

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<u>http://bmjopen.bmj.com</u>).

If you have any questions on BMJ Open's open peer review process please email <u>info.bmjopen@bmj.com</u>

BMJ Open

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

Journal:	BMJ Open
Journal.	
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





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 <u>licence</u>.

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 <u>Creative Commons</u> 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.

terez oni

Enseignement Superieur (ABES) Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

2		
3 4	1	Real-world impact of a comprehensive deep learning
5 6 7	2	model designed to assist chest radiograph reporting
8 9	3	
10 11	4	
12	5	
13	6	
14	7	
15 16	8	Cathering M. Janas 2. Luka Danahar? Michael D. Milnel?* Cyril Tangl. Jorral Seeh 3. Luka
17		Catherine M Jones ^{1,2} , Luke Danaher ² , Michael R Milne ^{1,2*} , Cyril Tang ¹ , Jarrel Seah ^{1,3} , Luke
18	9	Oakden-Rayner ⁴ , Andrew Johnson ¹ , Quinlan D Buchlak ^{1,5} , Nazanin Esmaili ^{5,6}
19	10	
20	11	¹ Annalise-AI, Sydney, NSW, Australia
21	12	² I-MED Radiology Network, Sydney, NSW, Australia
22 23	13	³ Department of Radiology, Alfred Health, Melbourne, VIC, Australia
23 24	14	⁴ Australian Institute for Machine Learning, The University of Adelaide, Adelaide, SA, Australia
25	15	⁵ School of Medicine, University of Notre Dame Australia, Sydney, NSW, Australia
26	16 17	⁶ Faculty of Engineering and IT, University of Technology Sydney, Ultimo, NSW, Australia
27	18	
28	19	*Correspondence to: michael.milne@annalise.ai
29 30	20	· · · · · · · · · · · · · · · · · · ·
31	20	
32	22	
33	23	
34	24	
35	25	
36 37	26	
38	27	
39	28	Corresponding author:
40	29	Name: Michael Milne
41	30	Annalise-AI
42 43	31	Sydney, Australia
44	32	Annalise-AI Sydney, Australia E-mail: michael.milne@annalise.ai
45	33	
46	34	
47	35	
48 49	36	
50	37	
51	38	
52	39	
53	40	
54 55	41	
55 56	42	
57	43 44	Keywords: Machine learning; radiomics; chest X-ray, deep learning.
58	44 45	Word Count: 4,201
59 60		

46	ABSTRACT
47 48	
	Objectives: AI algorithms have been developed to detect imaging features on chest X-ray
49	(CXR), however most of these algorithms are limited to detecting a single finding or a small set of
50	findings. Recently, a comprehensive AI model capable of detecting 124 CXR findings was developed and
51	cleared for clinical use. The aim of this study was to evaluate the real-world performance of the model as
52	a diagnostic assistance device for radiologists.
53	Design: This prospective real-world multicentre study involved a group of radiologists using the
54	model in their daily reporting workflow to report consecutive chest X-rays and recording their case-by-
55	case feedback on level of agreement with the model findings and whether this significantly affected their
56	reporting.
57	Setting: The study took place at multiple radiology clinics and hospitals within a large radiology
58	network in Australia between November and December, 2020.
59	Participants: Eleven consultant radiologists of general diagnostic and interventional
60	backgrounds, and varying levels of experience participated in this study.
61	Primary outcome measures: Proportion of CXR cases that had significant material changes to
62	the radiologist report, to patient management, or to imaging recommendations due to the model's
63	recommendations. Level of agreement between the radiologist and the model findings.
64	Results: Of 2,972 cases reviewed with the model, 92 cases (3.1%) had significant
65	report changes, 43 cases (1.4%) had changed patient management and 29 cases (1.0%) had further
66	imaging recommendations. In terms of agreement with the model, 2,572 cases showed complete
67	agreement (86.5%). 390 (13%) cases had one or more findings rejected by the radiologist. There
68	were 16 findings across 13 cases (0.5%) that were deemed to be missed by the model.
69	Conclusions: Use of an AI model in a real-world reporting environment significantly improved
70	radiologist reporting and showed good agreement with radiologists, highlighting the potential for AI
71	decision support to improve clinical practice.
72	
73	

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

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 • n of ca. 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].

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

7 109

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

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

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.

137 METHODS

139 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.

144 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.

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

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

		T 1 4 1 60 1	C 1
	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	
165			
166	Taskaisel Internetical		
167 168	Technical Integration	study technical integration of the	e software into existing radiology
108	Filor to the start of the	study, technical integration of the	e software into existing radiology
169	practice systems and testing occ	curred over several weeks. First,	an integration adapter was installed
170	on the IT network of each radio	logy clinic and acted as a gatewa	ay between the internal IT
171	infrastructure and the AI model	Auto routing rules were establi	shed ensuring only CXR studies were
1/1	initiastructure and the At model	. Auto-routing fules were establi	shed ensuring only CAR studies were
172	forwarded to the Integration Ad	lapter from the picture archiving	and communication system (PACS).
1 = 2			
173	Following a successful testing p	period, the Annalise CXR viewer	was installed and configured on
174	workstations for the group of st	udy radiologists.	
1,1	workbaatons for the group of be		
175			
170			
176	Study Participants		
177	Eleven consultant radio	logists working for a large Austr	alian radiology network were invited to
		6 6 6	
178	participate in the study through	their local radiologist network. 7	This group included a mix of general
170	1 1 1	1.1	
179	diagnostic and interventional ra	diologists who had completed sp	becialist radiology training. The group
180	included radiologists with a ran	ge of experience levels: five radi	ologists had 0–5 years post-training
100			
181	experience, three radiologists have	ad 6-10 years of experience, and	three radiologists had more than 10 yea
	of experience. Radiologists wer	re situated across four states in A	ustralia and worked in public hospitals,
182			
	private hospitals and communit	v alinia sattings. Writtan informe	ad consent was obtained from each
182 183	private hospitals and communit	y clinic settings. Written informe	ed consent was obtained from each
183			
			ed consent was obtained from each diologist attended a training seminar and

BMJ Open

2		
3 4	186	participating radiologists were able to familiarise themselves with the viewer prior to commencement of
5 6	187	data collection.
7 8	188	
9 10	189	CXR Case Selection
11 12	190	In this multicentre real-world prospective study, all consecutive chest radiographs reported by the
13 14	191	radiologists originating from inpatient, outpatient, and emergency settings were included for a period
15 16	192	covering nearly six weeks. The CXR cases were reported with the assistance of the AI tool in real-world
17 18	193	clinical practice, using high resolution diagnostic radiology monitors within the radiologists' normal
19 20 21	194	reporting environment.
22	195	
23 24 25	196	At least one frontal chest radiograph was required for analysis by the model, and cases that did
26 27	197	not include at least one were excluded. Chest radiographs from patients aged younger than 16 years were
28 29	198	excluded, as the CXR viewer has not been validated in these patients. Data from all sources was de-
30 31	199	identified for analysis.
32 33	200	
34 35	201	AI-Assisted Reporting
36 37	202	For each CXR case, radiologists produced their clinical report with access to clinical information,
38 39	203	the referral and available patient history, in line with the normal workflow. Model output was displayed
40 41 42	204	to the radiologist in a customised image viewer, linked to the image in the PACS, automatically
42 43 44	205	launching when a CXR case was opened (Figure 1)
45 46	206	
47 48	207	The modified version of the commercially available AI software gathered feedback from
49 50	208	radiologists during the reporting process. For each case, the model provided a list of suggested findings,
51 52	209	listed as "priority" or "other", along with a confidence indicator and, in some cases, a region of interest
53 54	210	localiser overlayed on the image. The CXR viewer was configured to display its findings after the
55 56 57	211	radiologists initial read of the case. For each case, the radiologist was asked to review the CXR viewer's
57 58 59	212	findings and provide feedback within the viewer. The options presented to the radiologists in the viewer
60	213	are listed in Table 2.

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

REVIEW OPTION		DESCRIPTION
Rejected clinical fin	ding	A model-detected finding disputed by the radiologist
Missed clinical find	ng	A model-detected finding missed by the radiologist
Add additional findi	ngs	Finding(s) identified by the radiologist but not identified by the
These findings sign impacted my report	ificantly	A yes/no binary question relating to the effect of the model out the radiologist report
These findings may patient managemen	-	A yes/no binary question relating to the effect of the model out patient management, as perceived by the reporting radiologist
These findings led t imaging recommend		A binary yes/no question related to whether the radiologist recommended further imaging based on the model output
The outcome	measure of 'sig	nificant impact on the report' was the primary outcome measure
A significant change v	vas described a	s the inclusion of findings recommended by the model, which
altered the radiologists	report in a me	aningful way. As this varied by patient and clinical setting, it
was left to the discretion	on of the radiol	ogist. For example, missing a pneumothorax in a ventilated ICU
patient with known pn	eumothorax we	ould not have the same significance as a previously unknown
pneumothorax in an o	ıtpatient. Durir	ng the analysis of radiologist feedback, it was assumed that a
change in patient man	agement or furt	her imaging recommendation would not occur without
radiologists indicating	a material char	nge in the CXR report, and thus management and imaging
questions were depend	ent on a signif	icant change in the report. Free text input describing missed
findings or other relev	ant data were n	nanually added after data collection was complete.

BMJ Open

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

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

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.

278 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)	Interventi onal?	Report changes (%)	Patient management changes (%)	Imaging reco mmendations (%)
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)
		2,972		92 (3.1)	43 (1.4)	29 (1.0)

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

To

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

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.

1				
1 2				
3 4	306			
5	307	Table 4 - Findings added by the radiolog	rist, and their respective counts. Crit	tical findings are highlighted.
6 7		Fi	nding Added	Count
8			electasis	4
9 10		So	litary Lung Nodule	3
11			urdiac valve prosthesis	2
12		So	litary Lung Mass	1
13 14		Pn	eumomediastinum	1
15		Pn	eumothorax	1
16		Sp	inal Wedge Fracture	1
17		Pu	lmonary Congestion	1
18 19			ribronchial Thickening	1
20		Su	bdiaphragmatic Gas	1
21	308			
22 23	309 310			
24	510			
25	311	Factors influencing reporting, 1	nanagement, or imaging rec	commendation
26 27				
27	312	The number of critical fin	ndings displayed by the mode	l was significantly higher in cases where
29	313	there was a change in report pati	ent management or imaging i	recommendation ($p < 0.001, p = 0.001, p =$
30 31	010			
32	314	0.004; Table 5). The presence of	a lateral projection image in the	he CXR case interpreted by the model was
33 34	315	associated with a significantly gro	eater likelihood of changes to	imaging recommendation ($p = 0.005$), but
35 36	316	not to the report or patient manag	ement ($p = 0.105$ and $p = 0.00$	61, respectively).
37			u i i	
38	317			
39 40	318	Radiologists with fewer t	han 5 years consultant experie	ence contributed 1,347 cases, and indicated
41		c .	•	
42 43	319	a rate of 5.0% for significant repo	ort change, 2.4% patient mana	gement change, and 1.5%
44 45	320	recommendations for further image	ging. These numbers were hig	gher than for the radiologists with 6-10
46 47	321	years of experience (1.3%, 0.4%,	0.5% respectively over 748 c	ases) and also for radiologists with greater
48	322	than 10 years of experience (1.6%	%, 0.9%, 0.6% over 877 cases). However, a likelihood ratio test applied
49 50			· · · · ,	
51	323	to binomial logistic regression an	alysis indicated that the level	of radiologist experience did not
52 53	324	significantly influence the rate of	change in report, patient man	agement, or imaging recommendation ($p =$
54 55	325	0.120, p = 0.262, and p = 0.516, r	respectively). Whether a pati	ent was imaged as an inpatient or
56 57	326	outpatient was not significantly a	ssociated with any change in	report, patient management, or imaging
58				
59 60	327	recommendation ($p = 0.358, p = 0$	0.572, p = 0.326, respectively).
00				

Table 5 - 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 E Value	Benjamini-Adjuste Threshold	^d 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

53 334 **Survey Results** 54

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

 $\overline{3}\overline{3}0$

BMJ Open

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.

to been teries only

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

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

considered an acceptable value. Furthermore, these studies report false-positive rates for CXR models

case reported in other studies investigating CXR models, which range from 14 - 88% [14,45,46], it is

BMJ Open

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

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

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

401 radiologists report having little knowledge of AI [48].

403 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.

416 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.

426 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

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

450 work through free provision of the model to participating radiologists for the duration of the study and

451 financing of an external biostatistician.

