positives, with 13% of cases having one or more model findings rejected by the radiologist. When this false positive rate is compared against the false positive rates per case reported in other studies investigating CXR models, which range from 14 – 88% [14,49,50], it is considered acceptable. Furthermore, these studies report false-positive rates for CXR models that only detect lung nodules, while in the current study this represents the false positive rate across 124 findings. Notably, on a per finding basis, only 5.3% of critical findings detected by the model were rejected by the radiologist. However, there were several outliers in the critical findings group that had noticeably higher rates of rejection, including acute rib fracture, hilar lymphadenopathy, malpositioned NGT/PAC, shoulder dislocation and tension pneumothorax. Several explanations for this are low sample size, the subjectivity of diagnosis and heightened model sensitivity at the expense of specificity. Overall, this trade-off appears to be reasonable to the participating radiologists, who reported a high level of satisfaction with the model.

In this study, analysis of radiologists by experience level using logistic regression found no statistically significant relationship between experience level and increased changes to reports, patient management changes, or imaging recommendations as a result of the model. Statistical analysis of the relationship between experience level and change in report was associated with a p value of 0.12, suggesting that, with further research, a significant relationship may be identified. It is expected that the inclusion of a larger group of radiologists may lead to a significant finding, as the association between experience and level of change has been noted in other studies. For example Jang et al., showed that less experienced radiologists benefited the most from the diagnostic assistance in detecting lung nodules on CXR [14]. In this study, three of the 11 radiologists contributed a higher than average incidence of the primary outcome of report change, and these were all less experienced radiologists compared to the cohort average experience level. Whilst this may be due to variations in individual radiologist interpretation of 'significant report change', the consistency of experience level across these three radiologists suggests a relationship with experience level and tool impact.

The primary factor that influenced the likelihood of the model findings leading to a change in the report was the presence of critical findings in the model's recommendation. This is particularly notable because it indicates that the changes to the report are significant. They did not simply involve the inclusion of additional non-critical findings in the report, which may be interpreted as overestimating the impact of the model. The inpatient or outpatient status of a case was found not to significantly affect the likelihood of significant changes to the radiologists' report, to patient management, or to imaging recommendations.

The post-study survey provided further insight into the impact that the CXR viewer had on participant reporting, in addition to the level of agreement and changes to the radiology report and patient management recommendations outlined above. The first notable response was that the CXR viewer may have negatively affected reporting times (albeit only mildly) for the majority of radiologists. This outcome was expected in this study setting because the radiologists were taking additional time to provide feedback on the model's recommendations for each case. Previous studies that surveyed radiologists reported that 74.4% thought AI would lower the interpretation time [51]. It is notable that even with the negative impact the model had on reporting time, the majority of radiologists (70%) were still satisfied with reporting time while using the CXR viewer, suggesting that the diagnostic improvements offered by the model were enough to offset the additional perceived reporting time. Additional insight from the survey suggested that very little training was required before radiologists felt comfortable using the tool. This is useful as education on AI has been a primary concern amongst clinicians, as a large proportion of radiologists report having little knowledge of AI [52].

Limitations and future research

The results presented in this study are self-reported by participating radiologists and are likely an underestimation of the model's actual impact. It is expected that radiologists would not report every instance in which they made an interpretive error. Another limitation is that there was no objective gold standard against which the radiologist and model interpretation could be measured. This is a small-scale study involving a limited sample size, conducted over several weeks. As a result, it lacks the statistical power to examine the benefit of the model on a finding-by-finding basis. In future, it would be beneficial to conduct a similar study with a larger sample size to allow for more powerful statistical analysis and examination of specific finding changes. Another useful next step would be to include a gold standard to determine the ground truth for the CXR findings, as this would prevent any under reporting which may occur with self-reported results, as well as enable the detection of false negatives as a result of the CXR viewer.

Although none of the cases evaluated in this study had been seen by the model previously, we note that one of the five data sources used for model training originated from the same radiology network. This therefore cannot be considered as true external evaluation. Further work in truly external institutions in the future are welcomed.

Conclusion

The present study indicated that the integration of a comprehensive AI model capable of detecting 124 findings on CXR into a radiology workflow led to significant changes in reports and patient management, with an acceptable rate of additional imaging recommendations. These results were not affected by the inpatient status of the patient, and although approaching significance, the experience level of the radiologists did not significantly relate to the primary endpoint outcomes. In secondary endpoint outcomes, the model output showed good agreement with radiologists, and radiologists showed high rates of satisfaction with their reporting times and diagnostic accuracy when using the CXR viewer as a diagnostic assist device. Results highlight the usefulness of AI-driven diagnostic assist tools in improving clinical practice and patient outcomes.

AUTHOR STATEMENT

CJ contributed to conception and design of the work, acquisition of data, analysis and visualisation of data, interpretation of data, drafting of the work, and project management. LD contributed to design of the work and acquisition of data. MM contributed to conception and design of the work, interpretation and visualisation of data, development of diagrams, drafting of the work, and project management. CT and JS contributed to analysis and visualisation of data, interpretation of data, development of diagrams, and drafting of the work. LO, AJ, QB and NE contributed to interpretation of data. All authors revised the work critically for important intellectual content, gave final approval of the version to be published, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

ACKNOWLEDGEMENTS

The authors would like to thank Mark Wilson, Marc Northrop, Nicolaus Carr and Trina Shnier for their valuable contributions to designing and managing the study.

COMPETING INTERESTS

CJ is a radiologist employed by the radiology practice and a clinical consultant for Annalise-AI. LD, LO and NE are independent of Annalise-AI and have no interests to declare. MM, JS, CT, AJ and QB are employed by or seconded to Annalise-AI. Study conception, study design, ethics approval and data security were conducted independent of Annalise-AI.

FUNDING STATEMENT

This work was supported by Annalise-AI Pty Ltd. Annalise-AI supported this work through free provision of the model to participating radiologists for the duration of the study and financing of an external biostatistician.

PATIENT AND PUBLIC INVOLVEMENT

Patients and public were not involved in the design, conduct, or reporting of this study.

DATA AVAILABILITY STATEMENT

All data relevant to the study are included in the article or uploaded as online supplemental information. No additional data are available.

References

1 Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 2016;**278**:563–77. doi:10.1148/radiol.2015151169

2 Greene R. Francis H. Williams, MD: father of chest radiology in North America. *RadioGraphics* 1991;**11**:325–32. doi:10.1148/radiographics.11.2.2028067

3 Schaefer-Prokop C, Neitzel U, Venema HW, *et al.* Digital chest radiography: an update on modern technology, dose containment and control of image quality. *Eur Radiol* 2008;**18**:1818–30. doi:10.1007/s00330-008-0948-3

4 Lee CS, Nagy PG, Weaver SJ, *et al.* Cognitive and System Factors Contributing to Diagnostic Errors in Radiology. *American Journal of Roentgenology* 2013;**201**:611–7. doi:10.2214/AJR.12.10375

5 Chotas HG, Ravin CE. Chest radiography: estimated lung volume and projected area obscured by the heart, mediastinum, and diaphragm. *Radiology* 1994;**193**:403–4. doi:10.1148/radiology.193.2.7972752

6 Berlin L. Accuracy of Diagnostic Procedures: Has It Improved Over the Past Five Decades? *American Journal of Roentgenology* 2007;**188**:1173–8. doi:10.2214/AJR.06.1270

7 Zaorsky NG, Churilla TM, Egleston BL, *et al.* Causes of death among cancer patients. *Annals of Oncology* 2017;**28**:400–7. doi:10.1093/annonc/mdw604

8 del Ciello A, Franchi P, Contegiacomo A, *et al.* Missed lung cancer: when, where, and why? *Diagn Interv Radiol* 2017;**23**:118–26. doi:10.5152/dir.2016.16187

9 Fazal MI, Patel ME, Tye J, *et al.* The past, present and future role of artificial intelligence in imaging. *European Journal of Radiology* 2018;**105**:246–50. doi:10.1016/j.ejrad.2018.06.020

10 Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science* 2015;**349**:255–60. doi:10.1126/science.aaa8415

11 Hosny A, Parmar C, Quackenbush J, *et al.* Artificial intelligence in radiology. *Nat Rev Cancer* 2018;**18**:500–10. doi:10.1038/s41568-018-0016-5

12 Erickson BJ, Korfiatis P, Akkus Z, *et al.* Machine Learning for Medical Imaging. *RadioGraphics* 2017;**37**:505–15. doi:10.1148/rg.2017160130

13 Esteva A, Chou K, Yeung S, *et al.* Deep learning-enabled medical computer vision. *npj Digital Medicine* 2021;**4**:1–9. doi:10.1038/s41746-020-00376-2

14 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. doi:10.1148/radiol.2020200165

15 Liang C-H, Liu Y-C, Wu M-T, *et al.* Identifying pulmonary nodules or masses on chest radiography using deep learning: external validation and strategies to improve clinical practice. *Clinical Radiology* 2020;**75**:38–45. doi:10.1016/j.crad.2019.08.005

16 Hurt B, Kligerman S, Hsiao A. Deep Learning Localization of Pneumonia: 2019 Coronavirus (COVID-19) Outbreak. *J Thorac Imaging* 2020;**35**:W87–9.

17 Kim JY, Choe PG, Oh Y, *et al.* The First Case of 2019 Novel Coronavirus Pneumonia Imported into Korea from Wuhan, China: Implication for Infection Prevention and Control Measures. *J Korean Med Sci* 2020;**35**. doi:10.3346/jkms.2020.35.e61

18 Bassi PRAS, Attux R. A Deep Convolutional Neural Network for COVID-19 Detection Using Chest X-Rays. *arXiv:200501578 [cs, eess]* Published Online First: 12 January 2021.http://arxiv.org/abs/2005.01578 (accessed 23 Mar 2021).

19 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. *Investigative Radiology* 2020;**55**:792–8. doi:10.1097/RLI.0000000000000707

20 Sze-To A, Wang Z. tCheXNet: Detecting Pneumothorax on Chest X-Ray Images Using Deep Transfer Learning. In: Karray F, Campilho A, Yu A, eds. *Image Analysis and Recognition*. Cham: : Springer International Publishing 2019. 325–32. doi:10.1007/978-3-030-27272-2_28

21 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. doi:10.1007/s00330-020-06771-3

22 Park S, Lee SM, Kim N, *et al.* Application of deep learning–based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy. *Eur Radiol* 2019;**29**:5341–8. doi:10.1007/s00330-019-06130-x

23 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. doi:10.1136/oemed-2019-106386

24 S Z, X Z, R Z. Identifying Cardiomegaly in ChestX-ray8 Using Transfer Learning. *Stud Health Technol Inform* 2019;**264**:482–6. doi:10.3233/shti190268

25 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. doi:10.1371/journal.pone.0236378

26 Pasa F, Golkov V, Pfeiffer F, *et al.* Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. *Scientific Reports* 2019;**9**:6268. doi:10.1038/s41598-019-42557-4

27 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. *Scientific Reports* 2020;**10**:210. doi:10.1038/s41598-019-56589-3

28 Heo S-J, Kim Y, Yun S, *et al.* Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers' Health Examination Data. *International Journal of Environmental Research and Public Health* 2019;**16**:250. doi:10.3390/ijerph16020250

29 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. *Scientific Reports* 2019;**9**:15000. doi:10.1038/s41598-019-51503-3

30 Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology* 2017;**284**:574–82. doi:10.1148/radiol.2017162326

31 Seah JCY, Tang CHM, Buchlak QD, *et al.* Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *The Lancet Digital Health* 2021;**3**:e496–506. doi:10.1016/S2589-7500(21)00106-0

32 Annalise.ai - Annalise CXR comprehensive medical imaging AI. Annalise.ai. https://annalise.ai/products/annalise-cxr/ (accessed 23 Mar 2021).

33 Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv:190511946 [cs, stat]* Published Online First: 11 September 2020.http://arxiv.org/abs/1905.11946 (accessed 30 Mar 2021).

34 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:150504597 [cs]* Published Online First: 18 May 2015.http://arxiv.org/abs/1505.04597 (accessed 30 Mar 2021).

35 xmlmillr6.pdf. https://www.ebs.tga.gov.au/servlet/xmlmillr6?dbid=ebs/PublicHTML/pdfStore.nsf&docid=F7ADAEBB76CEDD47CA2585E500424A43&agid=(PrintDetailsPublic)&actionid=1 (accessed 25 Aug 2021).

36 ace_lung_pathways_final_report_v1.4.pdf. https://www.cancerresearchuk.org/sites/default/files/ace_lung_pathways_final_report_v1.4.pdf (accessed 31 Aug 2021).

