








OPEN ACCESS

Evaluating the impact of AI assistance on decision-making in emergency doctors interpreting chest X-rays: a multi-reader multi-case study

David Lyell ¹, Michael Dinh,^{2,3} Mark Gillett,⁴ Nidhi Abraham,^{2,3} Emily Rose Symes,^{3,5} Anindya Pradipta Susanto ^{1,6}, Bashir Antoine Chakar ³, Radhika V Seimon ³, Enrico Coiera,¹ Farah Magrabi ¹

Handling editor Alex Novak

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/emered-2024-214781>).

¹Australian Institute of Health Innovation, Macquarie University, North Ryde, New South Wales, Australia

²Emergency Department, Royal Prince Alfred Hospital, Sydney Local Health District, Sydney, New South Wales, Australia

³RPA Green Light Institute for Emergency Care, Sydney Local Health District, Sydney, New South Wales, Australia

⁴Emergency Department, Royal North Shore Hospital, Northern Sydney Local Health District, Sydney, New South Wales, Australia

⁵Emergency Department, Canterbury Hospital, Sydney Local Health District, Sydney, New South Wales, Australia

⁶Medical Technology Cluster, Indonesian Medical Education and Research Institute, Universitas Indonesia, Jakarta, Indonesia

Correspondence to

Dr David Lyell;
david.lyell@mq.edu.au

Received 12 December 2024
Accepted 16 September 2025
Published Online First
13 October 2025



Check for updates

© Author(s) (or their employer(s)) 2025. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ Group.

To cite: Lyell D, Dinh M, Gillett M, et al. *Emerg Med J* 2025;**42**:774–782.

ABSTRACT

Background Artificial intelligence (AI) tools could assist emergency doctors interpreting chest X-rays to inform urgent care. However, the impact of AI assistance on clinical decision-making, a precursor to enhanced care and patient outcomes, remains understudied. This study evaluates the effect of AI assistance on clinical decisions of emergency doctors interpreting chest X-rays.

Method Junior and senior residents, emergency registrars and consultants working in Australian emergency departments were eligible. Doctors completed 18 clinical vignettes involving chest X-ray interpretation, representative of typical patient presentations. Vignettes were randomly selected from a bank of 49 based on the emergency medicine curriculum and contained a chest X-ray, presenting complaint, relevant symptoms and observations. Of the 18 vignettes, each doctor was randomly assigned to have half assisted by a commercial AI tool capable of detecting 124 different chest X-ray findings. Four vignettes contained X-rays known to produce incorrect AI findings. Primary outcomes were correct diagnosis and patient management. X-ray interpretation time, confidence of diagnosis, perceptions about the AI tool and the differential impact of AI assistance by seniority were also examined.

Results 200 doctors participated. AI assistance increased correct diagnosis by 5.9% (95% CI 2.7 to 9.2%) compared with unassisted vignettes, with the largest increase among senior residents (11.8%; 95% CI 5.2% to 18.3%). Patient management increased by 3.2% (95% CI 0.1% to 6.4%). Confidence in diagnosis increased by 5% (95% CI 3.4% to 6.6%; $p<0.001$) and interpretation time increased by 4.9 s ($p=0.08$). Incorrect AI findings decreased correct diagnosis by 1% for false-positive ($p=0.9$) and 9% for false-negative findings ($p=0.1$). Participants found the AI tool helpful for interpreting chest X-rays, highlighting missed findings, but were neutral on its accuracy.

Conclusion Improvements in diagnosis and patient management without meaningful increases in interpretation time suggest AI assistance could benefit clinical decisions involving chest X-ray interpretation. Further studies are required to ascertain if such improvements translate to improved patient care.

INTRODUCTION

Artificial intelligence (AI) technologies have the potential to enhance care delivery in emergency

WHAT IS ALREADY KNOWN ON THIS TOPIC

- ⇒ Use of artificial intelligence (AI) tools has been shown to improve the detection of common chest X-ray pathologies by emergency doctors.
- ⇒ However, the effects of AI assistance on diagnostic and patient management decisions remain understudied yet are critical to improving care delivery and patient outcomes.

WHAT THIS STUDY ADDS

- ⇒ We evaluated the impact of AI assistance on decision-making in emergency doctors interpreting chest X-rays. Diagnoses and patient management were compared, with and without AI, using clinical vignettes representative of emergency department patient presentations.
- ⇒ Accuracy of clinical decisions was greater among vignettes that were completed with AI assistance.
- ⇒ False-positive and false-negative AI findings did not significantly affect diagnosis or patient management.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

- ⇒ This study provides evidence that use of AI tools can enhance decision-making in emergency doctors interpreting chest X-rays, warranting further research in real-world clinical settings to evaluate their routine use as well as their impact on care delivery and patient outcomes.

departments (EDs) which are under increasing pressure worldwide, due to rising demand and overcrowding.^{1–3} Chest X-rays are one of the most frequently performed examinations and are often interpreted by emergency doctors to inform urgent clinical decisions in advance of formal radiology reporting. Accurate interpretation of chest X-rays is crucial for timely diagnosis and management of patients with critical conditions such as pneumothorax, pneumonia, heart failure and pleural effusion. One promising approach to support ED doctors, especially less experienced residents, is to provide access to AI tools that can assist with chest X-ray interpretation.