1 2 3	450	
4 5	452	DATIENT AND DUDI IC INVOLVEMENT
6 7	453	PATIENT AND PUBLIC INVOLVEMENT
8 9	454	Patients and public were not involved in the design, conduct, or reporting of this study.
10 11	455	
12 13	456	DATA AVAILABILITY STATEMENT
14 15 16	457	All data relevant to the study are included in the article or uploaded as online supplemental
10 17 18	458	information. No additional data are available.
$\begin{array}{c} 19\\ 20\\ 21\\ 22\\ 23\\ 24\\ 25\\ 26\\ 27\\ 28\\ 9\\ 30\\ 31\\ 32\\ 33\\ 34\\ 35\\ 36\\ 37\\ 38\\ 9\\ 41\\ 42\\ 43\\ 445\\ 46\\ 7\\ 48\\ 9\\ 50\\ 1\\ 52\\ 53\\ 56\\ 57\\ 58\\ 9\\ 60\\ \end{array}$	459	

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

2			
3 4	460	R	eferences
4 5	461	1	Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are
6	462	1	Data. <i>Radiology</i> 2016; 278 :563–77. doi:10.1148/radiol.2015151169
7	102		Duu. Ruulology 2010,270.305 77. doi:10.1110/100101.2013131105
8	463	2	Greene R. Francis H. Williams, MD: father of chest radiology in North America.
9 10	464		<i>RadioGraphics</i> 1991;11:325–32. doi:10.1148/radiographics.11.2.2028067
11			
12	465	3	Schaefer-Prokop C, Neitzel U, Venema HW, et al. Digital chest radiography: an update
13	466		on modern technology, dose containment and control of image quality. Eur Radiol
14	467		2008; 18 :1818–30. doi:10.1007/s00330-008-0948-3
15 16	468	4	Lee CS, Nagy PG, Weaver SJ, et al. Cognitive and System Factors Contributing to
17	468	4	Diagnostic Errors in Radiology. American Journal of Roentgenology 2013;201:611–7.
18	409		doi:10.2214/AJR.12.10375
19	470		doi:10.2214/15R.12.10575
20	471	5	Chotas HG, Ravin CE. Chest radiography: estimated lung volume and projected area
21 22	472		obscured by the heart, mediastinum, and diaphragm. Radiology 1994;193:403-4.
23	473		doi:10.1148/radiology.193.2.7972752
24			
25	474	6	Berlin L. Accuracy of Diagnostic Procedures: Has It Improved Over the Past Five
26 27	475		Decades? American Journal of Roentgenology 2007;188:1173-8.
27	476		doi:10.2214/AJR.06.1270
29	477	7	Zaorsky NG, Churilla TM, Egleston BL, et al. Causes of death among cancer patients.
30	478	/	Annals of Oncology 2017;28:400–7. doi:10.1093/annonc/mdw604
31	470		Annuis of Oncology 2017,20.400-7. doi.10.1095/annone/indw004
32 33	479	8	del Ciello A, Franchi P, Contegiacomo A, et al. Missed lung cancer: when, where, and
34	480		why? Diagn Interv Radiol 2017;23:118-26. doi:10.5152/dir.2016.16187
35			
36	481	9	Fazal MI, Patel ME, Tye J, et al. The past, present and future role of artificial intelligence
37 38	482		in imaging. European Journal of Radiology 2018;105:246–50.
39	483		doi:10.1016/j.ejrad.2018.06.020
40	484	10	Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science
41	484 485	10	2015; 349 :255–60. doi:10.1126/science.aaa8415
42	403		2015, 547 .255–60. doi:10.1126/science.ada8415
43 44	486	11	Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology. Nat Rev
45	487		Cancer 2018;18:500–10. doi:10.1038/s41568-018-0016-5
46			
47	488	12	Erickson BJ, Korfiatis P, Akkus Z, et al. Machine Learning for Medical Imaging.
48 40	489		RadioGraphics 2017;37:505-15. doi:10.1148/rg.2017160130
49 50	400	10	
51	490	13	Esteva A, Chou K, Yeung S, <i>et al.</i> Deep learning-enabled medical computer vision. <i>npj</i>
52	491		<i>Digital Medicine</i> 2021; 4 :1–9. doi:10.1038/s41746-020-00376-2
53	492	14	Jang S, Song H, Shin YJ, et al. Deep Learning–based Automatic Detection Algorithm for
54 55	493	14	Reducing Overlooked Lung Cancers on Chest Radiographs. <i>Radiology</i>
55 56	494		2020; 296 :652–61. doi:10.1148/radiol.2020200165
57	• •		· · , · · · · · · · · · · · · · · · · ·
58			
59 60			
60			

2			
3 4 5 6	495 496 497	15	Liang C-H, Liu Y-C, Wu M-T, <i>et al.</i> Identifying pulmonary nodules or masses on chest radiography using deep learning: external validation and strategies to improve clinical practice. <i>Clinical Radiology</i> 2020; 75 :38–45. doi:10.1016/j.crad.2019.08.005
7 8 9 10	498 499	16	Hurt B, Kligerman S, Hsiao A. Deep Learning Localization of Pneumonia: 2019 Coronavirus (COVID-19) Outbreak. <i>J Thorac Imaging</i> 2020; 35 :W87–9.
11 12 13 14	500 501 502	17	Kim JY, Choe PG, Oh Y, <i>et al.</i> The First Case of 2019 Novel Coronavirus Pneumonia Imported into Korea from Wuhan, China: Implication for Infection Prevention and Control Measures. <i>J Korean Med Sci</i> 2020; 35 . doi:10.3346/jkms.2020.35.e61
15 16 17 18 19	503 504 505	18	Bassi PRAS, Attux R. A Deep Convolutional Neural Network for COVID-19 Detection Using Chest X-Rays. <i>arXiv:200501578 [cs, eess]</i> Published Online First: 12 January 2021.http://arxiv.org/abs/2005.01578 (accessed 23 Mar 2021).
20 21 22 23 24	506 507 508 509	19	Rueckel J, Trappmann L, Schachtner B, <i>et al.</i> Impact of Confounding Thoracic Tubes and Pleural Dehiscence Extent on Artificial Intelligence Pneumothorax Detection in Chest Radiographs. <i>Investigative Radiology</i> 2020; 55 :792–8. doi:10.1097/RLI.000000000000707
25 26 27 28 29 30	510 511 512 513	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. <i>Image Analysis and Recognition</i> . Cham: : Springer International Publishing 2019. 325–32. doi:10.1007/978-3-030-27272-2_28
31 32 33 34	514 515 516	21	Hwang EJ, Hong JH, Lee KH, <i>et al.</i> Deep learning algorithm for surveillance of pneumothorax after lung biopsy: a multicenter diagnostic cohort study. <i>Eur Radiol</i> 2020; 30 :3660–71. doi:10.1007/s00330-020-06771-3
35 36 37 38 39	517 518 519	22	Park S, Lee SM, Kim N, <i>et al.</i> Application of deep learning–based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy. <i>Eur Radiol</i> 2019; 29 :5341–8. doi:10.1007/s00330-019-06130-x
40 41 42 43	520 521 522	23	Wang X, Yu J, Zhu Q, <i>et al.</i> Potential of deep learning in assessing pneumoconiosis depicted on digital chest radiography. <i>Occup Environ Med</i> 2020;77:597–602. doi:10.1136/oemed-2019-106386
44 45 46 47	523 524	24	S Z, X Z, R Z. Identifying Cardiomegaly in ChestX-ray8 Using Transfer Learning. <i>Stud Health Technol Inform</i> 2019; 264 :482–6. doi:10.3233/shti190268
48 49 50 51	525 526 527	25	Zou X-L, Ren Y, Feng D-Y, <i>et al.</i> A promising approach for screening pulmonary hypertension based on frontal chest radiographs using deep learning: A retrospective study. <i>PLOS ONE</i> 2020; 15 :e0236378. doi:10.1371/journal.pone.0236378
52 53 54 55	528 529 530	26	Pasa F, Golkov V, Pfeiffer F, <i>et al.</i> Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. <i>Scientific Reports</i> 2019; 9 :6268. doi:10.1038/s41598-019-42557-4
56 57 58 59 60	531 532 533	27	Nash M, Kadavigere R, Andrade J, <i>et al.</i> Deep learning, computer-aided radiography reading for tuberculosis: a diagnostic accuracy study from a tertiary hospital in India. <i>Scientific Reports</i> 2020; 10 :210. doi:10.1038/s41598-019-56589-3

1 2			
- 3 4 5 6 7 8	534 535 536 537	28	Heo S-J, Kim Y, Yun S, <i>et al.</i> Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers' Health Examination Data. <i>International Journal of Environmental Research and Public Health</i> 2019; 16 :250. doi:10.3390/ijerph16020250
9 10 11 12 13	538 539 540 541	29	Qin ZZ, Sander MS, Rai B, <i>et al.</i> Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. <i>Scientific Reports</i> 2019; 9 :15000. doi:10.1038/s41598-019-51503-3
14 15 16 17 18	542 543 544	30	Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. <i>Radiology</i> 2017; 284 :574–82. doi:10.1148/radiol.2017162326
19 20 21 22	545 546 547	31	Seah J, Tang C, Buchlak QD, <i>et al.</i> Radiologist chest X-ray diagnostic accuracy performance improvements when augmented by a comprehensive deep learning model. <i>In press</i> 2021.
23 24 25	548 549	32	Annalise.ai - Annalise CXR comprehensive medical imaging AI. Annalise.ai. https://annalise.ai/products/annalise-cxr/ (accessed 23 Mar 2021).
26 27 28 29 30	550 551 552	33	Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. <i>arXiv:190511946 [cs, stat]</i> Published Online First: 11 September 2020.http://arxiv.org/abs/1905.11946 (accessed 30 Mar 2021).
31 32 33 34	553 554 555	34	Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. <i>arXiv:150504597 [cs]</i> Published Online First: 18 May 2015.http://arxiv.org/abs/1505.04597 (accessed 30 Mar 2021).
35 36 37 38 39	556 557 558	35	Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. <i>Journal of the Royal Statistical Society Series B</i> (<i>Methodological</i>) 1995; 57 :289–300.
40 41 42	559 560	36	Mckinney W. pandas: a Foundational Python Library for Data Analysis and Statistics. <i>Python High Performance Science Computer</i> 2011.
43 44 45	561 562	37	Harris CR, Millman KJ, van der Walt SJ, <i>et al.</i> Array programming with NumPy. <i>Nature</i> 2020; 585 :357–62. doi:10.1038/s41586-020-2649-2
46 47	563	38	Jones E, Oliphant T, Peterson P. SciPy: Open Source Scientific Tools for Python. 2001.
48 49 50 51 52	564 565 566	39	Pedregosa F, Varoquaux G, Gramfort A, <i>et al.</i> Scikit-learn: Machine Learning in Python. <i>Journal of Machine Learning Research</i> Published Online First: 12 October 2011.https://hal.inria.fr/hal-00650905 (accessed 23 Mar 2021).
52 53 54 55 56 57 58 59 60	567 568	40	Jolly E. Pymer4: Connecting R and Python for linear mixed modeling. <i>Journal of Open Source Software</i> 2018; 3 :862.
	569 570	41	Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. Austin, Texas: 2010. 92–6. doi:10.25080/Majora-92bf1922-011

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

1 2			
3 4 5 6 7 8	571 572 573 574	42	Hwang EJ, Park S, Jin K-N, <i>et al.</i> Development and Validation of a Deep Learning- Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. <i>JAMA Netw Open</i> 2019; 2 :e191095. doi:10.1001/jamanetworkopen.2019.1095
8 9 10 11 12	575 576 577	43	Hwang EJ, Nam JG, Lim WH, <i>et al.</i> Deep Learning for Chest Radiograph Diagnosis in the Emergency Department. <i>Radiology</i> 2019; 293 :573–80. doi:10.1148/radiol.2019191225
13 14 15 16	578 579 580	44	Singh R, Kalra MK, Nitiwarangkul C, <i>et al.</i> Deep learning in chest radiography: Detection of findings and presence of change. <i>PLOS ONE</i> 2018; 13 :e0204155. doi:10.1371/journal.pone.0204155
17 18 19 20 21	581 582 583	45	Dellios N, Teichgraeber U, Chelaru R, <i>et al.</i> Computer-aided Detection Fidelity of Pulmonary Nodules in Chest Radiograph. <i>J Clin Imaging Sci</i> 2017; 7 . doi:10.4103/jcis.JCIS_75_16
22 23 24 25	584 585 586	46	Sim Y, Chung MJ, Kotter E, <i>et al.</i> Deep Convolutional Neural Network–based Software Improves Radiologist Detection of Malignant Lung Nodules on Chest Radiographs. <i>Radiology</i> Published Online First: 12 November 2019. doi:10.1148/radiol.2019182465
26 27 28 29 30	587 588 589	47	Waymel Q, Badr S, Demondion X, <i>et al.</i> Impact of the rise of artificial intelligence in radiology: What do radiologists think? <i>Diagnostic and Interventional Imaging</i> 2019; 100 :327–36. doi:10.1016/j.diii.2019.03.015
31 32 33 34	590 591 592	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. <i>Journal of the American College of Radiology</i> 2018; 15 :1753–7. doi:10.1016/j.jacr.2017.12.021
35 36 37 38	593 594		
39 40 41 42			
43 44 45			
46 47 48 49			
50 51 52			
53 54 55 56			
57 58 59 60			

FIGURE LEGENDS

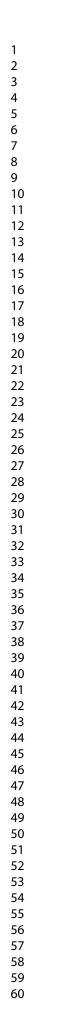
597 Figure 1 - Flow diagram illustrating the AI-assisted reporting process described in this study. (RIS: Radiological information system)

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.

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

37 BMJ Open: first published as 10.1136/bmjopen-2021-052902 on 20 December 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies. Page Pa



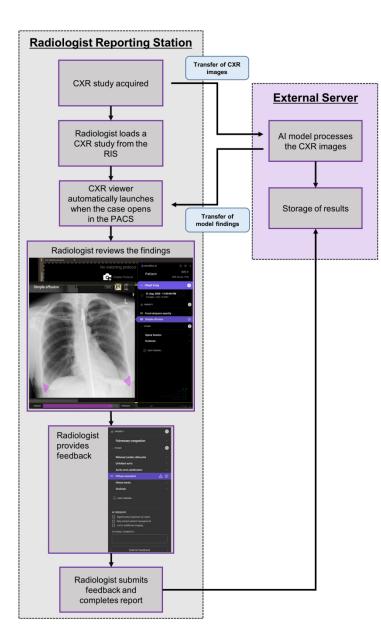


Figure 1 - Flow diagram illustrating the AI-assisted reporting process described in this study. RIS: Radiological information system.