37 Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)* 1995;**57**:289–300.

38 Mckinney W. pandas: a Foundational Python Library for Data Analysis and Statistics. *Python High Performance Science Computer* 2011.

39 Harris CR, Millman KJ, van der Walt SJ, *et al.* Array programming with NumPy. *Nature* 2020;**585**:357–62. doi:10.1038/s41586-020-2649-2

40 Jones E, Oliphant T, Peterson P. SciPy: Open Source Scientific Tools for Python. 2001.

41 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* Published Online First: 12 October 2011.https://hal.inria.fr/hal-00650905 (accessed 23 Mar 2021).

42 Jolly E. Pymer4: Connecting R and Python for linear mixed modeling. *Journal of Open Source Software* 2018;**3**:862.

43 Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. Austin, Texas: 2010. 92–6. doi:10.25080/Majora-92bf1922-011

44 Seah J, Tang C, Buchlak QD, *et al.* Radiologist chest X-ray diagnostic accuracy performance improvements when augmented by a comprehensive deep learning model. *The Lancet Digital Health* 2021.

45 Hwang EJ, Park S, Jin K-N, *et al.* Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. *JAMA Netw Open* 2019;**2**:e191095. doi:10.1001/jamanetworkopen.2019.1095

46 Hwang EJ, Nam JG, Lim WH, *et al.* Deep Learning for Chest Radiograph Diagnosis in the Emergency Department. *Radiology* 2019;**293**:573–80. doi:10.1148/radiol.2019191225

47 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. doi:10.1371/journal.pone.0204155

48 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. *The Lancet Digital Health* 2020;**2**:e573–81. doi:10.1016/S2589-7500(20)30221-1

49 Dellios N, Teichgraeber U, Chelaru R, *et al.* Computer-aided Detection Fidelity of Pulmonary Nodules in Chest Radiograph. *J Clin Imaging Sci* 2017;**7**. doi:10.4103/jcis.JCIS_75_16

50 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* Published Online First: 12 November 2019. doi:10.1148/radiol.2019182465

51 Waymel Q, Badr S, Demondion X, *et al.* Impact of the rise of artificial intelligence in radiology: What do radiologists think? *Diagnostic and Interventional Imaging* 2019;**100**:327–36. doi:10.1016/j.diii.2019.03.015

52 Collado-Mesa F, Alvarez E, Arheart K. The Role of Artificial Intelligence in Diagnostic Radiology: A Survey at a Single Radiology Residency Training Program. *Journal of the American College of Radiology* 2018;**15**:1753–7. doi:10.1016/j.jacr.2017.12.021

FIGURE LEGENDS

Figure 1 – Flow diagram illustrating the AI-assisted reporting process described in this study. (RIS: Radiological information system)

Figure 2 – Example of the modified user interface used by the participating radiologists in this study. The red box highlights the feedback options added to the interface for this study.

Figure 3 – Counts of numbers of critical findings for the cases seen by the radiologist, defined as the number of critical findings agreed + the number of critical findings added. The number of cases which returned zero findings was 1,513.

Figure 4 – Diverging stacked bar chart depicting the first set of radiologist survey responses.

Figure 5 – Diverging stacked bar chart visualising the second set of survey responses of the radiologists.

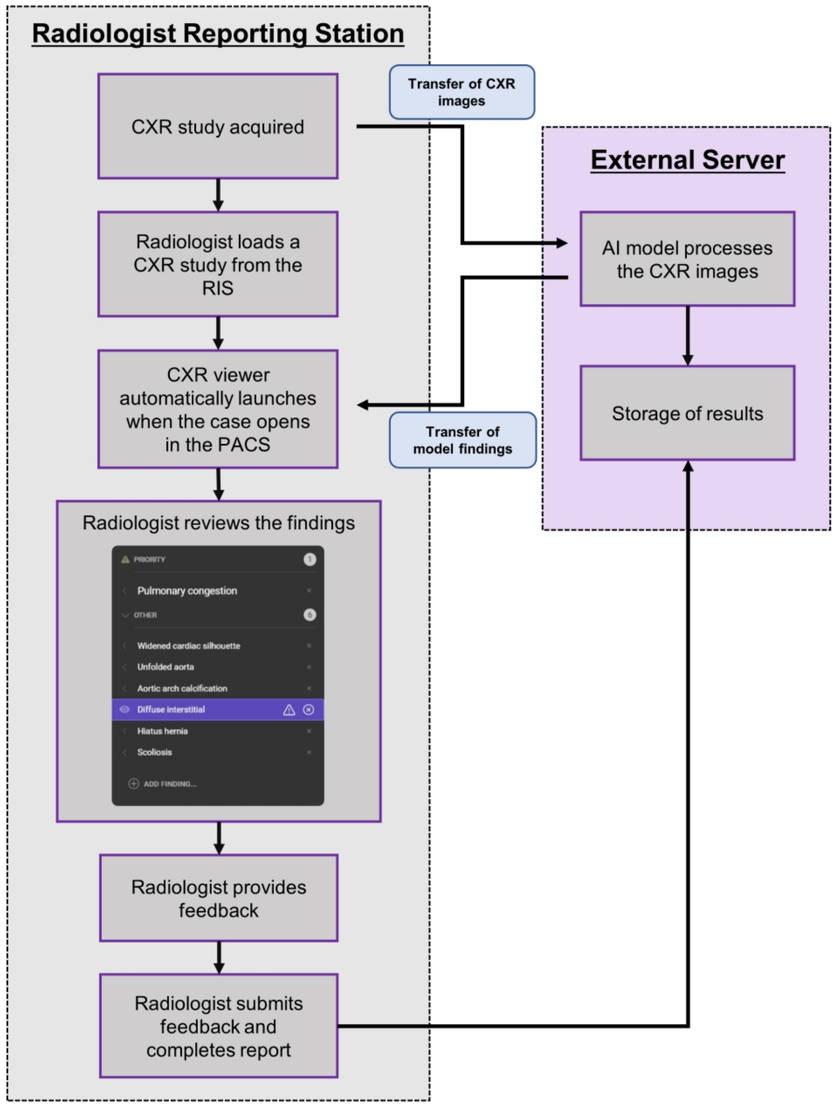

Figure 1 – Flow diagram illustrating the AI-assisted reporting process described in this study. (RIS: Radiological information system)

484x610mm (118 x 118 DPI)

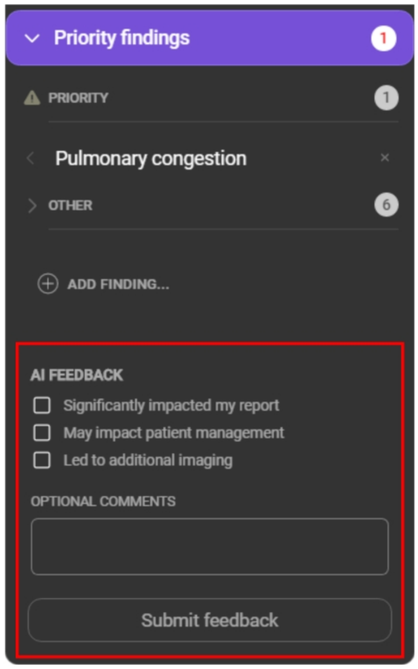

Figure 2 – Example of the modified user interface used by the participating radiologists in this study. The red box highlights the feedback options added to the interface for this study.

645x484mm (118 x 118 DPI)

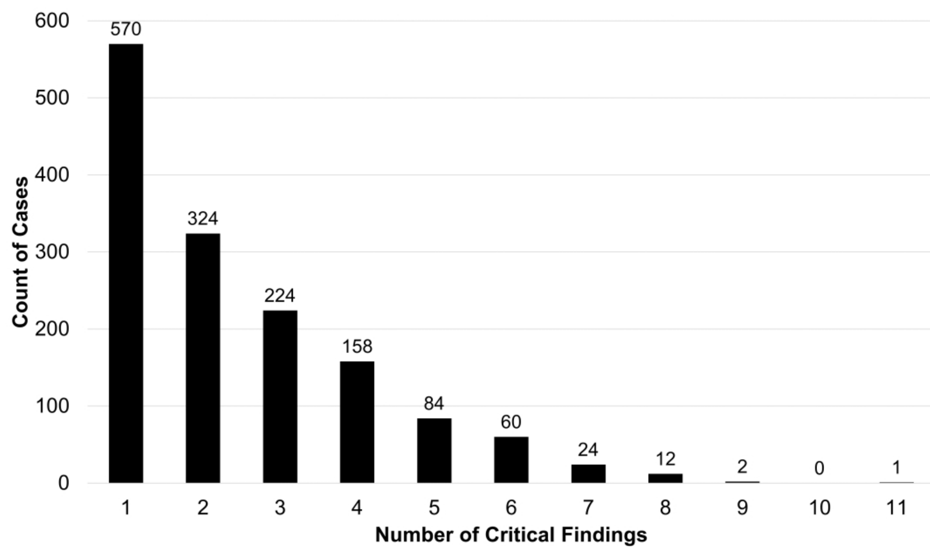

Figure 3 – Counts of numbers of critical findings for the cases seen by the radiologist, defined as the number of critical findings agreed + the number of critical findings added. The number of cases which returned zero findings was 1,513.

861x484mm (118 x 118 DPI)

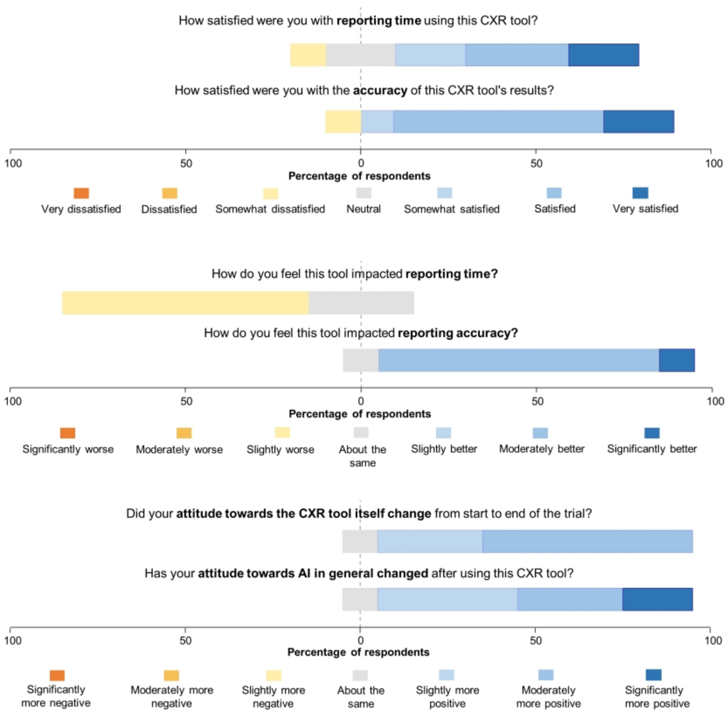

Figure 4 - Diverging stacked bar chart depicting the first set of radiologist survey responses.

645x484mm (118 x 118 DPI)

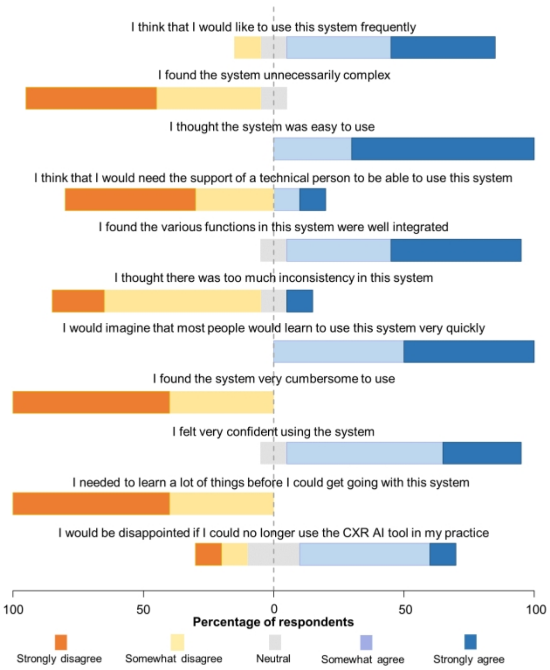

Figure 5 – Diverging stacked bar chart visualising the second set of survey responses of the radiologists.