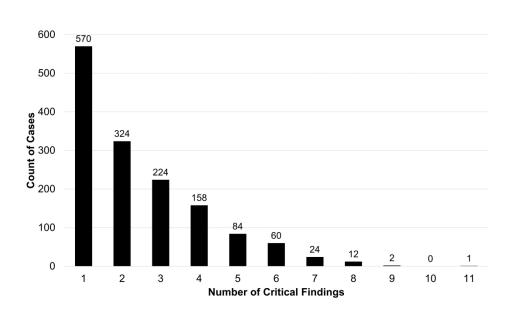
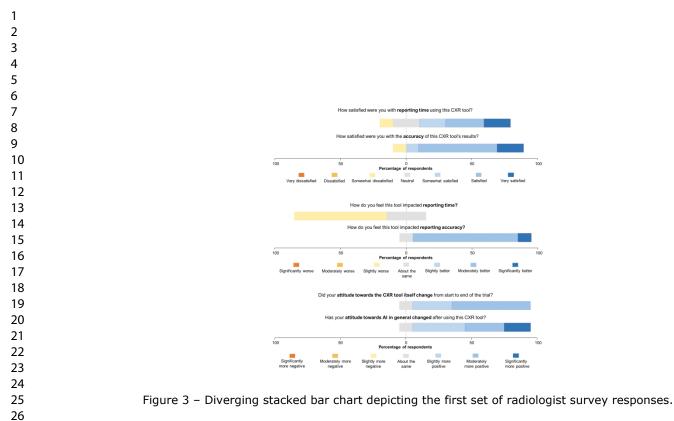


Figure 2 - Counts of numbers of critical findings for the cases seen by the radiologist, defined as the number of critical findings added. The number of cases which returned zero findings was 1,513.

BMJ Open: first published as 10.1136/bmjopen-2021-052902 on 20 December 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de I Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.



BMJ Open

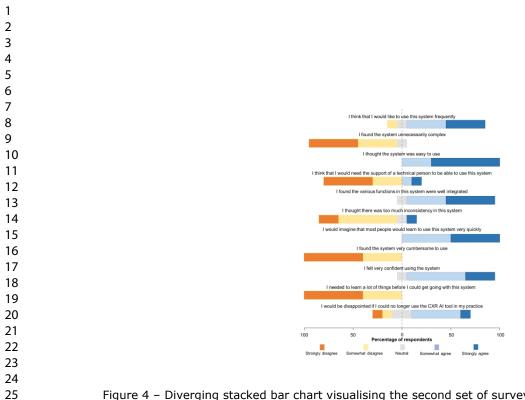


Figure 4 – Diverging stacked bar chart visualising the second set of survey responses of the radiologists.

Supplementary Table 1 - List of the 124 findings, including 34 critical findings which the model is validated to detect. ETT:

Enseignement Superieur (ABES) Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

	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 Enlargem
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 Interstit
Clavicle Fixation	Osteopaenia	Upper Zone Fibrotic Volum Loss
Clavicle Lesion	Pectus Carinatum	

Page 33 of 37

Hyperinflation Overexposed Underinflation Patient Rotation	Chest Incompletely Imaged Image Obscured Underexposed Hyperinflation Patient Rotation Patient Rotation	Chest Incompletely ImagedImage ObscuredUnderexposedHyperinflationOverexposedUnderinflationPatient RotationPatient Rotation

Enseignement Superieur (ABES) Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

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?	0	0	0	0	0	0	0
How do you feel this tool impacted reporting accuracy?	ο	ο	ο	ο	ο	0	0
	Very dissatisfied	Dissatisfied	Somewhat dissatisfied	Neutral	Somewhat satisfied	Satisfied	Very dissatisfied
How satisfied were you with reporting time using this CXR tool?	0	0	0	0	0	0	0
How satisfied were you with the accuracy of this CXR tool's results?	0	ο	ο	0	0	0	0
	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?	0	0	0	0	0	0	0
Has your attitude towards Al in general changed after using this CXR tool?	0	0	0	0	0	0	ο
		Strongly disagree	Somewhat disagree	Neutral	Somewhat agree	Strongly agree	
I think that I would like to use this system frequently.		0	0	ο	ο	0	
I found the system unnecessarily complex.		ο	o	ο	ο	0	
I thought the system was easy to use.		0	0	ο	ο	0	
I think that I would need the support of a technical person to be able to use this system.		0	o	0	ο	ο	
I found the various functions in this system were well integrated.		0	ο	ο	ο	0	
I thought there was too much inconsistency in this system.		0	0	ο	0	0	
I would imagine that most people would learn to use this system very quickly.		ο	ο	ο	ο	0	
I found the system very cumbersome to use.		ο	ο	ο	ο	ο	
I felt very confident using the system.		0	0	ο	ο	0	
I needed to learn a lot of things before I could get going with this system.		0	ο	ο	ο	0	

1	
2	
3	
4	
5	
ć	
6	
7	
8	
9	
10	
10	
11	
12	
13	
13 14	
15	
15 16	
17	
17	
18	
19	
20	
21	
22	
22	
23	
24	
25	
24 25 26	
27	
28	
Zŏ	
29	
30	
31	
32	
33	
24	
34 35 36	
35	
36	
37	
38	
39	
40	
41	
42	
43	
44	
45	
46	
47	
48	
49	
50	
50 51	
52	
53	
54	
55	
56	
50 57	
58	
50	

I would be disappointed if I could no longer use the CXR AI tool in my practice.	0	0	0	0	0

Section / Topic	No.	ltem	
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 Da		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: "Al- 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

³⁷ BMJ Open: first published as 10.1136/bmjopen-2021-052902 on 20 December 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de I
 Enseignement Superieur (ABES) .
 Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

Data Partitions	19	Intended sample size and how it was determined	Yes – page 10 (under: "statistics and data analys
	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: "moo development and validatio and described in reference
	23	Software libraries, frameworks, and packages	Yes – page 6 (under: "moo development and validatio and described in reference
	24	Initialization of model parameters (e.g., randomization, transfer learning)	Yes – page 6 (under: "moo development and validatio and described in reference
Training	25	Details of training approach, including data augmentation, hyperparameters, number of models trained	Yes – page 6 (under: "moo development and validatio and described in reference
	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: "moo development and validatio and described in reference
	29	Statistical measures of significance and uncertainty (e.g., confidence intervals)	Yes – page 6 (under: "moo development and validatio and described in reference
	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. <u>https://doi.org/10.1148/ryai.2020200029</u>

RSNA[®]

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 <u>licence</u>.

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 <u>Creative Commons</u> 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.

terez oni

Enseignement Superieur (ABES) Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies



2 3 4	1	Assessment of the effect of a comprehensive chest
5 6 7	2	radiograph deep learning model on radiologist
, 8 9	3	reports and patient outcomes: a real-world
10 11	4	observational study
12 13	5	
14	6	
15	7	Catherine M Jones ^{1,2} , Luke Danaher ² , Michael R Milne ^{1,2*} , Cyril Tang ¹ , Jarrel Seah ^{1,3} , Luke
16 17	8	Oakden-Rayner ⁴ , Andrew Johnson ¹ , Quinlan D Buchlak ^{1,5} , Nazanin Esmaili ^{5,6}
18		Oakuch-Rayner, Andrew Johnson, Quintan D Buchak *, Wazanin Esmani**
19	9	
20	10	¹ Annalise-AI, Sydney, NSW, Australia
21	11	² I-MED Radiology Network, Sydney, NSW, Australia
22	12	³ Department of Radiology, Alfred Health, Melbourne, VIC, Australia
23	13	⁴ Australian Institute for Machine Learning, The University of Adelaide, Adelaide, SA, Australia
24	14	⁵ School of Medicine, University of Notre Dame Australia, Sydney, NSW, Australia
25	15 16	⁶ Faculty of Engineering and IT, University of Technology Sydney, Ultimo, NSW, Australia
26 27	10	
28	17	*Commente de michael mila Querrelia ei
29		*Correspondence to: michael.milne@annalise.ai
30	19	
31	20	
32	21	
33	22	
34	23	
35	24	Corresponding author:
36 37	25	Name: Michael Milne
38	26	Annalise-AI
39	27	Sydney, Australia
40	28	E-mail: michael.milne@annalise.ai
41	29	
42	30	
43	31	
44	32	
45 46	33	
46 47	34	
48	35	
49	36	
50	37	
51	38	
52	39	Keywords: Machine learning; chest X-ray, deep learning.
53	40	
54	41	Word Count: 4,426
55 56		
57		
58		
59		
60		

42	ABSTRACT
43 44	Objectives: AI algorithms have been developed to detect imaging features on chest X-ray (CXR)
45	with a comprehensive AI model capable of detecting 124 CXR findings being recently developed. The
46	aim of this study was to evaluate the real-world usefulness of the model as a diagnostic assistance device
47	for radiologists.
48	Design: This prospective real-world multicentre study involved a group of radiologists using the
49	model in their daily reporting workflow to report consecutive chest X-rays and recording their feedback
50	on level of agreement with the model findings and whether this significantly affected their reporting.
51	Setting: The study took place at radiology clinics and hospitals within a large radiology network
52	in Australia between November and December 2020.
53	Participants: Eleven consultant diagnostic radiologists of varying levels of experience
54	participated in this study.
55	Primary and secondary outcome measures: Proportion of CXR cases where use of the AI
56	model led to significant material changes to the radiologist report, to patient management, or to imaging
57	recommendations. Additionally, level of agreement between radiologists and the model findings, and
58	radiologist attitudes towards the model were assessed.
59	Results: Of 2,972 cases reviewed with the model, 92 cases (3.1%) had significant
60	report changes, 43 cases (1.4%) had changed patient management and 29 cases (1.0%) had further
61	imaging recommendations. In terms of agreement with the model, 2,572 cases showed complete
62	agreement (86.5%). 390 (13%) cases had one or more findings rejected by the radiologist. There
63	were 16 findings across 13 cases (0.5%) deemed to be missed by the model. Nine out of 10 radiologists
64	felt their accuracy was improved with the model and were more positive towards AI post-study.
65	Conclusions: Use of an AI model in a real-world reporting environment significantly improved
66	radiologist reporting and showed good agreement with radiologists, highlighting the potential for AI
67	diagnostic support to improve clinical practice.
68	
69	

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

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 • yn of ea... 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].

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

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

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

2	
3	
4	
5	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
16 17 18	
17	
18	
19	
20	
21	
22	
22	
25	
24	
25	
26	
27	
26 27 28	
29	
30	
30 31	
32	
33	
34 35	
35	
36	
37 38	
20	
39	
40	
41	
42	
43	
44	
45	
46	
47	
47 48	
49	
50	
51	
52	
53	
55 54	
55	
56	
57	
58	
59	
60	
00	

1

112 volume and availability of digital imaging datasets [11]. Of note is the rise of convolutional neural 113 networks (CNNs), a type of deep neural network that excels at image feature extraction and classification, 114 and demonstrates strong performance in medical image analysis, leading to the rapid advancement of 115 computer vision in medical imaging [12,13]. CNNs have been used to develop models to successfully 116 detect targeted clinical findings on CXR, including lung cancer [14,15], pneumonia [16,17], COVID-19 117 [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 118 119 interpretation, however most of these deep learning systems are limited in scope to a single finding or a 120 small set of findings, therefore lacking the broad utility that would make them useful in clinical practice. 121 122 Recently, our group developed a comprehensive deep learning CXR diagnostic assist device, 123 which was designed to assist clinicians in CXR interpretation and improve diagnostic accuracy, validated 124 for 124 clinically relevant findings seen on frontal and lateral chest radiographs [31]. The primary 125 objective of the current study was to evaluate the real-world usefulness of the model as a diagnostic assist 126 device for radiologists in both hospital and community clinic settings. This involved examining the 127 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

136 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.

142 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].

159Technical Integration160Prior to the start of the star

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

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

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.

168 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.

185 CXR Case Selection

In this multicentre real-world prospective study, all consecutive chest radiographs reported by the
 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

1 2		
3 4	190	reporting environment. As per usual workflow across a large radiology network spanning a
5 6	191	geographically large area with many regional and remote clinics, both on-site and remote reporting of
7 8	192	CXR cases was undertaken. A total of 106 sites contributed cases with case numbers varying from one
9 10	193	case up to a maximum of 271 cases at the busiest site.
11 12	194	
13 14	195	At least one frontal chest radiograph was required for analysis by the model, and cases that did
15 16	196	not include at least one were excluded. Chest radiographs from patients aged younger than 16 years were
17 18 19	197	excluded. Data from all sources was de-identified for analysis.
20 21	198	
22 23	199	AI-Assisted Reporting
24 25	200	For each CXR case, radiologists produced their clinical report with access to clinical information,
26 27	201	the referral and available patient history, in line with the normal workflow. The AI model analyses the
28 29	202	CXR image(s) for each case but does not incorporate clinical inputs (such as previous imaging, referral
30 31	203	information or patient demographic data) into the analysis. Model output was displayed to the radiologist
32 33	204	in a user interface, linked to the image in the PACS, automatically launching when a CXR case was
34 35 36	205	opened (Figure 1).
37 38	206	
39 40	207	A modified version of the commercially available AI software was employed for this study,
41 42	208	which incorporated changes into the user interface to allow radiologists to provide feedback on model
43 44	209	recommendations. No changes were made to the underlying model. An example of the modified model
45 46	210	user interface is presented in figure 2. For each case, the model provided a list of suggested findings,
47 48	211	listed as "priority" or "other", along with a confidence indicator. For a subset of findings, a region of
49 50	212	interest localiser was overlayed on the image and the model indicated whether the finding was on the left
51 52	213	or the right side, or both (see Supplementary Table 1). The CXR viewer was configured to display its
53 54 55	214	findings after the radiologists' initial read of the case. For each case, radiologists were asked to review the
56 57	215	CXR viewer's findings and provide feedback within the viewer. The options presented to the radiologists
58 59	216	in the viewer are listed in Table 1.
60	217	

Enseignement Superieur (ABES) Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

	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	
219 220			
221	The outcome measure of 'significant impact on the report' was the primary outcome measure.		
222	A significant change was described as the inclusion of findings recommended by the model, which		
223	altered the radiologists report in a meaningful way. As this varied by patient and clinical setting, it		
224	was left to the discretion of the radiologist. During the analysis of radiologist feedback, it was		
225	assumed that a change in patient management or further imaging recommendation would not occur		
226	without radiologists indicating a material change in the CXR report, and thus management and		
227	imaging questions were dependent on a significant change in the report. This was also patient-		
228	specific; for example, missing a pneumothorax in a ventilated patient with known pneumothorax		
229	would not have the same impact on patient management as a previously unknown pneumothorax in an		
230	outpatient. Free text input describing missed findings or other relevant data were manually added after		
231	data collection was complete.		
232	No formal adjudication of cases showing discrepancy between radiologist and model		
233	interpretation was performed. The study was not designed as a diagnostic accuracy validation. No		
234	review or ground truthing process was performed. Radiologists remained responsible for image		
235	interpretation and formulation of the report.		

1 2		
3 4	236	
5 6	237	Post-Study Survey
7 8	238	Upon completion of data collection, a post-study survey was distributed to all participating
9 10	239	radiologists to obtain feedback on the usefulness of the CXR viewer and how it affected their opinion of
11 12	240	AI in radiology. A table of the survey questions is presented in Supplementary Table 2.
13 14 15 16 17	241	
	242	Statistics and Data Analysis
17 18 19	243	A 1% rate of significant changes in reports (the primary outcome measure) was deemed to be
20 21	244	clinically significant prior to commencing the study. Based on estimations of the prevalence of missed
22 23	245	critical findings on CXR, preliminary power calculations estimated that the number of cases required to
24 25	246	detect at least a 1% rate of significant changes in reports was approximately 2000 cases in total, with
26 27	247	alpha value 0.05 and desired power of 0.90. To account for any dropout in radiologists or cases, a target
28 29	248	of 3000 cases was set for the study. Ten radiologists were recruited, with an eleventh included for any
30 31	249	unexpected participant drop out and to achieve this target in a reasonable time period.
31 32 33 34 35 36 37 38 39	250	
	251	A two-tailed binomial test was used to test the hypothesis that the rate of significant report
	252	change, patient management change, or imaging recommendation change was at least 1%. To ensure that
	253	the sampling of CXRs reasonably approximated a random snapshot of the true population, radiologists in
41 42	254	various states, experience levels as well as different conditions of practice (community clinic vs hospital
43 44	255	based) were selected. Additionally, the study was conducted prospectively which further aligned the
45 46	256	structure of the sampled data with the expected structure of the population, justifying the choice of
47 48	257	analysing the sample using a binomial test without adjustment for each radiologist.
50	258	Multivariate logistic regression using generalised linear mixed effect analysis was used to assess
45 46 47 48 49 50 51 52	259	the effect of several possible confounders on the measured outcomes, including the number of critical
53 54	260	clinical findings per case identified by the model, the inpatient/outpatient status of the patients, the
55 56 57	261	experience level of the radiologists, and the presence or absence of a lateral radiograph. The Wald test
57 58 59 60	262	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.

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)
tal			2,972	92 (3.1)	43 (1.4)	29 (1.0)

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

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).

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

315 316 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

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

324	not to the report or patient management ($p = 0.105$ and $p = 0.061$, respectively).
325	
326	Radiologists with fewer than 5 years consultant experience contributed 1,347 cases, and indicated
327	a rate of 5.0% for significant report change, 2.4% patient management change, and 1.5%
328	recommendations for further imaging. These numbers were higher than for the radiologists with 6-10
329	years of experience (1.3%, 0.4%, 0.5% respectively over 748 cases) and also for radiologists with greater
330	than 10 years of experience (1.6%, 0.9%, 0.6% over 877 cases). However, a likelihood ratio test applied
331	to binomial logistic regression analysis indicated that the level of radiologist experience did not
332	significantly influence the rate of change in report, patient management, or imaging recommendation ($p =$
333	0.120, $p = 0.262$, and $p = 0.516$, respectively). Whether a patient was imaged as an inpatient or
334	outpatient was not significantly associated with any change in report, patient management, or imaging
335	recommendation ($p = 0.358$, $p = 0.572$, $p = 0.326$, respectively).

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

R. R. ONL

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

52
53 342 Survey Results
54

55343The post-study survey was completed by ten out of the eleven radiologists (Figure 4 and Figure56573445). Notably, 7 (70%) participants felt that their reporting time was slightly worse, however when asked5859345how satisfied they were with their reporting time, 7 (70%) indicated that they were satisfied.

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

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.

tor occite iew only

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.

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

BMJ Open

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.

Page 21 of 40

BMJ Open

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].

428 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

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

434 power to examine the benefit of the model on a finding-by-finding basis. In future, it would be beneficial 435 to conduct a similar study with a larger sample size to allow for more powerful statistical analysis and 436 examination of specific finding changes. Another useful next step would be to include a gold standard to 437 determine the ground truth for the CXR findings, as this would prevent any under reporting which may 438 occur with self-reported results, as well as enable the detection of false negatives as a result of the CXR 439 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.

445 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.

455 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

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

479 provision of the model to participating radiologists for the duration of the study and financing of an

480 external biostatistician.

1 2		
3 4	481	
5 6 7	482	PATIENT AND PUBLIC INVOLVEMENT
, 8 9	483	Patients and public were not involved in the design, conduct, or reporting of this study.
10 11	484	
12 13 14	485	DATA AVAILABILITY STATEMENT
14 15 16	486	All data relevant to the study are included in the article or uploaded as online supplemental
10 17 18	487	information. No additional data are available.
19 20 21 22 23 25 26 27 28 20 31 23 34 35 37 38 90 41 23 44 45 46 47 89 51 52 53 45 56 78 90	488	information. No additional data are available.

2			
3	489	R	eferences
4 5	489	1	Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are
6	490 491	1	Data. <i>Radiology</i> 2016; 278 :563–77. doi:10.1148/radiol.2015151169
7	491		Data. <i>Rualology</i> 2010, 278 .303–77. doi:10.1148/18di01.2013131109
8	492	2	Greene R. Francis H. Williams, MD: father of chest radiology in North America.
9	493	2	RadioGraphics 1991;11:325–32. doi:10.1148/radiographics.11.2.2028067
10	175		nualo 6/ up/nes 1991,11.525 52. doi:10.1116/10060600000000000000000000000000
11 12	494	3	Schaefer-Prokop C, Neitzel U, Venema HW, et al. Digital chest radiography: an update
13	495		on modern technology, dose containment and control of image quality. Eur Radiol
14	496		2008; 18 :1818–30. doi:10.1007/s00330-008-0948-3
15			
16	497	4	Lee CS, Nagy PG, Weaver SJ, et al. Cognitive and System Factors Contributing to
17 18	498		Diagnostic Errors in Radiology. American Journal of Roentgenology 2013;201:611-7.
10	499		doi:10.2214/AJR.12.10375
20		-	
21	500	5	Chotas HG, Ravin CE. Chest radiography: estimated lung volume and projected area
22	501		obscured by the heart, mediastinum, and diaphragm. <i>Radiology</i> 1994; 193 :403–4.
23	502		doi:10.1148/radiology.193.2.7972752
24 25	503	6	Berlin L. Accuracy of Diagnostic Procedures: Has It Improved Over the Past Five
26	503	0	Decades? American Journal of Roentgenology 2007; 188 :1173–8.
27	505		doi:10.2214/AJR.06.1270
28	505		
29	506	7	Zaorsky NG, Churilla TM, Egleston BL, et al. Causes of death among cancer patients.
30 31	507		Annals of Oncology 2017;28:400-7. doi:10.1093/annonc/mdw604
32			
33	508	8	del Ciello A, Franchi P, Contegiacomo A, et al. Missed lung cancer: when, where, and
34	509		why? Diagn Interv Radiol 2017;23:118-26. doi:10.5152/dir.2016.16187
35			
36	510	9	Fazal MI, Patel ME, Tye J, et al. The past, present and future role of artificial intelligence
37 38	511		in imaging. European Journal of Radiology 2018; 105 :246–50.
39	512		doi:10.1016/j.ejrad.2018.06.020
40	513	10	Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science
41	515	10	2015; 349 :255–60. doi:10.1126/science.aaa8415
42	514		2015, 547 .255–60. doi:10.1120/science.add0415
43 44	515	11	Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology. Nat Rev
45	516		<i>Cancer</i> 2018; 18 :500–10. doi:10.1038/s41568-018-0016-5
46			
47	517	12	Erickson BJ, Korfiatis P, Akkus Z, et al. Machine Learning for Medical Imaging.
48	518		<i>RadioGraphics</i> 2017; 37 :505–15. doi:10.1148/rg.2017160130
49 50			
50 51	519	13	Esteva A, Chou K, Yeung S, et al. Deep learning-enabled medical computer vision. npj
52	520		Digital Medicine 2021;4:1–9. doi:10.1038/s41746-020-00376-2
53	501	14	Jana C. Cong H. Chin VI. et al. Deve Learning hand Aret (C. D. t. C. Al. 191 C.
54	521 522	14	Jang S, Song H, Shin YJ, <i>et al.</i> Deep Learning–based Automatic Detection Algorithm for Reducing Overlooked Lung Cancers on Chest Radiographs. <i>Radiology</i>
55			
56 57	523		2020; 296 :652–61. doi:10.1148/radiol.2020200165
58			
59			
60			

2			
3 4 5 6	524 525 526	15	Liang C-H, Liu Y-C, Wu M-T, <i>et al.</i> Identifying pulmonary nodules or masses on chest radiography using deep learning: external validation and strategies to improve clinical practice. <i>Clinical Radiology</i> 2020; 75 :38–45. doi:10.1016/j.crad.2019.08.005
7 8 9 10	527 528	16	Hurt B, Kligerman S, Hsiao A. Deep Learning Localization of Pneumonia: 2019 Coronavirus (COVID-19) Outbreak. <i>J Thorac Imaging</i> 2020; 35 :W87–9.
10 11 12 13 14	529 530 531	17	Kim JY, Choe PG, Oh Y, <i>et al.</i> The First Case of 2019 Novel Coronavirus Pneumonia Imported into Korea from Wuhan, China: Implication for Infection Prevention and Control Measures. <i>J Korean Med Sci</i> 2020; 35 . doi:10.3346/jkms.2020.35.e61
15 16 17 18 19	532 533 534	18	Bassi PRAS, Attux R. A Deep Convolutional Neural Network for COVID-19 Detection Using Chest X-Rays. <i>arXiv:200501578 [cs, eess]</i> Published Online First: 12 January 2021.http://arxiv.org/abs/2005.01578 (accessed 23 Mar 2021).
20 21 22 23 24	535 536 537 538	19	Rueckel J, Trappmann L, Schachtner B, <i>et al.</i> Impact of Confounding Thoracic Tubes and Pleural Dehiscence Extent on Artificial Intelligence Pneumothorax Detection in Chest Radiographs. <i>Investigative Radiology</i> 2020; 55 :792–8. doi:10.1097/RLI.000000000000707
25 26 27 28 29 30	539 540 541 542	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. <i>Image Analysis and Recognition</i> . Cham: : Springer International Publishing 2019. 325–32. doi:10.1007/978-3-030-27272-2_28
31 32 33 34	543 544 545	21	Hwang EJ, Hong JH, Lee KH, <i>et al.</i> Deep learning algorithm for surveillance of pneumothorax after lung biopsy: a multicenter diagnostic cohort study. <i>Eur Radiol</i> 2020; 30 :3660–71. doi:10.1007/s00330-020-06771-3
35 36 37 38 39	546 547 548	22	Park S, Lee SM, Kim N, <i>et al.</i> Application of deep learning–based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy. <i>Eur Radiol</i> 2019; 29 :5341–8. doi:10.1007/s00330-019-06130-x
40 41 42 43	549 550 551	23	Wang X, Yu J, Zhu Q, <i>et al.</i> Potential of deep learning in assessing pneumoconiosis depicted on digital chest radiography. <i>Occup Environ Med</i> 2020;77:597–602. doi:10.1136/oemed-2019-106386
44 45 46 47	552 553	24	S Z, X Z, R Z. Identifying Cardiomegaly in ChestX-ray8 Using Transfer Learning. <i>Stud Health Technol Inform</i> 2019; 264 :482–6. doi:10.3233/shti190268
48 49 50 51	554 555 556	25	Zou X-L, Ren Y, Feng D-Y, <i>et al.</i> A promising approach for screening pulmonary hypertension based on frontal chest radiographs using deep learning: A retrospective study. <i>PLOS ONE</i> 2020; 15 :e0236378. doi:10.1371/journal.pone.0236378
52 53 54 55	557 558 559	26	Pasa F, Golkov V, Pfeiffer F, <i>et al.</i> Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. <i>Scientific Reports</i> 2019; 9 :6268. doi:10.1038/s41598-019-42557-4
56 57 58 59 60	560 561 562	27	Nash M, Kadavigere R, Andrade J, <i>et al.</i> Deep learning, computer-aided radiography reading for tuberculosis: a diagnostic accuracy study from a tertiary hospital in India. <i>Scientific Reports</i> 2020; 10 :210. doi:10.1038/s41598-019-56589-3

1 2			
3 4 5 6 7 8 9 10 11 12 13	563 564 565 566	28	Heo S-J, Kim Y, Yun S, <i>et al.</i> Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers' Health Examination Data. <i>International Journal of Environmental Research and Public Health</i> 2019; 16 :250. doi:10.3390/ijerph16020250
	567 568 569 570	29	Qin ZZ, Sander MS, Rai B, <i>et al.</i> Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. <i>Scientific Reports</i> 2019; 9 :15000. doi:10.1038/s41598-019-51503-3
14 15 16 17 18	571 572 573	30	Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. <i>Radiology</i> 2017; 284 :574–82. doi:10.1148/radiol.2017162326
19 20 21 22 23	574 575 576 577	31	Seah JCY, Tang CHM, Buchlak QD, <i>et al.</i> Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. <i>The Lancet Digital Health</i> 2021; 3 :e496–506. doi:10.1016/S2589-7500(21)00106-0
24 25 26 27	578 579	32	Annalise.ai - Annalise CXR comprehensive medical imaging AI. Annalise.ai. https://annalise.ai/products/annalise-cxr/ (accessed 23 Mar 2021).
27 28 29 30 31	580 581 582	33	Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. <i>arXiv:190511946 [cs, stat]</i> Published Online First: 11 September 2020.http://arxiv.org/abs/1905.11946 (accessed 30 Mar 2021).
32 33 34 35 36 37 38 39 40 41 42 43 44 45	583 584 585	34	Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. <i>arXiv:150504597 [cs]</i> Published Online First: 18 May 2015.http://arxiv.org/abs/1505.04597 (accessed 30 Mar 2021).
	586 587 588 589	35	xmlmillr6.pdf. https://www.ebs.tga.gov.au/servlet/xmlmillr6?dbid=ebs/PublicHTML/pdfStore.nsf&doci d=F7ADAEBB76CEDD47CA2585E500424A43&agid=(PrintDetailsPublic)&actionid=1 (accessed 25 Aug 2021).
	590 591 592	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).
46 47 48 49 50	593 594 595	37	Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. <i>Journal of the Royal Statistical Society Series B (Methodological)</i> 1995; 57 :289–300.
51 52 53	596 597	38	Mckinney W. pandas: a Foundational Python Library for Data Analysis and Statistics. <i>Python High Performance Science Computer</i> 2011.
54 55 56	598 599	39	Harris CR, Millman KJ, van der Walt SJ, <i>et al.</i> Array programming with NumPy. <i>Nature</i> 2020; 585 :357–62. doi:10.1038/s41586-020-2649-2
57 58 59 60	600	40	Jones E, Oliphant T, Peterson P. SciPy: Open Source Scientific Tools for Python. 2001.