645x484mm (118 x 118 DPI)

Supplementary Table 1 - List of the 124 findings, including 34 critical findings which the model is validated to detect. The format used by the model to recommend each finding are presented in brackets (Laterality: indicates whether the predicted finding is present on the left or right side, or both. ROI: a predicted region of interest localiser is overlayed on the image. None: no segmentation). ETT: endotracheal tube, NGT: nasogastric tube, PAC: pulmonary artery catheter.

Critical Clinical Findings (Localisation)		
Acute humerus fracture (Laterality)	Loculated effusion (ROI)	Subcutaneous emphysema (Laterality)
Acute rib fracture (ROI)	Lung collapse (Laterality)	Subdiaphragmatic gas (None)
Air Space Opacity – Multifocal (ROI)	Multiple masses or nodules (ROI)	Suboptimal central line (ROI)
Cavitating mass with content (ROI)	Perihilar airspace opacity (Laterality)	Suboptimal ETT (None)
Cavitating mass(es) (ROI)	Pneumomediastinum (None)	Suboptimal NGT (ROI)
Diffuse airspace opacity (Laterality)	Pulmonary congestion (None)	Suboptimal PAC (None)
Diffuse lower airspace opacity (Laterality)	Segmental collapse (ROI)	Superior mediastinal mass (None)
Diffuse upper airspace opacity (Laterality)	Shoulder dislocation (Laterality)	Tension pneumothorax (ROI)
Focal airspace opacity (ROI)	Simple effusion (ROI)	Tracheal deviation (None)
Hilar lymphadenopathy (None)	Simple pneumothorax (ROI)	Widened aortic contour (None)
Inferior mediastinal mass (None)	Solitary lung mass (ROI)	Widened cardiac silhouette (None)
	Solitary lung nodule (ROI)	

Non-Critical Clinical Findings (Localisation)		
Abdominal Clips (None)	Coronary Stent (None)	Pectus Excavatum (None)
Acute Clavicle Fracture (Laterality)	Diaphragmatic Elevation (None)	Peribronchial Cuffing (None)
Airway Stent (None)	Diaphragmatic Eventration (None)	Pericardial Fat Pad (None)
Aortic Arch Calcification (None)	Diffuse Fibrotic Volume Loss (Laterality)	Pleural Mass (ROI)
Aortic Stent (None)	Diffuse Interstitial (Laterality)	Post Resection Volume Loss (Laterality)
Atelectasis (ROI)	Diffuse Nodular / Miliary Lesions (Laterality)	Pulmonary Arterial Catheter (None)
Axillary Clips (Laterality)	Diffuse Pleural Thickening (None)	Pulmonary Artery Enlargement (None)
Basal Predominant Interstitial (Laterality)	Diffuse Spinal Osteophytes (None)	Reduced Lung Markings (None)
Biliary Stent (None)	Distended Bowel (None)	Rib Fixation (Laterality)
Breast Implant (None)	Electronic Cardiac Devices (None)	Rib Lesion (ROI)
Bronchiectasis (None)	Endotracheal Tube (None)	Rib Resection (None)
Bullae Diffuse (None)	Gallstones (None)	Rotator Cuff Anchor (Laterality)

Bullae Lower (None)	Gastric Band (None)	Scapular Fracture (Laterality)
Bullae Upper (None)	Hiatus Hernia (None)	Scapular Lesion (ROI)
Calcified Axillary Nodes (None)	Humeral Lesion (ROI)	Scoliosis (None)
Calcified Granuloma (<5mm) (None)	Intercostal Drain (Laterality)	Shoulder Arthritis (None)
Calcified Hilar Lymphadenopathy (None)	Internal Foreign Body (ROI)	Shoulder Fixation (Laterality)
Calcified Mass (>5mm) (ROI)	Kyphosis (None)	Shoulder Replacement (Laterality)
Calcified Neck Nodes (None)	Lower Zone Fibrotic Volume Loss (Laterality)	Spinal Fixation (None)
Calcified Pleural Plaques (None)	Lung Sutures (None)	Spine Arthritis (None)
Cardiac Valve Prosthesis (None)	Mastectomy (None)	Spine Lesion (ROI)
Central Venous Catheter (ROI)	Mediastinal Clips (None)	Spine Wedge Fracture (ROI)
Cervical Flexion (None)	Nasogastric Tube (ROI)	Sternotomy Wires (None)
Chronic Clavicle Fracture (None)	Neck Clips (Laterality)	Suboptimal Gastric Band (None)
Chronic Humerus Fracture (None)	Nipple Shadow (None)	Unfolded Aorta (None)
Chronic Rib Fracture (None)	Oesophageal Stent (None)	Upper Predominant Interstitial (Laterality)
Clavicle Fixation (Laterality)	Osteopaenia (None)	Upper Zone Fibrotic Volume Loss (Laterality)
Clavicle Lesion (ROI)	Pectus Carinatum (None)	
	Technical Findings	
Chest Incompletely Imaged (None)	Image Obscured (None)	Underexposed (None)
Hyperinflation (None)	Overexposed (None)	Underinflation (None)
	Patient Rotation (None)	

Supplementary Table 2 – Example of the survey questions provided to the radiologists at the end of the study.

	Significantly worse	Moderately worse	Slightly worse	About the same	Slightly better	Moderately better	Significantly better
How do you feel this tool impacted **reporting time**?	o	o	o	o	o	o	o
How do you feel this tool impacted **reporting accuracy**?	o	o	o	o	o	o	o
	Very dissatisfied	Dissatisfied	Somewhat dissatisfied	Neutral	Somewhat satisfied	Satisfied	Very dissatisfied
How satisfied were you with **reporting time** using this CXR tool?	o	o	o	o	o	o	o
How satisfied were you with the **accuracy** of this CXR tool's results?	o	o	o	o	o	o	o
	Significantly more negative	Moderately more negative	Slightly more negative	About the same	Slightly more positive	Moderately more negative	Significantly more negative
Did your **attitude towards the CXR tool itself** change from start to end of the trial?	o	o	o	o	o	o	o
Has your **attitude towards AI in general changed** after using this CXR tool?	o	o	o	o	o	o	o
		Strongly disagree	Somewhat disagree	Neutral	Somewhat agree	Strongly agree	
I think that I would like to use this system frequently.		o	o	o	o	o	
I found the system unnecessarily complex.		o	o	o	o	o	
I thought the system was easy to use.		o	o	o	o	o	
I think that I would need the support of a technical person to be able to use this system.		o	o	o	o	o	
I found the various functions in this system were well integrated.		o	o	o	o	o	
I thought there was too much inconsistency in this system.		o	o	o	o	o	
I would imagine that most people would learn to use this system very quickly.		o	o	o	o	o	
I found the system very cumbersome to use.		o	o	o	o	o	
I felt very confident using the system.		o	o	o	o	o	
I needed to learn a lot of things before I could get going with this system.		o	o	o	o	o	

I would be disappointed if I could no longer use the CXR AI tool in my practice.	o	o	o	o	o

CLAIM: Checklist for Artificial Intelligence in Medical Imaging

Section / Topic	No.	Item	
TITLE / ABSTRACT			
	1	Identification as a study of AI methodology, specifying the category of technology used (e.g., deep learning)	**Yes**
	2	Structured summary of study design, methods, results, and conclusions	**Yes**
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the AI approach	**Yes – page 4/5**
	4	Study objectives and hypotheses	**Yes – page 5**
METHODS			
Study Design	5	Prospective or retrospective study	**Yes – page 8** (under: "CXR case section")
	6	Study goal, such as model creation, exploratory study, feasibility study, non-inferiority trial	**Yes – page 8** (under: "CXR case section")
Data	7	Data sources	**Yes – page 8** (under: "CXR case section")
	8	Eligibility criteria: how, where, and when potentially eligible participants or studies were identified (e.g., symptoms, results from previous tests, inclusion in registry, patient-care setting, location, dates)	**Yes – page 8** (under: "CXR case section")
	9	Data pre-processing steps	**N/A**
	10	Selection of data subsets, if applicable	**N/A**
	11	Definitions of data elements, with references to Common Data Elements	**Yes – page 8/9** (under: "AI-assisted reporting)
	12	De-identification methods	**Yes – page 8** (under: "CXR case section")
	13	How missing data were handled	**N/A**
Ground Truth	14	Definition of ground truth reference standard, in sufficient detail to allow replication	**Yes – page 6** (under: "model development and validation")
	15	Rationale for choosing the reference standard (if alternatives exist)	**N/A**
	16	Source of ground-truth annotations; qualifications and preparation of annotators	**N/A –** Described in reference 31
	17	Annotation tools	**N/A –** Described in reference 31
	18	Measurement of inter- and intrarater variability; methods to mitigate variability and/or resolve discrepancies	**N/A –** Described in reference 31

Data Partitions	**19**	Intended sample size and how it was determined	**Yes – page 10** (under: "statistics and data analysis")
	20	How data were assigned to partitions; specify proportions	**N/A**
	21	Level at which partitions are disjoint (e.g., image, study, patient, institution)	**N/A**
Model	**22**	Detailed description of model, including inputs, outputs, all intermediate layers and connections	**Yes – page 6** (under: "model development and validation") **and described in reference 31**
	23	Software libraries, frameworks, and packages	**Yes – page 6** (under: "model development and validation") **and described in reference 31**
	24	Initialization of model parameters (e.g., randomization, transfer learning)	**Yes – page 6** (under: "model development and validation") **and described in reference 31**
Training	**25**	Details of training approach, including data augmentation, hyperparameters, number of models trained	**Yes – page 6** (under: "model development and validation") **and described in reference 31**
	26	Method of selecting the final model	**N/A**
	27	Ensembling techniques, if applicable	**N/A**
Evaluation	**28**	Metrics of model performance	**Yes – page 6** (under: "model development and validation") **and described in reference 31**
	29	Statistical measures of significance and uncertainty (e.g., confidence intervals)	**Yes – page 6** (under: "model development and validation") **and described in reference 31**
	30	Robustness or sensitivity analysis	**N/A**
	31	Methods for explainability or interpretability (e.g., saliency maps), and how they were validated	**N/A**
	32	Validation or testing on external data	**N/A**
RESULTS			
Data	**33**	Flow of participants or cases, using a diagram to indicate inclusion and exclusion	**Yes – Figure 1**
	34	Demographic and clinical characteristics of cases in each partition	**N/A**
Model performance	**35**	Performance metrics for optimal model(s) on all data partitions	**N/A**
	36	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	**N/A**
	37	Failure analysis of incorrectly classified cases	**N/A**
DISCUSSION			
	38	Study limitations, including potential bias, statistical uncertainty, and generalizability	**Yes – page 13** (under: " limitations and future research")

	39	Implications for practice, including the intended use and/or clinical role	**Yes – page 13** (under: "conclusion")
OTHER INFORMATION			
	40	Registration number and name of registry	**N/A**
	41	Where the full study protocol can be accessed	**N/A**
	42	Sources of funding and other support; role of funders	**Yes – page 21**

Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020; 2(2):e200029. https://doi.org/10.1148/ryai.2020200029

BMJ Open

Assessment of the effect of a comprehensive chest radiograph deep learning model on radiologist reports and patient outcomes: a real-world observational study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-052902.R2
Article Type:	Original research
Date Submitted by the Author:	23-Nov-2021
Complete List of Authors:	Jones, Catherine; Annalise-AI; I-Med Network Danaher, Luke; I-Med Network Milne, Michael; Annalise-AI; I-Med Network Tang, Cyril; Annalise-AI Seah, Jarrel; Alfred Health, Radiology; Annalise AI, Oakden-Rayner, Luke; The University of Adelaide, Australian Institute for Machine Learning Johnson, Andrew; Annalise-AI Buchlak, Quinlan; Annalise-AI; The University of Notre Dame Australia School of Medicine Sydney Campus Esmaili, Nazanin; The University of Notre Dame Australia School of Medicine Sydney Campus; University of Technology Sydney
Primary Subject Heading:	Radiology and imaging
Secondary Subject Heading:	Emergency medicine, Radiology and imaging
Keywords:	Chest imaging < RADIOLOGY & IMAGING, RADIOLOGY & IMAGING, Diagnostic radiology < RADIOLOGY & IMAGING

Assessment of the effect of a comprehensive chest radiograph deep learning model on radiologist reports and patient outcomes: a real-world observational study

Catherine M Jones[1,2], Luke Danaher[2], Michael R Milne[1,2*], Cyril Tang[1], Jarrel Seah[1,3], Luke Oakden-Rayner[4], Andrew Johnson[1], Quinlan D Buchlak[1,5], Nazanin Esmaili[5,6]

[1]Annalise-AI, Sydney, NSW, Australia
[2]I-MED Radiology Network, Sydney, NSW, Australia
[3]Department of Radiology, Alfred Health, Melbourne, VIC, Australia
[4]Australian Institute for Machine Learning, The University of Adelaide, Adelaide, SA, Australia
[5]School of Medicine, University of Notre Dame Australia, Sydney, NSW, Australia
[6]Faculty of Engineering and IT, University of Technology Sydney, Ultimo, NSW, Australia

*Correspondence to: michael.milne@annalise.ai

Corresponding author:
Name: Michael Milne
Annalise-AI
Sydney, Australia
E-mail: michael.milne@annalise.ai

Keywords: Machine learning; chest X-ray, deep learning.

Word Count: 4,486

ABSTRACT

Objectives: AI algorithms have been developed to detect imaging features on chest X-ray (CXR) with a comprehensive AI model capable of detecting 124 CXR findings being recently developed. The aim of this study was to evaluate the real-world usefulness of the model as a diagnostic assistance device for radiologists.

Design: This prospective real-world multicentre study involved a group of radiologists using the model in their daily reporting workflow to report consecutive chest X-rays and recording their feedback on level of agreement with the model findings and whether this significantly affected their reporting.

Setting: The study took place at radiology clinics and hospitals within a large radiology network in Australia between November and December 2020.

Participants: Eleven consultant diagnostic radiologists of varying levels of experience participated in this study.

Primary and secondary outcome measures: Proportion of CXR cases where use of the AI model led to significant material changes to the radiologist report, to patient management, or to imaging recommendations. Additionally, level of agreement between radiologists and the model findings, and radiologist attitudes towards the model were assessed.

Results: Of 2,972 cases reviewed with the model, 92 cases (3.1%) had significant report changes, 43 cases (1.4%) had changed patient management and 29 cases (1.0%) had further imaging recommendations. In terms of agreement with the model, 2,572 cases showed complete agreement (86.5%). 390 (13%) cases had one or more findings rejected by the radiologist. There were 16 findings across 13 cases (0.5%) deemed to be missed by the model. Nine out of 10 radiologists felt their accuracy was improved with the model and were more positive towards AI post-study.

Conclusions: Use of an AI model in a real-world reporting environment significantly improved radiologist reporting and showed good agreement with radiologists, highlighting the potential for AI diagnostic support to improve clinical practice.

ARTICLE SUMMARY

Strengths and limitations of this study

- This study substantially adds to the limited literature on real-world evaluation of comprehensive CXR AI models in radiology workflow.
- This was a multicentre study conducted across a mix of public hospitals, private hospitals, and community clinic settings.
- Due to the design of the study, diagnostic accuracy of the decision support system was not a measurable outcome.
- Results of this study are self-reported and may therefore be prone to bias.
- Determination of the significance of report changes due to the model's recommendations was made at the discretion of each radiologist on a case-by-case basis.

INTRODUCTION

Radiology is a data-rich medical specialty and is well placed to embrace artificial intelligence [1].This is especially true in high volume imaging tasks such as chest X-ray imaging. The rapid application of X-ray technology to diagnosing chest diseases at the end of the 19th century led to the chest X-ray (CXR) becoming a first-line diagnostic imaging tool [2] and it remains an essential component of the diagnostic pathway for chest disease. Due to advancements in digital image acquisition, low ionising radiation dose and low cost, the chest radiograph is more easily accessible worldwide than any other imaging modality [3].

The challenges of interpreting CXR, however, have not lessened over the last half-century. CXR images are 2D representations of complex 3D structures, relying on soft tissue contrast between structures of different densities. Multiple overlapping structures lead to reduced visibility of both normal and abnormal structures [4], with up to 40% of the lung parenchyma obscured by overlying ribs and the mediastinum [5]. This can be further exacerbated by other factors including the degree of inspiration, other devices in the field of view, and patient positioning. In addition, there is a wide range of pathology in the chest which is visible to varying degrees on the CXR. These factors combine to make CXRs difficult to accurately interpret, with an error rate of 20-50% for CXRs containing radiographic evidence of disease reported in the literature [6]. Notably, lung cancer is one of the most common cancers worldwide and is the most common cause of cancer death [7], and CXR interpretation error accounts for 90% of cases where lung cancer is missed [8]. Despite technological advancements in CXR over the past 50 years, this level of diagnostic error has remained constant [6].

A rapidly developing field attempting to assist radiologists in radiological interpretation involves the application of machine learning, in particular deep neural networks [9]. Deep neural networks learn patterns in large, complex datasets, enabling the detection of subtle features and outcome prediction [10,11]. The potential of these algorithms has grown rapidly in the past decade thanks to the development of more useful neural network models, advancements in computational power, and an increase in the volume and availability of digital imaging datasets [11]. Of note is the rise of convolutional neural networks (CNNs), a type of deep neural network that excels at image feature extraction and classification, and demonstrates strong performance in medical image analysis, leading to the rapid advancement of computer vision in medical imaging [12,13]. CNNs have been used to develop models to successfully detect targeted clinical findings on CXR, including lung cancer [14,15], pneumonia [16,17], COVID-19 [18], pneumothorax [19–22], pneumoconiosis [23], cardiomegaly [24], pulmonary hypertension [25] and tuberculosis [26–30]. These studies highlight the effectiveness of applied machine learning in CXR interpretation, however most of these deep learning systems are limited in scope to a single finding or a small set of findings, therefore lacking the broad utility that would make them useful in clinical practice.

Recently, our group developed a comprehensive deep learning CXR diagnostic assist device, which was designed to assist clinicians in CXR interpretation and improve diagnostic accuracy, validated for 124 clinically relevant findings seen on frontal and lateral chest radiographs [31]. The primary objective of the current study was to evaluate the real-world usefulness of the model as a diagnostic assist device for radiologists in both hospital and community clinic settings. This involved examining the frequency at which the model's recommendations led to a ‘significant impact on the report’, defined as the inclusion of findings recommended by the model which altered the radiologists report in a meaningful way. The frequency of change in patient management and recommendations for further imaging were also evaluated. Secondary endpoints included: (1) investigating agreement between radiologists and the findings detected by the model; and (2) assessing radiologist attitudes towards the tool and AI models in general.

METHODS

Ethics Statement

This study was approved by the institutional human research ethics committee of the Wesley Hospital, Brisbane, Queensland Australia (2020.14.324). Written informed consent was obtained from each participating radiologist. The requirement of patient consent was waived by the ethics committee due to the low-risk nature of the study.

Model development and validation

A modified version of a commercially available AI tool for use as a diagnostic assist device displaying results within a viewer (CXR viewer; Annalise CXR ver 1.2, Annalise-AI, Sydney, Australia) was evaluated [32]. The AI tool deploys an underlying machine learning model, developed and validated by Seah et al [31], which consists of attribute and classification CNNs based on the EfficientNet architecture [33] and a segmentation CNN based on U-Net [34] with EfficientNet backbone. The model was trained on 821,681 de-identified CXR images from 284,649 patients originating from inpatient, outpatient and emergency settings across Australia, Europe, and North America. Training dataset labelling involved independent triple labelling of all images by three radiologists selected from a wider pool of 120 consultant radiologists (none of whom were employed by the radiology network involved in this current study). The model was validated for 124 clinical findings in a multi-reader, multi-case (MRMC) study [31]. Thirty-four of these findings were deemed priority findings based on their clinical importance. The full list of 124 findings is available in Supplementary Table 1. Ground truth labels for the validation study dataset were determined by a consensus of three independent radiologists drawn from a pool of seven fully credentialed subspecialty thoracic radiologists. The algorithm is publicly available at https://cxrdemo.annalise.ai. The AI model was used in line with pre-existing regulatory approval [35].

Technical Integration

Prior to the start of the study, technical integration of the software into existing radiology practice systems and testing occurred over several weeks. First, an integration adapter was installed on the IT network of each radiology clinic and acted as a gateway between the internal IT infrastructure and the AI model. Auto-routing rules were established ensuring only CXR studies were forwarded to the integration adapter from the picture archiving and communication system (PACS). Following a successful testing period, the Annalise CXR viewer was installed and configured on workstations for the group of study radiologists.

Study Participants

Eleven consultant radiologists working for a large Australian radiology network were invited to participate in the study through their local radiologist network. This group included general diagnostic radiologists who had completed specialist radiology training and passed all diagnostic radiology college examinations required for consultant accreditation in Australia. All radiologists reported the minimum of 2000 chest radiographs per year (either within the radiology network or through other institutions) suggested to maintain competency [36]. No subspecialist chest radiologists were included.

The group included radiologists with a range of experience levels: five radiologists had 0–5 years post-training experience, three radiologists had 6–10 years of experience, and three radiologists had more than 10 years of experience. Radiologists were situated across four states in Australia and worked in public hospitals, private hospitals and community clinic settings. Both on site and remote reporting was included, in line with regular workflow. Prior to study commencement, each radiologist attended a training seminar and a one-on-one training session to fully understand the CXR viewer and its features. In addition, the participating radiologists were able to familiarise themselves with the viewer prior to commencement of data collection.

CXR Case Selection

In this multicentre real-world prospective study, all consecutive chest radiographs reported by the radiologists originating from inpatient, outpatient, and emergency settings were included for a period covering nearly six weeks. The CXR cases were reported with the assistance of the AI tool in real-world clinical practice, using high resolution diagnostic radiology monitors within the radiologists' normal reporting environment. As per usual workflow across a large radiology network spanning a geographically large area with many regional and remote clinics, both on-site and remote reporting of CXR cases was undertaken. A total of 106 sites contributed cases with case numbers varying from one case up to a maximum of 271 cases at the busiest site.

At least one frontal chest radiograph was required for analysis by the model, and cases that did not include at least one were excluded. Chest radiographs from patients aged younger than 16 years were excluded. Data from all sources was de-identified for analysis.

AI-Assisted Reporting

For each CXR case, radiologists produced their clinical report with access to clinical information, the referral and available patient history, in line with the normal workflow. The AI model analyses the CXR image(s) for each case but does not incorporate clinical inputs (such as previous imaging, referral information or patient demographic data) into the analysis. Model output was displayed to the radiologist in a user interface, linked to the image in the PACS, automatically launching when a CXR case was opened (Figure 1).

A modified version of the commercially available AI software was employed for this study, which incorporated changes into the user interface to allow radiologists to provide feedback on model recommendations. No changes were made to the underlying model. An example of the modified model user interface is presented in figure 2. For each case, the model provided a list of suggested findings, listed as “priority” or “other”, along with a confidence indicator. For a subset of findings, a region of interest localiser was overlayed on the image and the model indicated whether the finding was on the left or the right side, or both (see Supplementary Table 1). The CXR viewer was configured to display its findings after the radiologists’ initial read of the case. For each case, radiologists were asked to review the CXR viewer’s findings and provide feedback within the viewer. The options presented to the radiologists in the viewer are listed in Table 1.

218 *Table 1 - List of review options presented to the radiologist with each case.*

REVIEW OPTION	DESCRIPTION
Rejected clinical finding	A model-detected finding disputed by the radiologist
Missed clinical finding	A model-detected finding missed by the radiologist
Add additional findings	Finding(s) identified by the radiologist but not identified by the model
These findings significantly impacted my report	A yes/no binary question relating to the effect of the model output on the radiologist report
These findings may impact patient management	A yes/no binary question relating to the effect of the model output on patient management, as perceived by the reporting radiologist
These findings led to additional imaging recommendations	A binary yes/no question related to whether the radiologist recommended further imaging based on the model output

The outcome measure of 'significant impact on the report' was the primary outcome measure. A significant change was described as the inclusion of findings recommended by the model, which altered the radiologists report in a meaningful way. As this varied by patient and clinical setting, it was left to the discretion of the radiologist. During the analysis of radiologist feedback, it was assumed that a change in patient management or further imaging recommendation would not occur without radiologists indicating a material change in the CXR report, and thus management and imaging questions were dependent on a significant change in the report. This was also patient-specific; for example, missing a pneumothorax in a ventilated patient with known pneumothorax would not have the same impact on patient management as a previously unknown pneumothorax in an outpatient. Free text input describing missed findings or other relevant data were manually added after data collection was complete.

No formal adjudication of cases showing discrepancy between radiologist and model interpretation was performed. The study was not designed as a diagnostic accuracy validation. No review or ground truthing process was performed. Radiologists remained responsible for image interpretation and formulation of the report.

Post-Study Survey

Upon completion of data collection, a post-study survey was distributed to all participating radiologists to obtain feedback on the usefulness of the CXR viewer and how it affected their opinion of AI in radiology. A table of the survey questions is presented in Supplementary Table 2.