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

2			
3 4 5 6	601 602 603	41	Pedregosa F, Varoquaux G, Gramfort A, <i>et al.</i> Scikit-learn: Machine Learning in Python. <i>Journal of Machine Learning Research</i> Published Online First: 12 October 2011.https://hal.inria.fr/hal-00650905 (accessed 23 Mar 2021).
7 8 9 10	604 605	42	Jolly E. Pymer4: Connecting R and Python for linear mixed modeling. <i>Journal of Open Source Software</i> 2018; 3 :862.
11 12 13	606 607	43	Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. Austin, Texas: 2010. 92–6. doi:10.25080/Majora-92bf1922-011
14 15 16 17	608 609 610	44	Seah J, Tang C, Buchlak QD, <i>et al.</i> Radiologist chest X-ray diagnostic accuracy performance improvements when augmented by a comprehensive deep learning model. <i>The Lancet Digital Health</i> 2021.
18 19 20 21 22 23	611 612 613 614	45	Hwang EJ, Park S, Jin K-N, <i>et al.</i> Development and Validation of a Deep Learning- Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. <i>JAMA Netw Open</i> 2019; 2 :e191095. doi:10.1001/jamanetworkopen.2019.1095
24 25 26 27	615 616 617	46	Hwang EJ, Nam JG, Lim WH, <i>et al.</i> Deep Learning for Chest Radiograph Diagnosis in the Emergency Department. <i>Radiology</i> 2019; 293 :573–80. doi:10.1148/radiol.2019191225
28 29 30 31 32	618 619 620	47	Singh R, Kalra MK, Nitiwarangkul C, <i>et al.</i> Deep learning in chest radiography: Detection of findings and presence of change. <i>PLOS ONE</i> 2018; 13 :e0204155. doi:10.1371/journal.pone.0204155
33 34 35 36 37	621 622 623 624	48	Khan FA, Majidulla A, Tavaziva G, <i>et al.</i> 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. <i>The Lancet Digital Health</i> 2020; 2 :e573–81. doi:10.1016/S2589-7500(20)30221-1
38 39 40 41 42	625 626 627	49	Dellios N, Teichgraeber U, Chelaru R, <i>et al.</i> Computer-aided Detection Fidelity of Pulmonary Nodules in Chest Radiograph. <i>J Clin Imaging Sci</i> 2017; 7 . doi:10.4103/jcis.JCIS_75_16
43 44 45 46	628 629 630	50	Sim Y, Chung MJ, Kotter E, <i>et al.</i> Deep Convolutional Neural Network–based Software Improves Radiologist Detection of Malignant Lung Nodules on Chest Radiographs. <i>Radiology</i> Published Online First: 12 November 2019. doi:10.1148/radiol.2019182465
47 48 49 50 51	631 632 633	51	Waymel Q, Badr S, Demondion X, <i>et al.</i> Impact of the rise of artificial intelligence in radiology: What do radiologists think? <i>Diagnostic and Interventional Imaging</i> 2019; 100 :327–36. doi:10.1016/j.diii.2019.03.015
52 53 54 55	634 635 636	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. <i>Journal of the American College of Radiology</i> 2018; 15 :1753–7. doi:10.1016/j.jacr.2017.12.021
56 57 58 59 60	637 638		

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.
- ,ictu., t visualising the s. Figure 5 – Diverging stacked bar chart visualising the second set of survey responses of the radiologists.



Radiologist Reporting Station Transfer of CXR images CXR study acquired **External Server** Radiologist loads a AI model processes CXR study from the the CXR images RIS CXR viewer automatically launches Storage of results when the case opens Transfer of model findings in the PACS Radiologist reviews the findings Radiologist provides feedback Radiologist submits feedback and completes report

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

484x610mm (118 x 118 DPI)

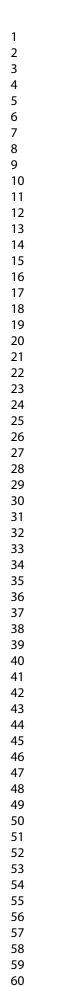
$1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14 \\ 15 \\ 16 \\ 17 \\ 18 \\ 19 \\ 20 \\ 21 \\ 22 \\ 23 \\ 24 \\ 25 \\ 26 \\ 27 \\ 28 \\ 29 \\ 30 \\ 31 \\ 32 \\ 33 \\ 34 \\ 35 \\ 36 \\ 37 \\ 38 \\ 39 \\ 40 \\ 41 \\ 42 \\ 43 \\ 44 \\ 45 \\ 46 \\ 47 \\ 48 \\ 49 \\ 49 \\ 41 \\ 45 \\ 46 \\ 47 \\ 48 \\ 49 \\ 40 \\ 41 \\ 45 \\ 46 \\ 47 \\ 48 \\ 49 \\ 40 \\ 41 \\ 45 \\ 46 \\ 47 \\ 48 \\ 49 \\ 40 \\ 41 \\ 45 \\ 46 \\ 47 \\ 48 \\ 49 \\ 40 \\ 41 \\ 45 \\ 46 \\ 47 \\ 48 \\ 49 \\ 40 \\ 40 \\ 40 \\ 40 \\ 40 \\ 40 \\ 40$	Figure 2 – Example o bc

60

✓ Priority findings	1
	1
< Pulmonary congestion	
> other	6
+ ADD FINDING	
AI FEEDBACK	
Significantly impacted my report	
May impact patient management	
Led to additional imaging	
OPTIONAL COMMENTS	
Submit feedback	

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)



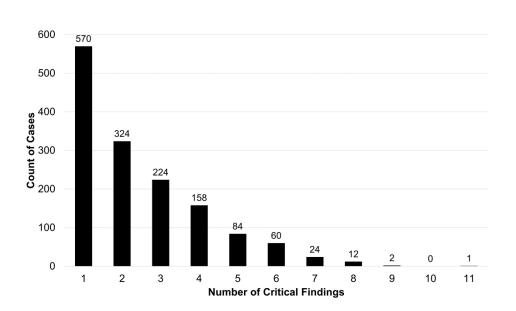


Figure 3 – Counts of numbers of critical findings for the cases seen by the radiologist, defined as the number of critical findings agreed + the number of critical findings added. The number of cases which returned zero findings was 1,513.

861x484mm (118 x 118 DPI)

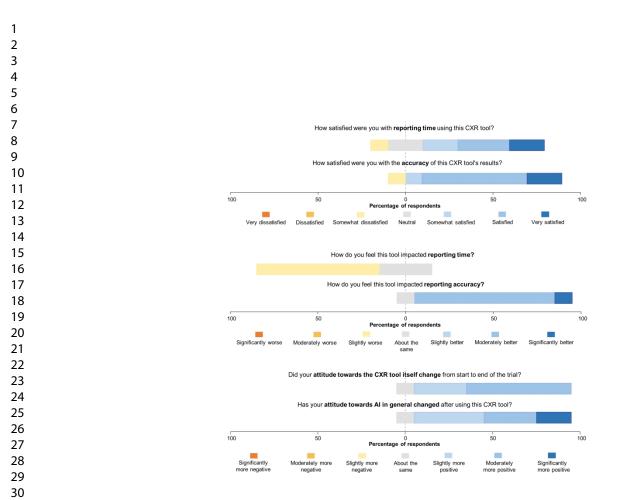


Figure 4 - Diverging stacked bar chart depicting the first set of radiologist survey responses.

645x484mm (118 x 118 DPI)

BMJ Open: first published as 10.1136/bmjopen-2021-052902 on 20 December 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de I Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

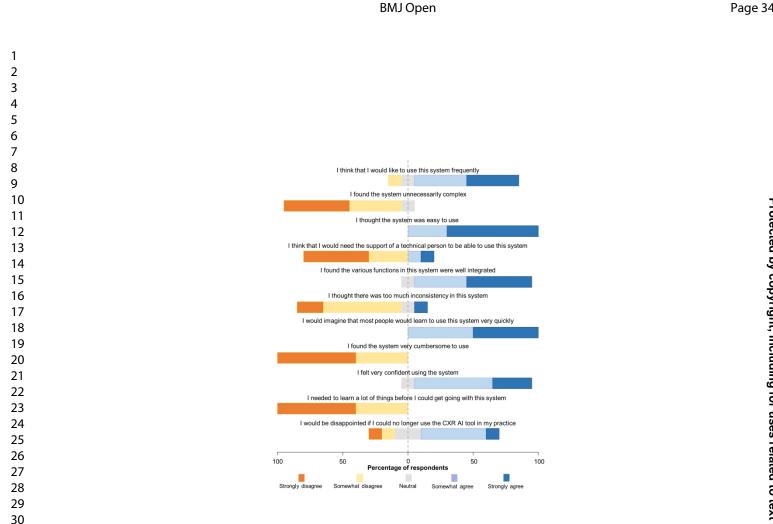


Figure 5 – Diverging stacked bar chart visualising the second set of survey responses of the radiologists.

645x484mm (118 x 118 DPI)

BMJ Open

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)

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

	Significantly worse	Moderately worse	Slightly worse	About the same	Slightly better	Moderately better	Significan better
How do you feel this tool impacted reporting time?	0	ο	0	ο	о	0	0
How do you feel this tool impacted reporting accuracy?	0	0	0	ο	ο	0	0
	Very dissatisfied	Dissatisfied	Somewhat dissatisfied	Neutral	Somewhat satisfied	Satisfied	Very dissatisfie
How satisfied were you with reporting time using this CXR tool?	0	0	0	0	0	0	0
How satisfied were you with the accuracy of this CXR tool's results?	ο	0	0	ο	ο	0	0
	Significantly more negative	Moderately more negative	Slightly more negative	About the same	Slightly more positive	Moderately more negative	Significan more negative
Did your attitude towards the CXR tool itself change from start to end of the trial?	0	0	0	ο	0	0	0
Has your attitude towards AI in general changed after using this CXR tool?	0	0	0	ο	0	0	0
	Ċ	Strongly disagree	Somewhat disagree	Neutral	Somewhat agree	Strongly agree	
I think that I would like to use this system frequently.		0	0	ο	0	0	
I found the system unnecessarily complex.		ο	ο	ο	ο	0	
I thought the system was easy to use.		0	ο	ο	ο	0	
I think that I would need the support of a technical person to be able to use this system.		ο	ο	٥	0	ο	
I found the various functions in this system were well integrated.		0	0	ο	ο	0	
I thought there was too much inconsistency in this system.		0	0	0	0	0	
I would imagine that most people would learn to use this system very quickly.		ο	ο	0	0	0	
I found the system very cumbersome to use.		ο	ο	ο	ο	ο	
I felt very confident using the system.		0	ο	ο	0	0	
I needed to learn a lot of things before I could get going with this system.		0	ο	ο	0	0	

I would be disappointed if I could no longer use the CXR AI tool in my practice.	0	0	0	0	0	
--	---	---	---	---	---	--

to occurrent of the terms of terms

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 32 24 25 26 27 28 29 30 31 32 33 34 35 36 37 37 37 37 37 37 37 37 37 37	
24 25 26 27 28 29 20	
30 31 32 33 34 35 36	
37 38 39 40 41 42 43	
44 45 46 47 48 49	
50 51 52 53 54 55 56 57 58 59	
60	

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: "Al- 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

40 BMJ Open: first published as 10.1136/bmjopen-2021-052902 on 20 December 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de I Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19	
20 21 22	
23 24 25 26	
20 27 28 29	
30 31 32	
33 34 35	
36 37 38	
39 40 41 42	
42 43 44 45	
46 47 48	
49 50 51	
52 53 54	
55 56 57	
58 59 60	

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")

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

1	
1	
2 3 4 5 6 7 8 9	
כ ⊿	
4 5	
5	
0	
/ 0	
0	
) 10	
11	
17	
12 13 14 15 16 17	
14	
15	
16	
17	
18	
18 19	
20	
21	
22	
23	
24	
25	
20 21 22 23 24 25 26 27 28 29	
27	
28	
29	
30	
31	
32	
32 33 34 35 36	
34	
35	
36	
37	
38	
39	
40	
41	
42	
43	
44	
45	
46	
47	
48	
49	
50	
51	
52	
53	
54	
55	

60

	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

secklist fo Artif Intell 20. **RSNA**[®]

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

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 <u>licence</u>.

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 <u>Creative Commons</u> 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.

terez oni

Enseignement Superieur (ABES) Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies



2 3 4	1	Assessment of the effect of a comprehensive chest
5 6 7	2	radiograph deep learning model on radiologist
, 8 9	3	reports and patient outcomes: a real-world
10 11	4	observational study
12 13	5	
14	6	
15	7	Catherine M Jones ^{1,2} , Luke Danaher ² , Michael R Milne ^{1,2*} , Cyril Tang ¹ , Jarrel Seah ^{1,3} , Luke
16		
17	8	Oakden-Rayner ⁴ , Andrew Johnson ¹ , Quinlan D Buchlak ^{1,5} , Nazanin Esmaili ^{5,6}
18 19	9	
20	10	¹ Annalise-AI, Sydney, NSW, Australia
20	11	² I-MED Radiology Network, Sydney, NSW, Australia
22	12	³ Department of Radiology, Alfred Health, Melbourne, VIC, Australia
23	13	⁴ Australian Institute for Machine Learning, The University of Adelaide, Adelaide, SA, Australia
24	14	⁵ School of Medicine, University of Notre Dame Australia, Sydney, NSW, Australia
25	15	⁶ Faculty of Engineering and IT, University of Technology Sydney, Ultimo, NSW, Australia
26	16	ruduký or Engineering und 11, om verský or reemiotogy bydney, oranie, 1600, 1600, 14000
27	17	
28	18	*Correspondence to: michael.milne@annalise.ai
29	19	
30		
31	20	
32	21	
33	22	
34	23	
35	24	Corresponding author:
36	25	Name: Michael Milne
37	26	Annalise-AI
38	27	Sydney, Australia
39 40	28	E-mail: michael.milne@annalise.ai
40 41	29	J
42	30	
43	31	
44		
45	32	
46	33	
47	34	
48	35	
49	36	
50	37	
51	38	
52	39	Keywords: Machine learning; chest X-ray, deep learning.
53	40	,
54	41	Word Count: 4,486
55	-	
56		
57 50		
58 59		
60		
00		

42	ABSTRACT
43 44	Objectives: AI algorithms have been developed to detect imaging features on chest X-ray (CXR)
45	with a comprehensive AI model capable of detecting 124 CXR findings being recently developed. The
46	aim of this study was to evaluate the real-world usefulness of the model as a diagnostic assistance device
47	for radiologists.
48	Design: This prospective real-world multicentre study involved a group of radiologists using the
49	model in their daily reporting workflow to report consecutive chest X-rays and recording their feedback
50	on level of agreement with the model findings and whether this significantly affected their reporting.
51	Setting: The study took place at radiology clinics and hospitals within a large radiology network
52	in Australia between November and December 2020.
53	Participants: Eleven consultant diagnostic radiologists of varying levels of experience
54	participated in this study.
55	Primary and secondary outcome measures: Proportion of CXR cases where use of the AI
56	model led to significant material changes to the radiologist report, to patient management, or to imaging
57	recommendations. Additionally, level of agreement between radiologists and the model findings, and
58	radiologist attitudes towards the model were assessed.
59	Results: Of 2,972 cases reviewed with the model, 92 cases (3.1%) had significant
60	report changes, 43 cases (1.4%) had changed patient management and 29 cases (1.0%) had further
61	imaging recommendations. In terms of agreement with the model, 2,572 cases showed complete
62	agreement (86.5%). 390 (13%) cases had one or more findings rejected by the radiologist. There
63	were 16 findings across 13 cases (0.5%) deemed to be missed by the model. Nine out of 10 radiologists
64	felt their accuracy was improved with the model and were more positive towards AI post-study.
65	Conclusions: Use of an AI model in a real-world reporting environment significantly improved
66	radiologist reporting and showed good agreement with radiologists, highlighting the potential for AI
67	diagnostic support to improve clinical practice.
68	
69	

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

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 • yn of ea... 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].

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

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

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

2	
3	
4	
5	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
16 17 18	
17	
18	
19	
20	
21	
22	
22	
25	
24	
25	
26	
27	
26 27 28	
29	
30	
30 31	
32	
33	
34 35	
35	
36	
37 38	
20	
39	
40	
41	
42	
43	
44	
45	
46	
47	
47 48	
49	
50	
51	
52	
53	
55 54	
55	
56	
57	
58	
59	
60	
00	

1

112 volume and availability of digital imaging datasets [11]. Of note is the rise of convolutional neural 113 networks (CNNs), a type of deep neural network that excels at image feature extraction and classification, 114 and demonstrates strong performance in medical image analysis, leading to the rapid advancement of 115 computer vision in medical imaging [12,13]. CNNs have been used to develop models to successfully 116 detect targeted clinical findings on CXR, including lung cancer [14,15], pneumonia [16,17], COVID-19 117 [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 118 119 interpretation, however most of these deep learning systems are limited in scope to a single finding or a 120 small set of findings, therefore lacking the broad utility that would make them useful in clinical practice. 121 122 Recently, our group developed a comprehensive deep learning CXR diagnostic assist device, 123 which was designed to assist clinicians in CXR interpretation and improve diagnostic accuracy, validated 124 for 124 clinically relevant findings seen on frontal and lateral chest radiographs [31]. The primary 125 objective of the current study was to evaluate the real-world usefulness of the model as a diagnostic assist 126 device for radiologists in both hospital and community clinic settings. This involved examining the 127 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

136 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.

142 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].

159Technical Integration160Prior to the start of the star

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

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

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.

168 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.

185 CXR Case Selection

In this multicentre real-world prospective study, all consecutive chest radiographs reported by the
 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

1 2		
3 4	190	reporting environment. As per usual workflow across a large radiology network spanning a
5 6	191	geographically large area with many regional and remote clinics, both on-site and remote reporting of
7 8	192	CXR cases was undertaken. A total of 106 sites contributed cases with case numbers varying from one
9 10	193	case up to a maximum of 271 cases at the busiest site.
11 12	194	
13 14	195	At least one frontal chest radiograph was required for analysis by the model, and cases that did
15 16	196	not include at least one were excluded. Chest radiographs from patients aged younger than 16 years were
17 18 19	197	excluded. Data from all sources was de-identified for analysis.
20 21	198	
22 23	199	AI-Assisted Reporting
24 25	200	For each CXR case, radiologists produced their clinical report with access to clinical information,
26 27	201	the referral and available patient history, in line with the normal workflow. The AI model analyses the
28 29	202	CXR image(s) for each case but does not incorporate clinical inputs (such as previous imaging, referral
30 31	203	information or patient demographic data) into the analysis. Model output was displayed to the radiologist
32 33	204	in a user interface, linked to the image in the PACS, automatically launching when a CXR case was
34 35 36	205	opened (Figure 1).
37 38	206	
39 40	207	A modified version of the commercially available AI software was employed for this study,
41 42	208	which incorporated changes into the user interface to allow radiologists to provide feedback on model
43 44	209	recommendations. No changes were made to the underlying model. An example of the modified model
45 46	210	user interface is presented in figure 2. For each case, the model provided a list of suggested findings,
47 48	211	listed as "priority" or "other", along with a confidence indicator. For a subset of findings, a region of
49 50	212	interest localiser was overlayed on the image and the model indicated whether the finding was on the left
51 52 53	213	or the right side, or both (see Supplementary Table 1). The CXR viewer was configured to display its
55 54 55	214	findings after the radiologists' initial read of the case. For each case, radiologists were asked to review the
56 57	215	CXR viewer's findings and provide feedback within the viewer. The options presented to the radiologists
58 59	216	in the viewer are listed in Table 1.
60	217	

Enseignement Superieur (ABES) Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

	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 reportA yes/no binary question relating to the effect of the mo the radiologist report				
	These findings may impact patient managementA yes/no binary question relating to the effect of the model our 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			
219 220					
221	The outcome measure of 'sign	nificant impact on the report' was the primary outcome measure.			
222	A significant change was described as the inclusion of findings recommended by the model, which				
223	altered the radiologists report in a meaningful way. As this varied by patient and clinical setting, it				
224	was left to the discretion of the radiolo	ogist. During the analysis of radiologist feedback, it was			
225	assumed that a change in patient mana	agement or further imaging recommendation would not occur			
226	without radiologists indicating a mater	rial change in the CXR report, and thus management and			
227	imaging questions were dependent on	a significant change in the report. This was also patient-			
228	specific; for example, missing a pneur	nothorax in a ventilated patient with known pneumothorax			
229	would not have the same impact on pa	tient management as a previously unknown pneumothorax in an			
230	outpatient. Free text input describing r	nissed findings or other relevant data were manually added after			
231	data collection was complete.				
232	No formal adjudication of cases showing discrepancy between radiologist and model				
233	interpretation was performed. The stud	dy was not designed as a diagnostic accuracy validation. No			
234	review or ground truthing process was performed. Radiologists remained responsible for image				
235	interpretation and formulation of the r	eport.			

1 2		
3 4	236	
5 6	237	Post-Study Survey
7 8	238	Upon completion of data collection, a post-study survey was distributed to all participating
9 10	239	radiologists to obtain feedback on the usefulness of the CXR viewer and how it affected their opinion of
11 12	240	AI in radiology. A table of the survey questions is presented in Supplementary Table 2.
13 14	241	
15 16 17 18	242	Statistics and Data Analysis
	243	A 1% rate of significant changes in reports (the primary outcome measure) was deemed to be
19 20 21	244	clinically significant prior to commencing the study. Based on estimations of the prevalence of missed
22 23	245	critical findings on CXR, preliminary power calculations estimated that the number of cases required to
24 25	246	detect at least a 1% rate of significant changes in reports was approximately 2000 cases in total, with
26 27	247	alpha value 0.05 and desired power of 0.90. To account for any dropout in radiologists or cases, a target
28 29	248	of 3000 cases was set for the study. Ten radiologists were recruited, with an eleventh included for any
30 31	249	unexpected participant drop out and to achieve this target in a reasonable time period.
32 33	250	
	251	A two-tailed binomial test was used to test the hypothesis that the rate of significant report
	252	change, patient management change, or imaging recommendation change was at least 1%. To ensure that
39 40	253	the sampling of CXRs reasonably approximated a random snapshot of the true population, radiologists in
41 42	254	various states, experience levels as well as different conditions of practice (community clinic vs hospital
43 44	255	based) were selected. Additionally, the study was conducted prospectively which further aligned the
45 46	256	structure of the sampled data with the expected structure of the population, justifying the choice of
47 48	257	analysing the sample using a binomial test without adjustment for each radiologist.
49 50	258	Multivariate logistic regression using generalised linear mixed effect analysis was used to assess
51 52	259	the effect of several possible confounders on the measured outcomes, including the number of critical
53 54	260	clinical findings per case identified by the model, the inpatient/outpatient status of the patients, the
55 56 57	261	experience level of the radiologists, and the presence or absence of a lateral radiograph. The Wald test
57 58 59 60	262	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.

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)
tal			2,972	92 (3.1)	43 (1.4)	29 (1.0)

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

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).

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

315 316 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

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

324	not to the report or patient management ($p = 0.105$ and $p = 0.061$, respectively).
325	
326	Radiologists with fewer than 5 years consultant experience contributed 1,347 cases, and indicated
327	a rate of 5.0% for significant report change, 2.4% patient management change, and 1.5%
328	recommendations for further imaging. These numbers were higher than for the radiologists with 6-10
329	years of experience (1.3%, 0.4%, 0.5% respectively over 748 cases) and also for radiologists with greater
330	than 10 years of experience (1.6%, 0.9%, 0.6% over 877 cases). However, a likelihood ratio test applied
331	to binomial logistic regression analysis indicated that the level of radiologist experience did not
332	significantly influence the rate of change in report, patient management, or imaging recommendation ($p =$
333	0.120, $p = 0.262$, and $p = 0.516$, respectively). Whether a patient was imaged as an inpatient or
334	outpatient was not significantly associated with any change in report, patient management, or imaging
335	recommendation ($p = 0.358$, $p = 0.572$, $p = 0.326$, respectively).

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

R. R. ONL

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

52
53 342 Survey Results
54

55343The post-study survey was completed by ten out of the eleven radiologists (Figure 4 and Figure56573445). Notably, 7 (70%) participants felt that their reporting time was slightly worse, however when asked5859345how satisfied they were with their reporting time, 7 (70%) indicated that they were satisfied.

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies

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.

tor occite iew only

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.

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

BMJ Open

1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
12	
13	
14	
15	
16	
1-	
17	
18	
19	
20	
7 1	
21 22	
22	
23	
25	
24	
25	
25	
26	
27	
28	
29	
~ ~	
30	
31	
32	
33	
34	
35	
36	
30	
37	
38	
39	
40	
41	
42	
43	
44	
45	
46	
47	
48	
49	
50	
51	
52	
54	
55	
00	
56	
20	

396

378

379 The significant impact of the CXR viewer on radiologist reporting and recommendations did 380 however come at the cost of false positives, with 13% of cases having one or more model findings 381 rejected by the radiologist. When this false positive rate is compared against the false positive rates per 382 case reported in other studies investigating CXR models, which range from 14 - 88% [14,49,50], it is 383 considered acceptable. Furthermore, these studies report false-positive rates for CXR models that only 384 detect lung nodules, while in the current study this represents the false positive rate across 124 findings. 385 Notably, on a per finding basis, only 5.3% of critical findings detected by the model were rejected by the 386 radiologist. However, there were several outliers in the critical findings group that had noticeably higher 387 rates of rejection, including acute rib fracture, hilar lymphadenopathy, malpositioned NGT/PAC, shoulder 388 dislocation and tension pneumothorax. Several explanations for this are low sample size, the subjectivity 389 of diagnosis (especially for hilar lymphadenopathy and tension features of pneumothorax), and 390 heightened model sensitivity at the expense of specificity. In particular, the rate of 'overcalling' of 391 malposition of nasogastric tubes was related to both the threshold choice (favouring sensitivity given the 392 critical nature of NGT malposition) and the limitation in the model output in distinguishing malpositioned 393 NGTs from incompletely visualised NGTs. This limitation has subsequently been addressed with model 394 modifications. Overall, this trade-off appears to be reasonable to the participating radiologists, who 395 reported a high level of satisfaction with the model.