Statistics and Data Analysis

A 1% rate of significant changes in reports (the primary outcome measure) was deemed to be clinically significant prior to commencing the study. Based on estimations of the prevalence of missed critical findings on CXR, preliminary power calculations estimated that the number of cases required to detect at least a 1% rate of significant changes in reports was approximately 2000 cases in total, with alpha value 0.05 and desired power of 0.90. To account for any dropout in radiologists or cases, a target of 3000 cases was set for the study. Ten radiologists were recruited, with an eleventh included for any unexpected participant drop out and to achieve this target in a reasonable time period.

A two-tailed binomial test was used to test the hypothesis that the rate of significant report change, patient management change, or imaging recommendation change was at least 1%. To ensure that the sampling of CXRs reasonably approximated a random snapshot of the true population, radiologists in various states, experience levels as well as different conditions of practice (community clinic vs hospital based) were selected. Additionally, the study was conducted prospectively which further aligned the structure of the sampled data with the expected structure of the population, justifying the choice of analysing the sample using a binomial test without adjustment for each radiologist.

Multivariate logistic regression using generalised linear mixed effect analysis was used to assess the effect of several possible confounders on the measured outcomes, including the number of critical clinical findings per case identified by the model, the inpatient/outpatient status of the patients, the experience level of the radiologists, and the presence or absence of a lateral radiograph. The Wald test was applied to the derived regression coefficients to determine their significance.

Radiologists were grouped by experience level into 0-5 years post completion of radiology training, 6-10 years, and more than ten years. A likelihood ratio test comparing a binomial logistic regression with categorical radiologist experience against a null model was performed to assess the hypothesis that the outcomes (significant changes in reports, management, or imaging recommendation) were associated with experience.

A significance threshold of 0.05 was chosen, with the Benjamini-Hochberg procedure [37] applied to all reported outcomes to account for multiple hypothesis testing. Two clinically qualified researchers independently performed statistical analyses using different software. Calculations were performed in Excel 2016 with RealStatistics resource pack and cross-checked in Python 3.7 using the Pandas 1.0.5 [38], NumPy 1.18.5 [39], SciPy 1.4.1 [40], Scikit-Learn 0.24.0 [41], pymer4 0.7.1 (linked to R 3.4.1, Ime4 1.1.26) [42] and Statsmodels 0.12.1 [43] libraries.

RESULTS

A total of 2,972 cases were reported by 11 radiologists over a period of six weeks. These cases came from 2,665 unique patients (52.7% male), with a median age of 67 (IQR 50–77). Information on radiologist experience, number of cases reported, source of cases and outcome measures for each radiologist are listed in Table 2.

Table 2 - Demographics and results for the eleven radiologists involved in this study. Percentages (%) represent the associated value as a proportion of the total case number for that radiologist.

Radiologist ID	Number of years post-training	Cases reported (% outpatient)	Significant report impact (%)	Patient management changes (%)	Imaging reco mmendations (%)
1	19	136 (21.3)	1 (0.7)	1 (0.7)	0 (0.0)
2	1	325 (46.2)	4 (1.2)	0 (0.0	1 (0.3)
3	4	230 (86.1)	20 (8.6)	14 (6.1)	10 (4.3)
4	6	375 (22.7)	3 (1.0)	0 (0.0)	1 (0.2)
5	4	186 (45.7)	22 (11.8)	9 (4.8)	8 (4.3)
6	20	333 (11.1)	3 (1.0)	2 (0.6)	1 (0.3)
7	3	312 (48.4)	15 (4.8)	8 (2.5)	1 (0.3)
8	26	408 (39.7)	10 (2.4)	5 (1.2)	4 (1.0)
9	9	214 (43.0)	6 (2.8)	2 (0.9)	2 (0.9)
10	6	159 (98.1)	1 (0.6)	1 (0.6)	1 (0.6)
11	5	294 (40.1)	7 (2.4)	1 (0.3)	0 (0.0)
Total		**2,972**	**92 (3.1)**	**43 (1.4)**	**29 (1.0)**

Of the 2,972 cases, 1,825 (61.4%) cases had lateral (as well as frontal) radiographs available for interpretation. 1,709 (57.5%) cases were from an inpatient setting, and 1,263 (42.5%) from an outpatient setting. The median number of findings per case was five (mean: 5.1, SD: 3.9), with a wide range in the number of findings per case (maximum=20). A total of 364 cases returned zero findings predicted by the model from the complete 124 findings list. 1,526 of the 2,972 cases had one or more critical findings detected by the CXR viewer, with the critical findings in 1,459 (96%) of these cases being confirmed by the radiologist. The number of critical findings per case is summarised in Figure 3.

Influence of the AI model on radiologist reporting

Across all 2,972 cases, there were 92 cases identified by radiologists as having significant report changes (3.1%), 43 cases of changed patient management (1.4%) and 29 cases of additional imaging recommendations (1.0%) as a result of exposure to the AI model output. When compared to the hypothesised 1% rate of change, the findings were significantly higher for changed reports ($p < 0.01$) and changed patient management ($p<0.01$), and not significantly different for rate of imaging recommendation ($p=0.50$).

Agreement with model findings

Of the 2,972 cases, 2,569 had no findings rejected or added by the radiologists, indicating agreement with the model over all 124 possible findings in 86.5% of cases. 306 (10.2%) cases had one finding rejected by the radiologist and 84 (2.8%) had two or more findings rejected by the radiologist. 202 (5.3%) critical findings detected by the model were rejected by radiologists. The missed and rejected critical findings are detailed in Table 3.

13 cases (0.5%) had findings (16 in total) added by the radiologists which they deemed were missed by the model, of which 8 were critical findings (see Table 3). The remaining 8 non-critical missed findings were atelectasis (4 findings), cardiac valve prosthesis (2 findings), spinal wedge fracture (1 finding) and peribronchial thickening (1 finding).

Table 3 – Breakdown of the critical findings detected by the model and the level of radiologist agreement with each, including the number of findings reportedly missed by the model (and added by the radiologist) or missed by the radiologist. Percentages (%) represent the associated value as a proportion of the total number of findings displayed by the model.

Critical Finding	Displayed by model	Radiologist agreed with finding (%)	Radiologist rejected finding (%)	Added in by radiologist	Missed by radiologist
Acute aortic syndrome	2	2.0 (100.0)	0 (0.0)	0	0
Acute humerus fracture	5	5 (100.0)	0 (0.0)	0	0
Acute rib fracture	54	39 (72.2)	15 (27.8)	0	5
Cardiomegaly	1,008	979 (97.1)	29 (2.9)	0	0
Cavitating mass	14	13 (92.9)	1 (7.1)	0	0
Cavitating mass internal content	6	5 (83.3)	1 (16.7)	0	0
Diffuse airspace opacity	13	13 (100.0)	0 (0.0)	0	0
Diffuse lower airspace opacity	153	148 (96.7)	5 (3.3)	0	0
Diffuse perihilar airspace opacity	45	45 (100.0)	0 (0.0)	0	0
Diffuse upper airspace opacity	2	2 (100.0)	0 (0.0)	0	0
Focal airspace opacity	341	321 (94.1)	20 (5.9)	0	2
Hilar lymphadenopathy	8	6 (75.0)	2 (25.0)	0	0
Inferior mediastinal mass	8	7 (87.5)	1 (12.5)	0	0
Loculated effusion	87	80 (92.0)	7 (8.0)	0	1
Lung collapse	11	10 (90.9)	1 (9.1)	0	0
Malpositioned CVC	85	78 (91.8)	7 (8.2)	0	1
Malpositioned ETT	52	43 (82.7)	9 (17.3)	0	0
Malpositioned NGT	39	31 (79.5)	8 (20.5)	0	0
Malpositioned PAC	13	9 (69.2)	4 (30.8)	0	0
Multifocal airspace opacity	125	120 (96.0)	5 (4.0)	0	1
Multiple pulmonary masses	43	38 (88.4)	5 (11.6)	0	0
Pneumomediastinum	5	5 (100.0)	0 (0.0)	1	0
Pulmonary congestion	220	215 (97.7)	5 (2.3)	1	0
Segmental collapse	292	290 (99.3)	2 (0.7)	0	1
Shoulder dislocation	1	0 (0.0)	1 (100.0)	0	0
Simple effusion	687	650 (94.6)	37 (5.4)	0	1
Simple pneumothorax	90	77 (85.6)	13 (14.4)	1	1
Single pulmonary mass	41	38 (92.7)	3 (7.3)	1	1
Single pulmonary nodule	105	95 (90.5)	10 (9.5)	3	5
Subcutaneous emphysema	53	51 (96.2)	2 (3.8)	0	1
Subdiaphragmatic gas	7	7 (100.0)	0 (0.0)	1	0
Superior mediastinal mass	37	32 (86.5)	5 (13.5)	0	0
Tension pneumothorax	11	7 (63.6)	4 (36.4)	0	0
Tracheal deviation	133	133 (100.0)	0 (0.0)	0	0
Total	3,796	3,594 (94.7)	202 (5.3)	8	20

Factors influencing reporting, management, or imaging recommendation

The number of critical findings displayed by the model was significantly higher in cases where there was a change in report, patient management, or imaging recommendation ($p < 0.001$, $p = 0.001$, $p = 0.004$; Table 4). The presence of a lateral projection image in the CXR case interpreted by the model was associated with a significantly greater likelihood of changes to imaging recommendation ($p = 0.005$), but not to the report or patient management ($p = 0.105$ and $p = 0.061$, respectively).

Radiologists with fewer than 5 years consultant experience contributed 1,347 cases, and indicated a rate of 5.0% for significant report change, 2.4% patient management change, and 1.5% recommendations for further imaging. These numbers were higher than for the radiologists with 6-10 years of experience (1.3%, 0.4%, 0.5% respectively over 748 cases) and also for radiologists with greater than 10 years of experience (1.6%, 0.9%, 0.6% over 877 cases). However, a likelihood ratio test applied to binomial logistic regression analysis indicated that the level of radiologist experience did not significantly influence the rate of change in report, patient management, or imaging recommendation ($p = 0.120$, $p = 0.262$, and $p = 0.516$, respectively). Whether a patient was imaged as an inpatient or outpatient was not significantly associated with any change in report, patient management, or imaging recommendation ($p = 0.358$, $p = 0.572$, $p = 0.326$, respectively).

Table 4 - Factors affecting AI model influence on report, patient management, or imaging recommendation. Significance testing by the Benjamini-Hochberg algorithm to account for multiple hypotheses. Odds ratios derived from stepwise logistic regression coefficients with confidence intervals calculated with Benjamini-adjusted thresholds. Radiologist experience analysed as a categorical variable with odds ratios representing effect of changing experience levels from the baseline (0 to 5 years) to a different level.

Predictor	Change	Odds Ratios (Adjusted CI)	P Value	Benjamini-Adjusted Threshold	Significance
Number of Critical Findings	Report	1.306 (1.132-1.507)	0	0.0042	YES
Number of Critical Findings	Patient Management	1.267 (1.056-1.521)	0.001	0.0083	YES
Number of Critical Findings	Imaging Recommendation	1.319 (1.035-1.681)	0.004	0.0125	YES
Lateral CXR	Imaging Recommendation	6.495 (1.297-32.530)	0.005	0.0167	YES
Lateral CXR	Patient Management	2.158 (0.837-5.565)	0.061	0.0208	NO
Lateral CXR	Report	1.542 (0.848-2.805)	0.105	0.025	NO
Radiologist Experience	Report	0 to 5 years: Baseline 6 to 10 years: 0.255 (0.043-1.521) > 10 years: 0.305 (0.065-1.439)	0.120	0.0292	NO
Radiologist Experience	Patient Management	0 to 5 years: Baseline 6 to 10 years: 0.165 (0.009-3.214) > 10 years: 0.378 (0.054-2.654)	0.262	0.0333	NO
Radiologist Experience	Imaging Recommendation	0 to 5 years: Baseline 6 to 10 years: 0.357 (0.034-3.783) > 10 years: 0.380 (0.044-3.287)	0.516	0.0458	NO
Inpatient/Outpatient	Imaging Recommendation	1.550 (0.613-3.919)	0.326	0.0375	NO
Inpatient/Outpatient	Report	0.794 (0.476-1.323)	0.358	0.0417	NO
Inpatient/Outpatient	Patient Management	0.818 (0.408-1.640)	0.572	0.0500	NO

Survey Results

The post-study survey was completed by ten out of the eleven radiologists (Figure 4 and Figure 5). Notably, 7 (70%) participants felt that their reporting time was slightly worse, however when asked how satisfied they were with their reporting time, 7 (70%) indicated that they were satisfied.