397 In this study, analysis of radiologists by experience level using logistic regression found no 398 statistically significant relationship between experience level and increased changes to reports, patient 399 management changes, or imaging recommendations as a result of the model. Statistical analysis of the 400 relationship between experience level and change in report was associated with a p value of 0.12, 401 suggesting that, with further research, a significant relationship may be identified. It is expected that the 402 inclusion of a larger group of radiologists may lead to a significant finding, as the association between 403 experience and level of change has been noted in other studies. For example Jang et al., showed that less 57 58 404 experienced radiologists benefited the most from the diagnostic assistance in detecting lung nodules on 59 60 405 CXR [14]. In this study, three of the 11 radiologists contributed a higher than average incidence of the

1 2

BMJ Open

3	
4	
5	
6	
7	
8	
9	
9 10	
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	
∠∪ ว1	
20 21 22 23 24 25 26 27 28 29 30 31	
22	
23	
24	
25	
26	
27	
28	
29	
20	
50 21	
31	
32	
33	
34	
35	
36	
37	
38	
39	
40	
40 41	
42	
43	
44	
45	
46	
47	
48	
49	
50	
51	
52	
53	
54	
55	
56	
57	
58	
59	
60	

406 primary outcome of report change, and these were all less experienced radiologists compared to the 407 cohort average experience level. Whilst this may be due to variations in individual radiologist 408 interpretation of 'significant report change', the consistency of experience level across these three 409 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.

418 419 The post-study survey provided further insight into the impact that the CXR viewer had on 420 participant reporting, in addition to the level of agreement and changes to the radiology report and patient 421 management recommendations outlined above. The first notable response was that the CXR viewer may 422 have negatively affected reporting times (albeit only mildly) for the majority of radiologists. This 423 outcome was expected in this study setting because the radiologists were taking additional time to provide 424 feedback on the model's recommendations for each case. Previous studies that surveyed radiologists 425 reported that 74.4% thought AI would lower the interpretation time [51]. It is notable that even with the 426 negative impact the model had on reporting time, the majority of radiologists (70%) were still satisfied 427 with reporting time while using the CXR viewer, suggesting that the diagnostic improvements offered by 428 the model were enough to offset the additional perceived reporting time. Additional insight from the 429 survey suggested that very little training was required before radiologists felt comfortable using the tool. 430 This is useful as education on AI has been a primary concern amongst clinicians, as a large proportion of 431 radiologists report having little knowledge of AI [52]. 432

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

433 Limitations and future research

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, AI training, and similar technologies.

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 ies 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.

460 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.

1 2		
3 4	486	
5 6	487	PATIENT AND PUBLIC INVOLVEMENT
7 8 9	488	Patients and public were not involved in the design, conduct, or reporting of this study.
10 11	489	
12 13	490	DATA AVAILABILITY STATEMENT
14 15 16	491	All data relevant to the study are included in the article or uploaded as online supplemental
17 18	492	information. No additional data are available.
19 20 22 23 24 25 26 27 28 29 30 32 33 35 36 37 39 40 41 42 43 445 46 78 90 152 54 55 57 89 60	493	information. No additional data are available.

1 2				
3	494	494 References		
4 5	495	1	Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are	
6 7	496	1	Data. <i>Radiology</i> 2016; 278 :563–77. doi:10.1148/radiol.2015151169	
8 9	497	2	Greene R. Francis H. Williams, MD: father of chest radiology in North America.	
9 10	498		RadioGraphics 1991;11:325-32. doi:10.1148/radiographics.11.2.2028067	
11	499	3	Schaefer-Prokop C, Neitzel U, Venema HW, et al. Digital chest radiography: an update	
12 13	500	-	on modern technology, dose containment and control of image quality. <i>Eur Radiol</i>	
14 15	501		2008; 18 :1818–30. doi:10.1007/s00330-008-0948-3	
16	502	4	Lee CS, Nagy PG, Weaver SJ, et al. Cognitive and System Factors Contributing to	
17 18	503		Diagnostic Errors in Radiology. American Journal of Roentgenology 2013;201:611–7.	
19	504		doi:10.2214/AJR.12.10375	
20 21	505	5	Chotas HG, Ravin CE. Chest radiography: estimated lung volume and projected area	
22	506		obscured by the heart, mediastinum, and diaphragm. Radiology 1994;193:403-4.	
23	507		doi:10.1148/radiology.193.2.7972752	
24 25	508	6	Berlin L. Accuracy of Diagnostic Procedures: Has It Improved Over the Past Five	
26	509		Decades? American Journal of Roentgenology 2007;188:1173-8.	
27 28	510		doi:10.2214/AJR.06.1270	
29	511	7	Zaorsky NG, Churilla TM, Egleston BL, et al. Causes of death among cancer patients.	
30 31	512		Annals of Oncology 2017;28:400-7. doi:10.1093/annonc/mdw604	
32	512	0	del Cielle A. Energhi D. Cautering and A. Miner diana segueration relation and	
33 34	513 514	8	del Ciello A, Franchi P, Contegiacomo A, <i>et al.</i> Missed lung cancer: when, where, and why? <i>Diagn Interv Radiol</i> 2017; 23 :118–26. doi:10.5152/dir.2016.16187	
35	517		wily ! Diagn Interv Radioi 2017,23.116 20. doi:10.5152/dii.2010.10107	
36	515	9	Fazal MI, Patel ME, Tye J, et al. The past, present and future role of artificial intelligence	
37 38	516		in imaging. European Journal of Radiology 2018; 105 :246–50.	
39	517		doi:10.1016/j.ejrad.2018.06.020	
40 41	518	10	Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science	
42	519		2015; 349 :255–60. doi:10.1126/science.aaa8415	
43 44	520	11	Hosny A, Parmar C, Quackenbush J, et al. Artificial intelligence in radiology. Nat Rev	
45	521		Cancer 2018;18:500-10. doi:10.1038/s41568-018-0016-5	
46 47	522	12	Erickson BJ, Korfiatis P, Akkus Z, et al. Machine Learning for Medical Imaging.	
48	523	12	RadioGraphics 2017; 37 :505–15. doi:10.1148/rg.2017160130	
49 50				
51	524	13	Esteva A, Chou K, Yeung S, <i>et al.</i> Deep learning-enabled medical computer vision. <i>npj</i>	
52	525		<i>Digital Medicine</i> 2021;4:1–9. doi:10.1038/s41746-020-00376-2	
53 54	526	14	Jang S, Song H, Shin YJ, et al. Deep Learning-based Automatic Detection Algorithm for	
55	527		Reducing Overlooked Lung Cancers on Chest Radiographs. <i>Radiology</i>	
56 57	528		2020; 296 :652–61. doi:10.1148/radiol.2020200165	
58				
59 60				
00				

2			
3 4 5 6	529 530 531	15	Liang C-H, Liu Y-C, Wu M-T, <i>et al.</i> Identifying pulmonary nodules or masses on chest radiography using deep learning: external validation and strategies to improve clinical practice. <i>Clinical Radiology</i> 2020; 75 :38–45. doi:10.1016/j.crad.2019.08.005
7 8 9 10	532 533	16	Hurt B, Kligerman S, Hsiao A. Deep Learning Localization of Pneumonia: 2019 Coronavirus (COVID-19) Outbreak. <i>J Thorac Imaging</i> 2020; 35 :W87–9.
11 12 13 14	534 535 536	17	Kim JY, Choe PG, Oh Y, <i>et al.</i> The First Case of 2019 Novel Coronavirus Pneumonia Imported into Korea from Wuhan, China: Implication for Infection Prevention and Control Measures. <i>J Korean Med Sci</i> 2020; 35 . doi:10.3346/jkms.2020.35.e61
15 16 17 18 19	537 538 539	18	Bassi PRAS, Attux R. A Deep Convolutional Neural Network for COVID-19 Detection Using Chest X-Rays. <i>arXiv:200501578 [cs, eess]</i> Published Online First: 12 January 2021.http://arxiv.org/abs/2005.01578 (accessed 23 Mar 2021).
20 21 22 23 24	540 541 542 543	19	Rueckel J, Trappmann L, Schachtner B, <i>et al.</i> Impact of Confounding Thoracic Tubes and Pleural Dehiscence Extent on Artificial Intelligence Pneumothorax Detection in Chest Radiographs. <i>Investigative Radiology</i> 2020; 55 :792–8. doi:10.1097/RLI.000000000000707
25 26 27 28 29 30	544 545 546 547	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. <i>Image Analysis and Recognition</i> . Cham: : Springer International Publishing 2019. 325–32. doi:10.1007/978-3-030-27272-2_28
31 32 33 34	548 549 550	21	Hwang EJ, Hong JH, Lee KH, <i>et al.</i> Deep learning algorithm for surveillance of pneumothorax after lung biopsy: a multicenter diagnostic cohort study. <i>Eur Radiol</i> 2020; 30 :3660–71. doi:10.1007/s00330-020-06771-3
35 36 37 38 39	551 552 553	22	Park S, Lee SM, Kim N, <i>et al.</i> Application of deep learning–based computer-aided detection system: detecting pneumothorax on chest radiograph after biopsy. <i>Eur Radiol</i> 2019; 29 :5341–8. doi:10.1007/s00330-019-06130-x
40 41 42 43	554 555 556	23	Wang X, Yu J, Zhu Q, <i>et al.</i> Potential of deep learning in assessing pneumoconiosis depicted on digital chest radiography. <i>Occup Environ Med</i> 2020;77:597–602. doi:10.1136/oemed-2019-106386
44 45 46 47	557 558	24	S Z, X Z, R Z. Identifying Cardiomegaly in ChestX-ray8 Using Transfer Learning. <i>Stud Health Technol Inform</i> 2019; 264 :482–6. doi:10.3233/shti190268
48 49 50 51	559 560 561	25	Zou X-L, Ren Y, Feng D-Y, <i>et al.</i> A promising approach for screening pulmonary hypertension based on frontal chest radiographs using deep learning: A retrospective study. <i>PLOS ONE</i> 2020; 15 :e0236378. doi:10.1371/journal.pone.0236378
52 53 54 55	562 563 564	26	Pasa F, Golkov V, Pfeiffer F, <i>et al.</i> Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. <i>Scientific Reports</i> 2019; 9 :6268. doi:10.1038/s41598-019-42557-4
56 57 58 59 60	565 566 567	27	Nash M, Kadavigere R, Andrade J, <i>et al.</i> Deep learning, computer-aided radiography reading for tuberculosis: a diagnostic accuracy study from a tertiary hospital in India. <i>Scientific Reports</i> 2020; 10 :210. doi:10.1038/s41598-019-56589-3

1 2			
2 3 4 5 6 7 8	568 569 570 571	28	Heo S-J, Kim Y, Yun S, <i>et al.</i> Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers' Health Examination Data. <i>International Journal of Environmental Research and Public Health</i> 2019; 16 :250. doi:10.3390/ijerph16020250
8 9 10 11 12 13	572 573 574 575	29	Qin ZZ, Sander MS, Rai B, <i>et al.</i> Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. <i>Scientific Reports</i> 2019; 9 :15000. doi:10.1038/s41598-019-51503-3
14 15 16 17 18	576 577 578	30	Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. <i>Radiology</i> 2017; 284 :574–82. doi:10.1148/radiol.2017162326
19 20 21 22 23	579 580 581 582	31	Seah JCY, Tang CHM, Buchlak QD, <i>et al.</i> Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. <i>The Lancet Digital Health</i> 2021; 3 :e496–506. doi:10.1016/S2589-7500(21)00106-0
24 25 26 27	583 584	32	Annalise.ai - Annalise CXR comprehensive medical imaging AI. Annalise.ai. https://annalise.ai/products/annalise-cxr/ (accessed 23 Mar 2021).
27 28 29 30 31	586 Networks. <i>arXiv:190511946 [cs, stat]</i> Pu		Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. <i>arXiv:190511946 [cs, stat]</i> Published Online First: 11 September 2020.http://arxiv.org/abs/1905.11946 (accessed 30 Mar 2021).
32 33 34 35	588 589 590	34	Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. <i>arXiv:150504597 [cs]</i> Published Online First: 18 May 2015.http://arxiv.org/abs/1505.04597 (accessed 30 Mar 2021).
36 37 38 39 40 41	591 592 593 594	35	xmlmillr6.pdf. https://www.ebs.tga.gov.au/servlet/xmlmillr6?dbid=ebs/PublicHTML/pdfStore.nsf&doci d=F7ADAEBB76CEDD47CA2585E500424A43&agid=(PrintDetailsPublic)&actionid=1 (accessed 25 Aug 2021).
42 43 44 45	595 596 597	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).
46 47 48 49 50	 598 37 Benjamini Y, Hochberg Y. Controlling 599 Powerful Approach to Multiple Testing 600 (<i>Methodological</i>) 1995;57:289–300. 	Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. <i>Journal of the Royal Statistical Society Series B (Methodological)</i> 1995; 57 :289–300.	
51 52 53	601 602	38	Mckinney W. pandas: a Foundational Python Library for Data Analysis and Statistics. <i>Python High Performance Science Computer</i> 2011.
54 55 56	603 604	39	Harris CR, Millman KJ, van der Walt SJ, <i>et al.</i> Array programming with NumPy. <i>Nature</i> 2020; 585 :357–62. doi:10.1038/s41586-020-2649-2
57 58 59 60	605	40	Jones E, Oliphant T, Peterson P. SciPy: Open Source Scientific Tools for Python. 2001.

Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

1 2			
3 4 5 6	606 607 608	41	Pedregosa F, Varoquaux G, Gramfort A, <i>et al.</i> Scikit-learn: Machine Learning in Python. <i>Journal of Machine Learning Research</i> Published Online First: 12 October 2011.https://hal.inria.fr/hal-00650905 (accessed 23 Mar 2021).
7 8 9 10	609 610	42	Jolly E. Pymer4: Connecting R and Python for linear mixed modeling. <i>Journal of Open Source Software</i> 2018; 3 :862.
10 11 12 13	611 612	43	Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. Austin, Texas: 2010. 92–6. doi:10.25080/Majora-92bf1922-011
14 15 16 17 18	613 614 615	44	Seah J, Tang C, Buchlak QD, <i>et al.</i> Radiologist chest X-ray diagnostic accuracy performance improvements when augmented by a comprehensive deep learning model. <i>The Lancet Digital Health</i> 2021.
19 20 21 22 23	616 617 618 619	45	Hwang EJ, Park S, Jin K-N, <i>et al.</i> Development and Validation of a Deep Learning- Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. <i>JAMA Netw Open</i> 2019; 2 :e191095. doi:10.1001/jamanetworkopen.2019.1095
24 25 26 27 28	620 621 622	46	Hwang EJ, Nam JG, Lim WH, <i>et al.</i> Deep Learning for Chest Radiograph Diagnosis in the Emergency Department. <i>Radiology</i> 2019; 293 :573–80. doi:10.1148/radiol.2019191225
29 30 31 32	623 624 625	47	Singh R, Kalra MK, Nitiwarangkul C, <i>et al.</i> Deep learning in chest radiography: Detection of findings and presence of change. <i>PLOS ONE</i> 2018; 13 :e0204155. doi:10.1371/journal.pone.0204155
33 34 35 36 37 38	626 627 628 629	48	Khan FA, Majidulla A, Tavaziva G, <i>et al.</i> 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. <i>The Lancet Digital Health</i> 2020; 2 :e573–81. doi:10.1016/S2589-7500(20)30221-1
39 40 41 42	630 631 632	49	Dellios N, Teichgraeber U, Chelaru R, <i>et al.</i> Computer-aided Detection Fidelity of Pulmonary Nodules in Chest Radiograph. <i>J Clin Imaging Sci</i> 2017;7. doi:10.4103/jcis.JCIS_75_16
43 44 45 46 47	633 634 635	50	Sim Y, Chung MJ, Kotter E, <i>et al.</i> Deep Convolutional Neural Network–based Software Improves Radiologist Detection of Malignant Lung Nodules on Chest Radiographs. <i>Radiology</i> Published Online First: 12 November 2019. doi:10.1148/radiol.2019182465
48 49 50 51	636 637 638	51	Waymel Q, Badr S, Demondion X, <i>et al.</i> Impact of the rise of artificial intelligence in radiology: What do radiologists think? <i>Diagnostic and Interventional Imaging</i> 2019; 100 :327–36. doi:10.1016/j.diii.2019.03.015
52 53 54 55 56	639 640 641	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. <i>Journal of the American College of Radiology</i> 2018; 15 :1753–7. doi:10.1016/j.jacr.2017.12.021
57 58 59 60	642 643		

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



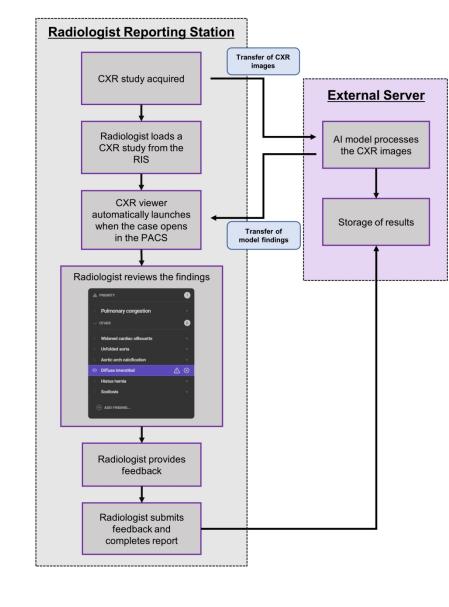


Figure 1 - Flow diagram illustrating the AI-assisted reporting process described in this study. (RIS: Radiological information system)

190x240mm (300 x 300 DPI)

$\begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \\ 12 \\ 13 \\ 14 \\ 15 \\ 16 \\ 17 \\ 18 \\ 19 \\ 20 \\ 21 \\ 22 \\ 23 \\ 24 \\ 25 \\ 26 \\ 27 \\ 28 \\ 29 \\ 30 \\ 31 \\ 32 \\ 33 \\ 34 \\ 35 \\ 36 \\ 37 \\ 38 \\ 39 \\ 40 \\ 41 \\ 42 \\ 43 \\ 44 \\ 5 \\ 46 \\ 47 \\ 48 \\ 49 \\ 50 \\ 51 \\ 52 \\ 53 \\ 54 \end{matrix}$	Figure 2 -
54 55	

60

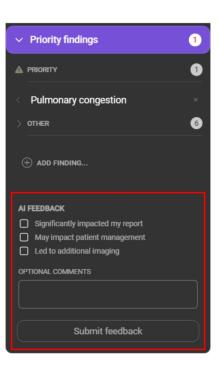
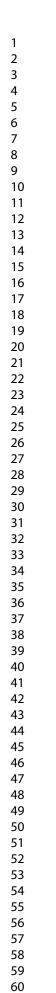


Figure 2 - Example of the modified user interface used by the participating radiologists in this study. The red box highlights the feedback options added to the interface for this study.

254x190mm (300 x 300 DPI)

BMJ Open: first published as 10.1136/bmjopen-2021-052902 on 20 December 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de l Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.



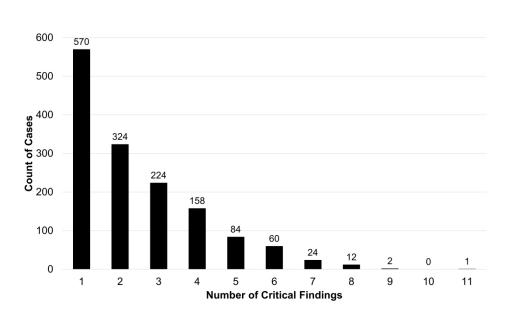
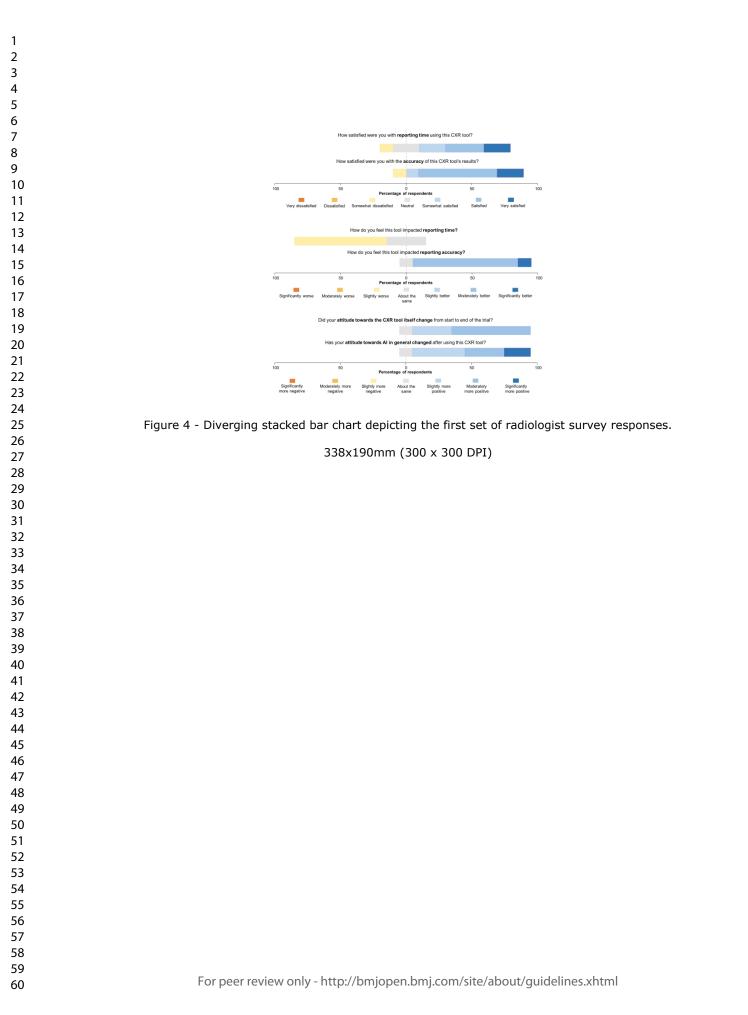
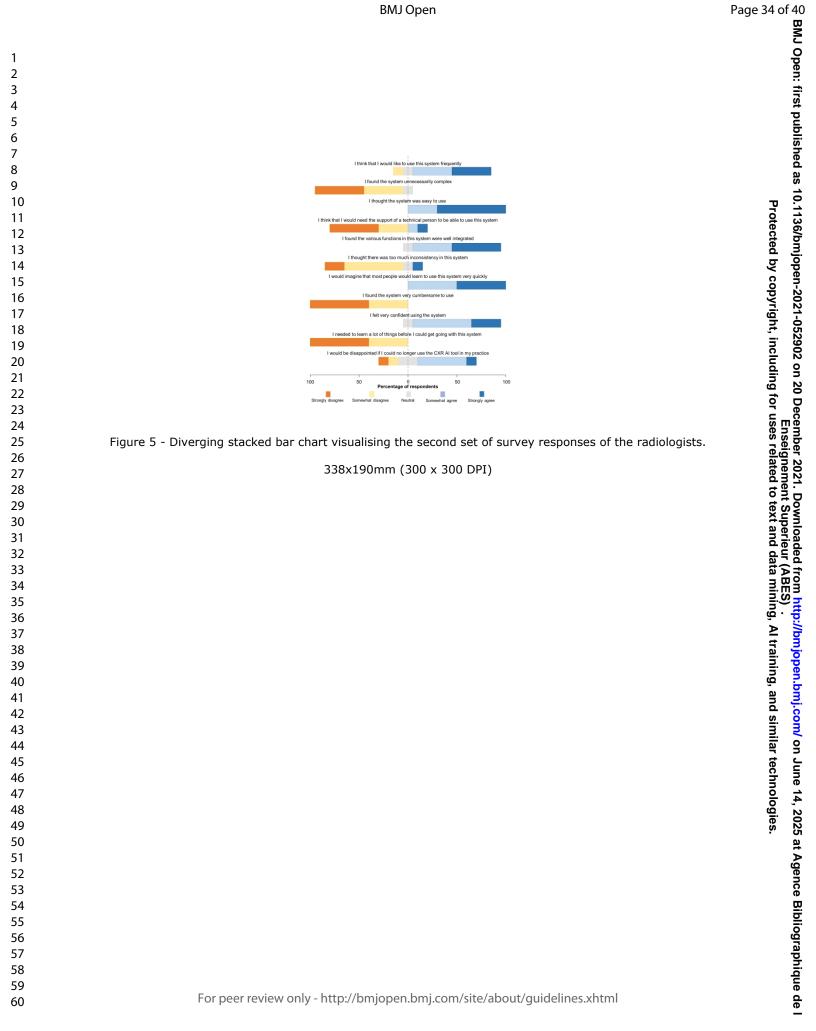


Figure 3 - Counts of numbers of critical findings for the cases seen by the radiologist, defined as the number of critical findings agreed + the number of critical findings added. The number of cases which returned zero findings was 1,513.

338x190mm (300 x 300 DPI)





BMJ Open

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)

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

	Significantly worse	Moderately worse	Slightly worse	About the same	Slightly better	Moderately better	Significan better
How do you feel this tool impacted reporting time?	0	ο	0	ο	о	0	0
How do you feel this tool impacted reporting accuracy?	0	0	0	ο	ο	0	0
	Very dissatisfied	Dissatisfied	Somewhat dissatisfied	Neutral	Somewhat satisfied	Satisfied	Very dissatisfie
How satisfied were you with reporting time using this CXR tool?	0	0	0	0	0	0	0
How satisfied were you with the accuracy of this CXR tool's results?	0	0	0	ο	ο	0	0
	Significantly more negative	Moderately more negative	Slightly more negative	About the same	Slightly more positive	Moderately more negative	Significan more negative
Did your attitude towards the CXR tool itself change from start to end of the trial?	0	0	0	ο	0	0	0
Has your attitude towards AI in general changed after using this CXR tool?	0	0	ο	ο	0	0	0
	Ċ	Strongly disagree	Somewhat disagree	Neutral	Somewhat agree	Strongly agree	
I think that I would like to use this system frequently.		0	0	ο	0	0	
I found the system unnecessarily complex.		ο	ο	ο	ο	0	
I thought the system was easy to use.		0	ο	ο	ο	0	
I think that I would need the support of a technical person to be able to use this system.		ο	ο	٥	0	ο	
I found the various functions in this system were well integrated.		0	0	ο	ο	0	
I thought there was too much inconsistency in this system.		0	0	0	0	0	
I would imagine that most people would learn to use this system very quickly.		ο	ο	0	0	0	
I found the system very cumbersome to use.		ο	ο	ο	ο	ο	
I felt very confident using the system.		0	ο	ο	0	0	
I needed to learn a lot of things before I could get going with this system.		0	ο	ο	0	0	

I would be disappointed if I could no longer use the CXR AI tool in my practice.	0	0	0	0	0	
--	---	---	---	---	---	--

to occurrent of the terms of terms



Section / Topic	No.	ltem	
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: "Al- 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

40 BMJ Open: first published as 10.1136/bmjopen-2021-052902 on 20 December 2021. Downloaded from http://bmjopen.bmj.com/ on June 14, 2025 at Agence Bibliographique de I Enseignement Superieur (ABES) . Protected by copyright, including for uses related to text and data mining, Al training, and similar technologies.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19	
20 21 22	
23 24 25 26	
20 27 28 29	
30 31 32	
33 34 35	
36 37 38	
39 40 41 42	
42 43 44 45	
46 47 48	
49 50 51	
52 53 54	
55 56 57	
58 59 60	

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")

For peer review only - http://bmjopen.bmj.com/site/about/guidelines.xhtml

1	
1 2	
2 3 4 5 6 7 8 9	
כ ⊿	
4	
5	
0	
/ Q	
0	
) 10	
11	
17	
12 13 14 15 16 17	
14	
15	
16	
17	
18	
18 19	
20	
21	
22	
23	
24	
25	
20 21 22 23 24 25 26 27 28 29	
27	
28	
29	
30	
31	
32	
32 33 34 35 36	
34	
35	
36	
37	
38	
39	
40	
41	
42	
43	
44	
45	
46	
47	
48	
49	
50	
51	
52	
53	
54	
55	

60

	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

secklist fo Artif Intell 20. **RSNA**[®]