Nine out of ten radiologists responded that their reporting accuracy was improved while using the CXR viewer, with nine out of ten (90%) participants being satisfied with accuracy of the CXR model's findings. Nine radiologists (90%) demonstrated an improved attitude towards the use of the AI diagnostic viewer by the end of the study and 9 (90%) demonstrated an improved attitude towards AI in general. No radiologists reported a more negative attitude towards the CXR viewer or towards AI in general.

DISCUSSION

We have previously shown that using the output of this comprehensive deep learning model improved radiologist diagnostic accuracy [44] in a non-clinical setting, but it is important to demonstrate that this improvement translates into meaningful change in a real-world environment. In this multicentre real-world prospective study, we determined how often the finding recommendations of the comprehensive deep learning model led to a material change in the radiologist's report, a change in the patient management recommendation, or a change in subsequent imaging recommendation. To the authors' knowledge, this is the first time that the impact of a comprehensive deep learning model developed to detect radiological findings on CXR has been studied in a real-world reporting environment. Other commercially available deep learning models able to detect multiple findings on CXR have been studied in the non-clinical setting, yielding encouraging results and outperforming physicians in the detection of major thoracic findings [45] as well as improving resident diagnostic sensitivity [46]. Other models have demonstrated diagnostic accuracy that is comparable to that of test radiologists [47]. Additionally, studies have yielded promising results for the use of models in population screening, particularly for tuberculosis, where several models have met the minimum WHO recommendations for tuberculosis triage tests [29,48].

We showed that radiologists agreed with all findings identified by the AI model in 86.5% of cases on a per case basis, while on a per finding basis, agreed with the critical findings identified by the model on 94.7% of findings. Notably, there was a significant change to the report in 3.1% of cases leading to changes in recommended patient management in 1.4% of cases, and changes to imaging recommendations in 1% of cases. Of note, 146 lung lesions (solitary lung nodule and solitary lung mass) were present in the dataset according to the model. Two lung lesions flagged by the model but missed by radiologists were recommended for additional imaging and changed management, subsequently diagnosed as lung carcinoma, highlighting the real-world value of integrating this type of system into the radiology workflow. However, four findings of lung nodule were flagged by the radiologists as missed by the model, indicating that the model alone is not intended to replace radiologist interpretation.

The significant impact of the CXR viewer on radiologist reporting and recommendations did however come at the cost of false positives, with 13% of cases having one or more model findings rejected by the radiologist. When this false positive rate is compared against the false positive rates per case reported in other studies investigating CXR models, which range from 14 – 88% [14,49,50], it is considered acceptable. Furthermore, these studies report false-positive rates for CXR models that only detect lung nodules, while in the current study this represents the false positive rate across 124 findings. Notably, on a per finding basis, only 5.3% of critical findings detected by the model were rejected by the radiologist. However, there were several outliers in the critical findings group that had noticeably higher rates of rejection, including acute rib fracture, hilar lymphadenopathy, malpositioned NGT/PAC, shoulder dislocation and tension pneumothorax. Several explanations for this are low sample size, the subjectivity of diagnosis (especially for hilar lymphadenopathy and tension features of pneumothorax), and heightened model sensitivity at the expense of specificity. In particular, the rate of 'overcalling' of malposition of nasogastric tubes was related to both the threshold choice (favouring sensitivity given the critical nature of NGT malposition) and the limitation in the model output in distinguishing malpositioned NGTs from incompletely visualised NGTs. This limitation has subsequently been addressed with model modifications. Overall, this trade-off appears to be reasonable to the participating radiologists, who reported a high level of satisfaction with the model.

In this study, analysis of radiologists by experience level using logistic regression found no statistically significant relationship between experience level and increased changes to reports, patient management changes, or imaging recommendations as a result of the model. Statistical analysis of the relationship between experience level and change in report was associated with a p value of 0.12, suggesting that, with further research, a significant relationship may be identified. It is expected that the inclusion of a larger group of radiologists may lead to a significant finding, as the association between experience and level of change has been noted in other studies. For example Jang et al., showed that less experienced radiologists benefited the most from the diagnostic assistance in detecting lung nodules on CXR [14]. In this study, three of the 11 radiologists contributed a higher than average incidence of the primary outcome of report change, and these were all less experienced radiologists compared to the cohort average experience level. Whilst this may be due to variations in individual radiologist interpretation of 'significant report change', the consistency of experience level across these three radiologists suggests a relationship with experience level and tool impact.

The primary factor that influenced the likelihood of the model findings leading to a change in the report was the presence of critical findings in the model's recommendation. This is particularly notable because it indicates that the changes to the report are significant. They did not simply involve the inclusion of additional non-critical findings in the report, which may be interpreted as overestimating the impact of the model. The inpatient or outpatient status of a case was found not to significantly affect the likelihood of significant changes to the radiologists' report, to patient management, or to imaging recommendations.

The post-study survey provided further insight into the impact that the CXR viewer had on participant reporting, in addition to the level of agreement and changes to the radiology report and patient management recommendations outlined above. The first notable response was that the CXR viewer may have negatively affected reporting times (albeit only mildly) for the majority of radiologists. This outcome was expected in this study setting because the radiologists were taking additional time to provide feedback on the model's recommendations for each case. Previous studies that surveyed radiologists reported that 74.4% thought AI would lower the interpretation time [51]. It is notable that even with the negative impact the model had on reporting time, the majority of radiologists (70%) were still satisfied with reporting time while using the CXR viewer, suggesting that the diagnostic improvements offered by the model were enough to offset the additional perceived reporting time. Additional insight from the survey suggested that very little training was required before radiologists felt comfortable using the tool. This is useful as education on AI has been a primary concern amongst clinicians, as a large proportion of radiologists report having little knowledge of AI [52].

Limitations and future research

The results presented in this study are self-reported by participating radiologists and are likely an underestimation of the model's actual impact. It is expected that radiologists would not report every instance in which they made an interpretive error. Another limitation is that there was no objective gold standard against which the radiologist and model interpretation could be measured. This is a small-scale study involving a limited sample size, conducted over several weeks. As a result, it lacks the statistical power to examine the benefit of the model on a finding-by-finding basis. In future, it would be beneficial to conduct a similar study with a larger sample size to allow for more powerful statistical analysis and examination of specific finding changes. Another useful next step would be to include a gold standard to determine the ground truth for the CXR findings, as this would prevent any under reporting which may occur with self-reported results, as well as enable the detection of false negatives as a result of the CXR viewer.

Although none of the cases evaluated in this study had been seen by the model previously, we note that one of the five data sources used for model training originated from the same radiology network. This therefore cannot be considered as true external evaluation. Further work in truly external institutions in the future are welcomed.

Conclusion

The present study indicated that the integration of a comprehensive AI model capable of detecting 124 findings on CXR into a radiology workflow led to significant changes in reports and patient management, with an acceptable rate of additional imaging recommendations. These results were not affected by the inpatient status of the patient, and although approaching significance, the experience level of the radiologists did not significantly relate to the primary endpoint outcomes. In secondary endpoint outcomes, the model output showed good agreement with radiologists, and radiologists showed high rates of satisfaction with their reporting times and diagnostic accuracy when using the CXR viewer as a diagnostic assist device. Results highlight the usefulness of AI-driven diagnostic assist tools in improving clinical practice and patient outcomes.

AUTHOR STATEMENT

CJ contributed to conception and design of the work, acquisition of data, analysis and visualisation of data, interpretation of data, drafting of the work, and project management. LD contributed to design of the work and acquisition of data. MM contributed to conception and design of the work, interpretation and visualisation of data, development of diagrams, drafting of the work, and project management. CT and JS contributed to analysis and visualisation of data, interpretation of data, development of diagrams, and drafting of the work. LO, AJ, QB and NE contributed to interpretation of data. All authors revised the work critically for important intellectual content, gave final approval of the version to be published, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

ACKNOWLEDGEMENTS

The authors would like to thank Mark Wilson, Marc Northrop, Nicolaus Carr and Trina Shnier for their valuable contributions to designing and managing the study.

COMPETING INTERESTS

CJ is a radiologist employed by the radiology practice and a clinical consultant for Annalise-AI. LD, LO and NE are independent of Annalise-AI and have no interests to declare. MM, JS, CT, AJ and QB are employed by or seconded to Annalise-AI. Study conception, study design, ethics approval and data security were conducted independent of Annalise-AI.

FUNDING STATEMENT

This work was supported by Annalise-AI Pty Ltd. Annalise-AI supported this work through free provision of the model to participating radiologists for the duration of the study and financing of an external biostatistician. Award/Grant number is not applicable.

PATIENT AND PUBLIC INVOLVEMENT

Patients and public were not involved in the design, conduct, or reporting of this study.

DATA AVAILABILITY STATEMENT

All data relevant to the study are included in the article or uploaded as online supplemental information. No additional data are available.

References

1 Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 2016;**278**:563–77. doi:10.1148/radiol.2015151169

2 Greene R. Francis H. Williams, MD: father of chest radiology in North America. *RadioGraphics* 1991;**11**:325–32. doi:10.1148/radiographics.11.2.2028067

3 Schaefer-Prokop C, Neitzel U, Venema HW, *et al.* Digital chest radiography: an update on modern technology, dose containment and control of image quality. *Eur Radiol* 2008;**18**:1818–30. doi:10.1007/s00330-008-0948-3

4 Lee CS, Nagy PG, Weaver SJ, *et al.* Cognitive and System Factors Contributing to Diagnostic Errors in Radiology. *American Journal of Roentgenology* 2013;**201**:611–7. doi:10.2214/AJR.12.10375

5 Chotas HG, Ravin CE. Chest radiography: estimated lung volume and projected area obscured by the heart, mediastinum, and diaphragm. *Radiology* 1994;**193**:403–4. doi:10.1148/radiology.193.2.7972752

6 Berlin L. Accuracy of Diagnostic Procedures: Has It Improved Over the Past Five Decades? *American Journal of Roentgenology* 2007;**188**:1173–8. doi:10.2214/AJR.06.1270

7 Zaorsky NG, Churilla TM, Egleston BL, *et al.* Causes of death among cancer patients. *Annals of Oncology* 2017;**28**:400–7. doi:10.1093/annonc/mdw604

8 del Ciello A, Franchi P, Contegiacomo A, *et al.* Missed lung cancer: when, where, and why? *Diagn Interv Radiol* 2017;**23**:118–26. doi:10.5152/dir.2016.16187

9 Fazal MI, Patel ME, Tye J, *et al.* The past, present and future role of artificial intelligence in imaging. *European Journal of Radiology* 2018;**105**:246–50. doi:10.1016/j.ejrad.2018.06.020

10 Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science* 2015;**349**:255–60. doi:10.1126/science.aaa8415

11 Hosny A, Parmar C, Quackenbush J, *et al.* Artificial intelligence in radiology. *Nat Rev Cancer* 2018;**18**:500–10. doi:10.1038/s41568-018-0016-5

12 Erickson BJ, Korfiatis P, Akkus Z, *et al.* Machine Learning for Medical Imaging. *RadioGraphics* 2017;**37**:505–15. doi:10.1148/rg.2017160130

13 Esteva A, Chou K, Yeung S, *et al.* Deep learning-enabled medical computer vision. *npj Digital Medicine* 2021;**4**:1–9. doi:10.1038/s41746-020-00376-2

14 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. doi:10.1148/radiol.2020200165

15 Liang C-H, Liu Y-C, Wu M-T, *et al.* Identifying pulmonary nodules or masses on chest radiography using deep learning: external validation and strategies to improve clinical practice. *Clinical Radiology* 2020;**75**:38–45. doi:10.1016/j.crad.2019.08.005

16 Hurt B, Kligerman S, Hsiao A. Deep Learning Localization of Pneumonia: 2019 Coronavirus (COVID-19) Outbreak. *J Thorac Imaging* 2020;**35**:W87–9.

17 Kim JY, Choe PG, Oh Y, *et al.* The First Case of 2019 Novel Coronavirus Pneumonia Imported into Korea from Wuhan, China: Implication for Infection Prevention and Control Measures. *J Korean Med Sci* 2020;**35**. doi:10.3346/jkms.2020.35.e61

18 Bassi PRAS, Attux R. A Deep Convolutional Neural Network for COVID-19 Detection Using Chest X-Rays. *arXiv:200501578 [cs, eess]* Published Online First: 12 January 2021.http://arxiv.org/abs/2005.01578 (accessed 23 Mar 2021).

19 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. *Investigative Radiology* 2020;**55**:792–8. doi:10.1097/RLI.0000000000000707

20 Sze-To A, Wang Z. tCheXNet: Detecting Pneumothorax on Chest X-Ray Images Using Deep Transfer Learning. In: Karray F, Campilho A, Yu A, eds. *Image Analysis and Recognition*. Cham: : Springer International Publishing 2019. 325–32. doi:10.1007/978-3-030-27272-2_28

21 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. doi:10.1007/s00330-020-06771-3

22 Park S, Lee SM, Kim N, *et al.* Application of deep learning–based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy. *Eur Radiol* 2019;**29**:5341–8. doi:10.1007/s00330-019-06130-x

23 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. doi:10.1136/oemed-2019-106386

24 S Z, X Z, R Z. Identifying Cardiomegaly in ChestX-ray8 Using Transfer Learning. *Stud Health Technol Inform* 2019;**264**:482–6. doi:10.3233/shti190268

25 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. doi:10.1371/journal.pone.0236378

26 Pasa F, Golkov V, Pfeiffer F, *et al.* Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. *Scientific Reports* 2019;**9**:6268. doi:10.1038/s41598-019-42557-4

27 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. *Scientific Reports* 2020;**10**:210. doi:10.1038/s41598-019-56589-3

28 Heo S-J, Kim Y, Yun S, *et al.* Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers' Health Examination Data. *International Journal of Environmental Research and Public Health* 2019;**16**:250. doi:10.3390/ijerph16020250

29 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. *Scientific Reports* 2019;**9**:15000. doi:10.1038/s41598-019-51503-3

30 Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology* 2017;**284**:574–82. doi:10.1148/radiol.2017162326

31 Seah JCY, Tang CHM, Buchlak QD, *et al.* Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *The Lancet Digital Health* 2021;**3**:e496–506. doi:10.1016/S2589-7500(21)00106-0

32 Annalise.ai - Annalise CXR comprehensive medical imaging AI. Annalise.ai. https://annalise.ai/products/annalise-cxr/ (accessed 23 Mar 2021).

33 Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv:190511946 [cs, stat]* Published Online First: 11 September 2020.http://arxiv.org/abs/1905.11946 (accessed 30 Mar 2021).

34 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:150504597 [cs]* Published Online First: 18 May 2015.http://arxiv.org/abs/1505.04597 (accessed 30 Mar 2021).

35 xmlmillr6.pdf. https://www.ebs.tga.gov.au/servlet/xmlmillr6?dbid=ebs/PublicHTML/pdfStore.nsf&docid=F7ADAEBB76CEDD47CA2585E500424A43&agid=(PrintDetailsPublic)&actionid=1 (accessed 25 Aug 2021).

36 ace_lung_pathways_final_report_v1.4.pdf. https://www.cancerresearchuk.org/sites/default/files/ace_lung_pathways_final_report_v1.4.pdf (accessed 31 Aug 2021).

37 Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)* 1995;**57**:289–300.

38 Mckinney W. pandas: a Foundational Python Library for Data Analysis and Statistics. *Python High Performance Science Computer* 2011.

39 Harris CR, Millman KJ, van der Walt SJ, *et al.* Array programming with NumPy. *Nature* 2020;**585**:357–62. doi:10.1038/s41586-020-2649-2

40 Jones E, Oliphant T, Peterson P. SciPy: Open Source Scientific Tools for Python. 2001.

41 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* Published Online First: 12 October 2011.https://hal.inria.fr/hal-00650905 (accessed 23 Mar 2021).

42 Jolly E. Pymer4: Connecting R and Python for linear mixed modeling. *Journal of Open Source Software* 2018;**3**:862.

43 Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. Austin, Texas: 2010. 92–6. doi:10.25080/Majora-92bf1922-011

44 Seah J, Tang C, Buchlak QD, *et al.* Radiologist chest X-ray diagnostic accuracy performance improvements when augmented by a comprehensive deep learning model. *The Lancet Digital Health* 2021.

45 Hwang EJ, Park S, Jin K-N, *et al.* Development and Validation of a Deep Learning-Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. *JAMA Netw Open* 2019;**2**:e191095. doi:10.1001/jamanetworkopen.2019.1095

46 Hwang EJ, Nam JG, Lim WH, *et al.* Deep Learning for Chest Radiograph Diagnosis in the Emergency Department. *Radiology* 2019;**293**:573–80. doi:10.1148/radiol.2019191225

47 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. doi:10.1371/journal.pone.0204155

48 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. *The Lancet Digital Health* 2020;**2**:e573–81. doi:10.1016/S2589-7500(20)30221-1

49 Dellios N, Teichgraeber U, Chelaru R, *et al.* Computer-aided Detection Fidelity of Pulmonary Nodules in Chest Radiograph. *J Clin Imaging Sci* 2017;**7**. doi:10.4103/jcis.JCIS_75_16

50 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* Published Online First: 12 November 2019. doi:10.1148/radiol.2019182465

51 Waymel Q, Badr S, Demondion X, *et al.* Impact of the rise of artificial intelligence in radiology: What do radiologists think? *Diagnostic and Interventional Imaging* 2019;**100**:327–36. doi:10.1016/j.diii.2019.03.015

52 Collado-Mesa F, Alvarez E, Arheart K. The Role of Artificial Intelligence in Diagnostic Radiology: A Survey at a Single Radiology Residency Training Program. *Journal of the American College of Radiology* 2018;**15**:1753–7. doi:10.1016/j.jacr.2017.12.021

FIGURE LEGENDS

Figure 1 – Flow diagram illustrating the AI-assisted reporting process described in this study. (RIS: Radiological information system)

Figure 2 – Example of the modified user interface used by the participating radiologists in this study. The red box highlights the feedback options added to the interface for this study.

Figure 3 – Counts of numbers of critical findings for the cases seen by the radiologist, defined as the number of critical findings agreed + the number of critical findings added. The number of cases which returned zero findings was 1,513.

Figure 4 – Diverging stacked bar chart depicting the first set of radiologist survey responses.

Figure 5 – Diverging stacked bar chart visualising the second set of survey responses of the radiologists.

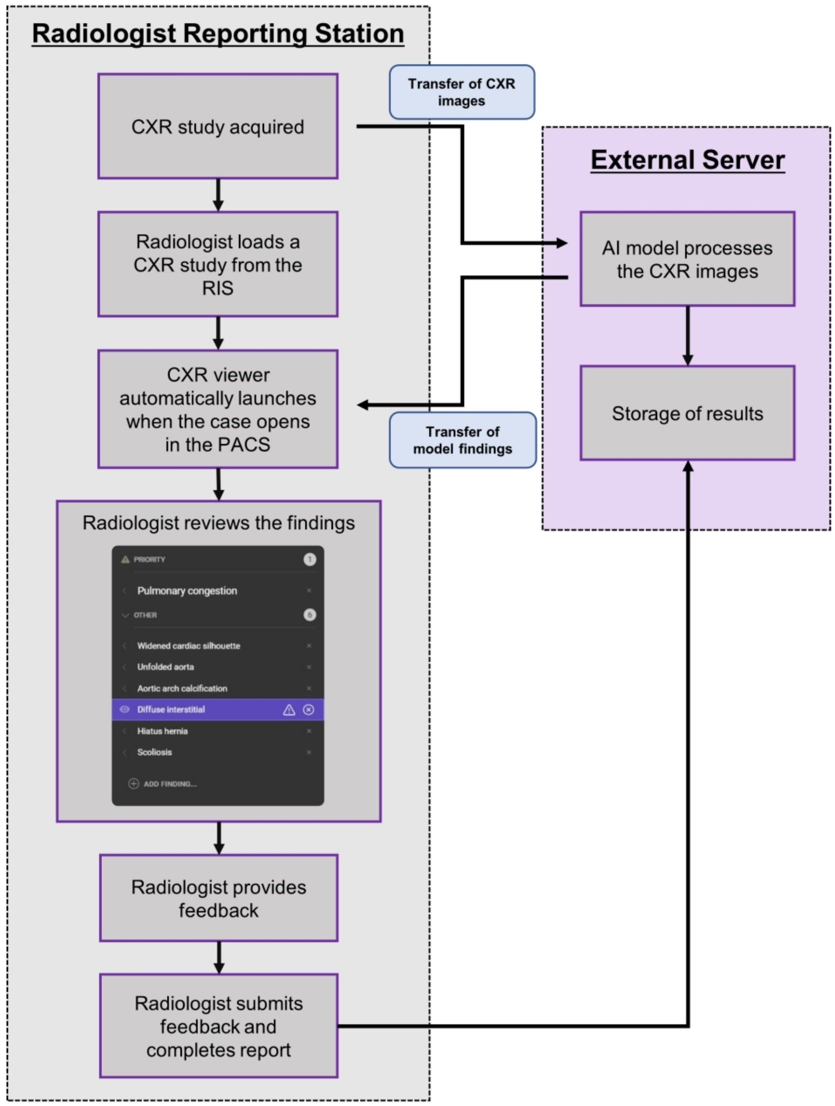

Figure 1 - Flow diagram illustrating the AI-assisted reporting process described in this study. (RIS: Radiological information system)

190x240mm (300 x 300 DPI)

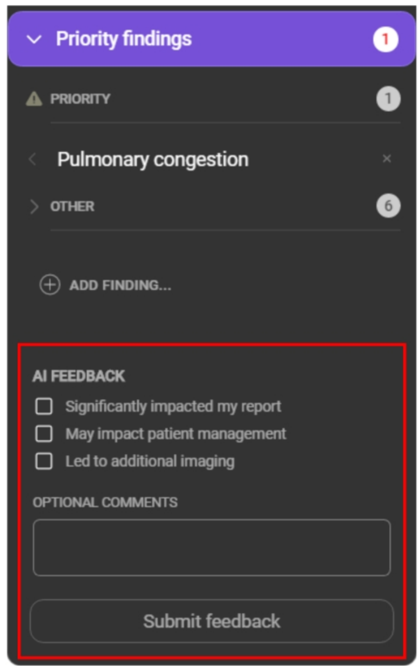

Figure 2 - Example of the modified user interface used by the participating radiologists in this study. The red box highlights the feedback options added to the interface for this study.

254x190mm (300 x 300 DPI)

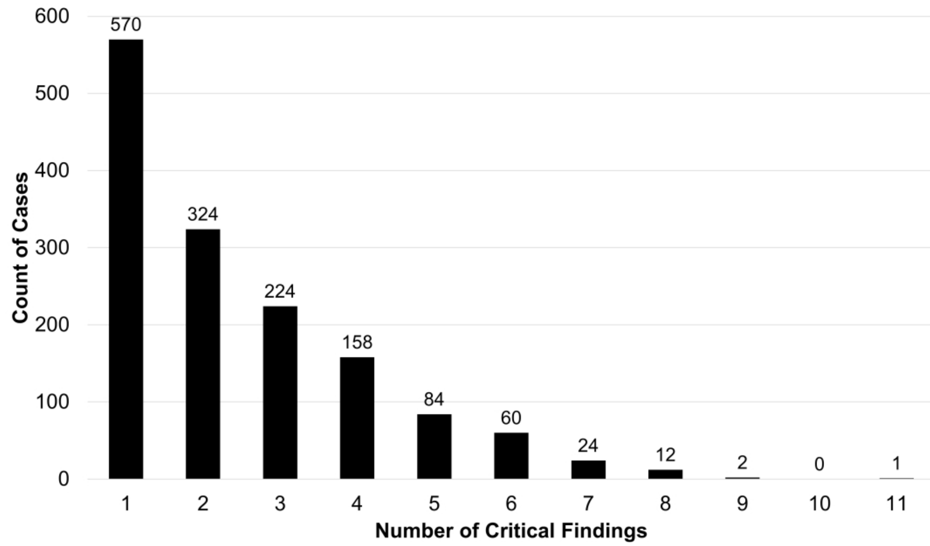

Figure 3 - Counts of numbers of critical findings for the cases seen by the radiologist, defined as the number of critical findings agreed + the number of critical findings added. The number of cases which returned zero findings was 1,513.

338x190mm (300 x 300 DPI)

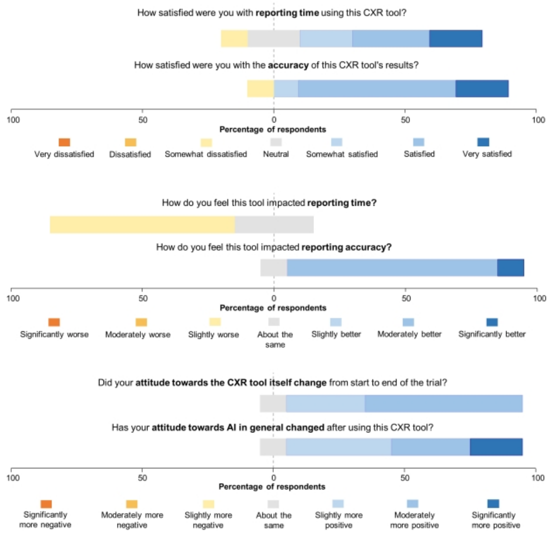

Figure 4 - Diverging stacked bar chart depicting the first set of radiologist survey responses.

338x190mm (300 x 300 DPI)

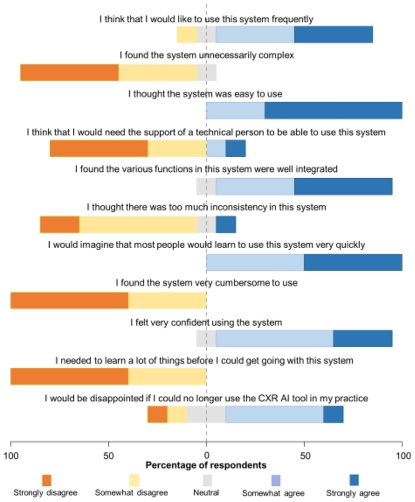

Figure 5 - Diverging stacked bar chart visualising the second set of survey responses of the radiologists.

338x190mm (300 x 300 DPI)

Supplementary Table 1 - List of the 124 findings, including 34 critical findings which the model is validated to detect. The format used by the model to recommend each finding are presented in brackets (Laterality: indicates whether the predicted finding is present on the left or right side, or both. ROI: a predicted region of interest localiser is overlayed on the image. None: no segmentation). ETT: endotracheal tube, NGT: nasogastric tube, PAC: pulmonary artery catheter.

Critical Clinical Findings (Localisation)

Acute humerus fracture (Laterality)	Loculated effusion (ROI)	Subcutaneous emphysema (Laterality)
Acute rib fracture (ROI)	Lung collapse (Laterality)	Subdiaphragmatic gas (None)
Air Space Opacity – Multifocal (ROI)	Multiple masses or nodules (ROI)	Suboptimal central line (ROI)
Cavitating mass with content (ROI)	Perihilar airspace opacity (Laterality)	Suboptimal ETT (None)
Cavitating mass(es) (ROI)	Pneumomediastinum (None)	Suboptimal NGT (ROI)
Diffuse airspace opacity (Laterality)	Pulmonary congestion (None)	Suboptimal PAC (None)
Diffuse lower airspace opacity (Laterality)	Segmental collapse (ROI)	Superior mediastinal mass (None)
Diffuse upper airspace opacity (Laterality)	Shoulder dislocation (Laterality)	Tension pneumothorax (ROI)
Focal airspace opacity (ROI)	Simple effusion (ROI)	Tracheal deviation (None)
Hilar lymphadenopathy (None)	Simple pneumothorax (ROI)	Widened aortic contour (None)
Inferior mediastinal mass (None)	Solitary lung mass (ROI)	Widened cardiac silhouette (None)
	Solitary lung nodule (ROI)	

Non-Critical Clinical Findings (Localisation)

Abdominal Clips (None)	Coronary Stent (None)	Pectus Excavatum (None)
Acute Clavicle Fracture (Laterality)	Diaphragmatic Elevation (None)	Peribronchial Cuffing (None)
Airway Stent (None)	Diaphragmatic Eventration (None)	Pericardial Fat Pad (None)
Aortic Arch Calcification (None)	Diffuse Fibrotic Volume Loss (Laterality)	Pleural Mass (ROI)
Aortic Stent (None)	Diffuse Interstitial (Laterality)	Post Resection Volume Loss (Laterality)
Atelectasis (ROI)	Diffuse Nodular / Miliary Lesions (Laterality)	Pulmonary Arterial Catheter (None)
Axillary Clips (Laterality)	Diffuse Pleural Thickening (None)	Pulmonary Artery Enlargement (None)
Basal Predominant Interstitial (Laterality)	Diffuse Spinal Osteophytes (None)	Reduced Lung Markings (None)
Biliary Stent (None)	Distended Bowel (None)	Rib Fixation (Laterality)
Breast Implant (None)	Electronic Cardiac Devices (None)	Rib Lesion (ROI)
Bronchiectasis (None)	Endotracheal Tube (None)	Rib Resection (None)
Bullae Diffuse (None)	Gallstones (None)	Rotator Cuff Anchor (Laterality)

Bullae Lower (None)	Gastric Band (None)	Scapular Fracture (Laterality)
Bullae Upper (None)	Hiatus Hernia (None)	Scapular Lesion (ROI)
Calcified Axillary Nodes (None)	Humeral Lesion (ROI)	Scoliosis (None)
Calcified Granuloma (<5mm) (None)	Intercostal Drain (Laterality)	Shoulder Arthritis (None)
Calcified Hilar Lymphadenopathy (None)	Internal Foreign Body (ROI)	Shoulder Fixation (Laterality)
Calcified Mass (>5mm) (ROI)	Kyphosis (None)	Shoulder Replacement (Laterality)
Calcified Neck Nodes (None)	Lower Zone Fibrotic Volume Loss (Laterality)	Spinal Fixation (None)
Calcified Pleural Plaques (None)	Lung Sutures (None)	Spine Arthritis (None)
Cardiac Valve Prosthesis (None)	Mastectomy (None)	Spine Lesion (ROI)
Central Venous Catheter (ROI)	Mediastinal Clips (None)	Spine Wedge Fracture (ROI)
Cervical Flexion (None)	Nasogastric Tube (ROI)	Sternotomy Wires (None)
Chronic Clavicle Fracture (None)	Neck Clips (Laterality)	Suboptimal Gastric Band (None)
Chronic Humerus Fracture (None)	Nipple Shadow (None)	Unfolded Aorta (None)
Chronic Rib Fracture (None)	Oesophageal Stent (None)	Upper Predominant Interstitial (Laterality)
Clavicle Fixation (Laterality)	Osteopaenia (None)	Upper Zone Fibrotic Volume Loss (Laterality)
Clavicle Lesion (ROI)	Pectus Carinatum (None)	
	Technical Findings	
Chest Incompletely Imaged (None)	Image Obscured (None)	Underexposed (None)
Hyperinflation (None)	Overexposed (None)	Underinflation (None)
	Patient Rotation (None)	

Supplementary Table 2 – Example of the survey questions provided to the radiologists at the end of the study.

	Significantly worse	Moderately worse	Slightly worse	About the same	Slightly better	Moderately better	Significantly better
How do you feel this tool impacted **reporting time**?	o	o	o	o	o	o	o
How do you feel this tool impacted **reporting accuracy**?	o	o	o	o	o	o	o
	Very dissatisfied	Dissatisfied	Somewhat dissatisfied	Neutral	Somewhat satisfied	Satisfied	Very dissatisfied
How satisfied were you with **reporting time** using this CXR tool?	o	o	o	o	o	o	o
How satisfied were you with the **accuracy** of this CXR tool's results?	o	o	o	o	o	o	o
	Significantly more negative	Moderately more negative	Slightly more negative	About the same	Slightly more positive	Moderately more negative	Significantly more negative
Did your **attitude towards the CXR tool itself** change from start to end of the trial?	o	o	o	o	o	o	o
Has your **attitude towards AI in general changed** after using this CXR tool?	o	o	o	o	o	o	o
		Strongly disagree	Somewhat disagree	Neutral	Somewhat agree	Strongly agree	
I think that I would like to use this system frequently.		o	o	o	o	o	
I found the system unnecessarily complex.		o	o	o	o	o	
I thought the system was easy to use.		o	o	o	o	o	
I think that I would need the support of a technical person to be able to use this system.		o	o	o	o	o	
I found the various functions in this system were well integrated.		o	o	o	o	o	
I thought there was too much inconsistency in this system.		o	o	o	o	o	
I would imagine that most people would learn to use this system very quickly.		o	o	o	o	o	
I found the system very cumbersome to use.		o	o	o	o	o	
I felt very confident using the system.		o	o	o	o	o	
I needed to learn a lot of things before I could get going with this system.		o	o	o	o	o	

I would be disappointed if I could no longer use the CXR AI tool in my practice.	o	o	o	o	o

CLAIM: Checklist for Artificial Intelligence in Medical Imaging

Section / Topic	No.	Item	
TITLE / ABSTRACT			
	1	Identification as a study of AI methodology, specifying the category of technology used (e.g., deep learning)	**Yes**
	2	Structured summary of study design, methods, results, and conclusions	**Yes**
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the AI approach	**Yes – page 4/5**
	4	Study objectives and hypotheses	**Yes – page 5**
METHODS			
Study Design	5	Prospective or retrospective study	**Yes – page 8** (under: "CXR case section")
	6	Study goal, such as model creation, exploratory study, feasibility study, non-inferiority trial	**Yes – page 8** (under: "CXR case section")
Data	7	Data sources	**Yes – page 8** (under: "CXR case section")
	8	Eligibility criteria: how, where, and when potentially eligible participants or studies were identified (e.g., symptoms, results from previous tests, inclusion in registry, patient-care setting, location, dates)	**Yes – page 8** (under: "CXR case section")
	9	Data pre-processing steps	**N/A**
	10	Selection of data subsets, if applicable	**N/A**
	11	Definitions of data elements, with references to Common Data Elements	**Yes – page 8/9** (under: "AI-assisted reporting)
	12	De-identification methods	**Yes – page 8** (under: "CXR case section")
	13	How missing data were handled	**N/A**
Ground Truth	14	Definition of ground truth reference standard, in sufficient detail to allow replication	**Yes – page 6** (under: "model development and validation")
	15	Rationale for choosing the reference standard (if alternatives exist)	**N/A**
	16	Source of ground-truth annotations; qualifications and preparation of annotators	**N/A –** Described in reference 31
	17	Annotation tools	**N/A –** Described in reference 31
	18	Measurement of inter- and intrarater variability; methods to mitigate variability and/or resolve discrepancies	**N/A –** Described in reference 31

Data Partitions	**19**	Intended sample size and how it was determined	**Yes – page 10** (under: ”statistics and data analysis”)
	20	How data were assigned to partitions; specify proportions	**N/A**
	21	Level at which partitions are disjoint (e.g., image, study, patient, institution)	**N/A**
Model	**22**	Detailed description of model, including inputs, outputs, all intermediate layers and connections	**Yes – page 6** (under: “model development and validation”) **and described in reference 31**
	23	Software libraries, frameworks, and packages	**Yes – page 6** (under: “model development and validation”) **and described in reference 31**
	24	Initialization of model parameters (e.g., randomization, transfer learning)	**Yes – page 6** (under: “model development and validation”) **and described in reference 31**
Training	**25**	Details of training approach, including data augmentation, hyperparameters, number of models trained	**Yes – page 6** (under: “model development and validation”) **and described in reference 31**
	26	Method of selecting the final model	**N/A**
	27	Ensembling techniques, if applicable	**N/A**
Evaluation	**28**	Metrics of model performance	**Yes – page 6** (under: “model development and validation”) **and described in reference 31**
	29	Statistical measures of significance and uncertainty (e.g., confidence intervals)	**Yes – page 6** (under: “model development and validation”) **and described in reference 31**
	30	Robustness or sensitivity analysis	**N/A**
	31	Methods for explainability or interpretability (e.g., saliency maps), and how they were validated	**N/A**
	32	Validation or testing on external data	**N/A**
RESULTS			
Data	**33**	Flow of participants or cases, using a diagram to indicate inclusion and exclusion	**Yes – Figure 1**
	34	Demographic and clinical characteristics of cases in each partition	**N/A**
Model performance	**35**	Performance metrics for optimal model(s) on all data partitions	**N/A**
	36	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	**N/A**
	37	Failure analysis of incorrectly classified cases	**N/A**
DISCUSSION			
	38	Study limitations, including potential bias, statistical uncertainty, and generalizability	**Yes – page 13** (under: “ limitations and future research”)

	39	Implications for practice, including the intended use and/or clinical role	**Yes – page 13** (under: "conclusion")
OTHER INFORMATION			
	40	Registration number and name of registry	**N/A**
	41	Where the full study protocol can be accessed	**N/A**
	42	Sources of funding and other support; role of funders	**Yes – page 21**

Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020; 2(2):e200029. https://doi.org/10.1148/ryai.2020200029