The utility of AI tools for chest X-ray interpretation in ED remains largely unexplored.^{4–6} Recent



studies have shown that AI can improve detection of the most common chest X-ray pathologies such as pleural effusions, pneumothoraces, consolidations suspicious for pneumonia and nodules, aiding doctors in their interpretation.^{7,8} However, few studies have evaluated the critical link between improved detection of X-ray findings and clinical decision-making, which is an essential precursor for improving care delivery and patient outcomes.⁹

Accordingly, we evaluated the effects of AI assistance on the diagnostic and patient management decisions of emergency doctors interpreting chest X-rays. A multi-reader multi-case study was conducted using clinical vignettes representative of patient presentations to the ED. Clinical vignettes allow evaluation of the effects of AI on decision-making in a patient-free and risk-free environment, providing greater experimental control than real-world clinical settings. They also enable testing of critical patient safety factors, such as false-positive and false-negative AI findings that cannot be ethically or feasibly tested in clinical settings. Crucially, vignettes enable the impact of AI on decision-making to be established ahead of expensive and potentially disruptive clinical deployment.

We also examined whether AI assistance had a differential impact on outcomes based on seniority of doctors and related outcomes such as interpretation time, confidence in diagnosis and participant perceptions about the AI. The risk of AI assistance adversely influencing clinical decisions was tested by comparing the effect of known false-positive and false-negative AI findings on diagnostic and patient management decisions; this aspect has not been previously studied.

METHODS

Participants

Doctors working in EDs were recruited. Junior medical officers, senior resident medical officers, emergency registrars and emergency consultants working in Australian EDs were eligible (see online supplemental appendix A for detailed criteria). Participants were recruited by publishing a call for volunteers via email at five metropolitan teaching hospitals in Sydney, Australia, and across state-based and national emergency medicine networks.

We estimated that 199 participants were required to detect a difference with a small effect size (Cohen's $d=0.2$)¹⁰ between the AI-assisted and unassisted vignettes, using paired two-tailed t-tests with 80% power and alpha of 0.05. A small effect size was selected based on a prior study of the detection of X-ray findings using the AI tool evaluated here.¹¹

Study design

A split-plot multi-reader multi-case study was conducted online.¹² Participants completed 18 clinical vignettes, half with AI assistance and half unassisted (figure 1). Fourteen vignettes were representative of emergency presentations (seven with and seven without AI assistance). The remaining four vignettes involved X-rays where the AI tool was known to produce incorrect findings relevant to the correct diagnosis for the vignette, two involved X-rays known to produce false-positive findings and two known to produce false-negative findings. The effects of AI assistance on decision-making were evaluated within-participants (repeated measures) as the difference between AI-assisted and unassisted vignettes.

Randomisation and masking

The allocation of vignettes to AI assistance was according to 16 pre-generated sequences, created so that each vignette was

allocated approximately evenly to being completed with and without AI assistance. Participants were allocated to sequences at the time of enrolment using balanced randomisation, with sequences presented in random order. Vignettes presenting incorrect AI findings were always presented last, in random order, to avoid incorrect AI findings inducing a first failure effect, whereby on recognising an initial failure, participant trust in and reliance on AI is immediately and substantially degraded and could adversely influence results given the rare, but safety-critical nature of the errors tested.¹³

Vignettes with AI assistance were readily apparent to participants; however, they were blinded to the presence of incorrect AI findings. Scorers assessing participant responses to the vignettes against the validated gold standard were blinded to whether AI assistance had been provided.

Clinical vignettes

The vignettes completed by participants were randomly selected from a bank of 49 based on the Australasian College for Emergency Medicine (ACEM) curriculum, covering typical and important conditions for which chest X-rays would be indicated in emergency medicine.¹⁴ Vignettes included cardiovascular presentations, such as congestive cardiac failure, disorders of pericardium and aortic dissection, as well as respiratory presentations like pneumonia, aspiration, pneumothorax, pneumomediastinum, pleural effusions, lung lesion/mass and chronic obstructive pulmonary disease. Additional areas included oncological and immunological presentations (eg, sarcoidosis), trauma and orthopaedic cases involving fractures and dislocations, and procedural indications such as advanced airway management, tube thoracostomy, nasogastric tube and central venous access. Vignettes with normal X-rays include several gastrointestinal presentations, pulmonary embolism and viral infections. This comprehensive coverage ensured the vignettes reflected the breadth of conditions encountered in the ED.

Vignettes included the presenting complaint, relevant symptoms and observations introducing participants to the hypothetical patient and providing information ordinarily obtained from the patient or their caregiver as well as preliminary observations, establishing why a chest X-ray was indicated (see online supplemental appendix B for an overview of vignettes and online supplemental appendix C for an example). The X-rays were sourced from publicly available datasets (see online supplemental appendix D).

The vignettes, X-rays, AI findings and gold standard responses for diagnosis and patient management as relevant to emergency medicine were validated by an independent expert panel of two senior consultant Fellows of the Australasian College for Emergency Medicine (FACEM) and one Fellow of the Royal Australian and New Zealand College of Radiologists (RANZCR). Disagreements were resolved by consensus.

The panel rated the difficulty of correctly diagnosing the patient based on the vignette and X-ray with typical emergency presentations classified as simple and those assessed as being slightly, moderately or substantially harder than the usual being classified as complex. The panel also rated the acuity of vignettes, which were primarily informed by patient observations. X-rays known to produce false-positive and false-negative AI findings were selected from examples provided by the developer of the AI tool and those discovered while curating vignettes, all were subsequently validated by the expert panel radiologist.

Chest X-ray AI

The AI tool evaluated in this study, Annalise Enterprise CXR (AE V2.2) is a commercially available tool that is intended to

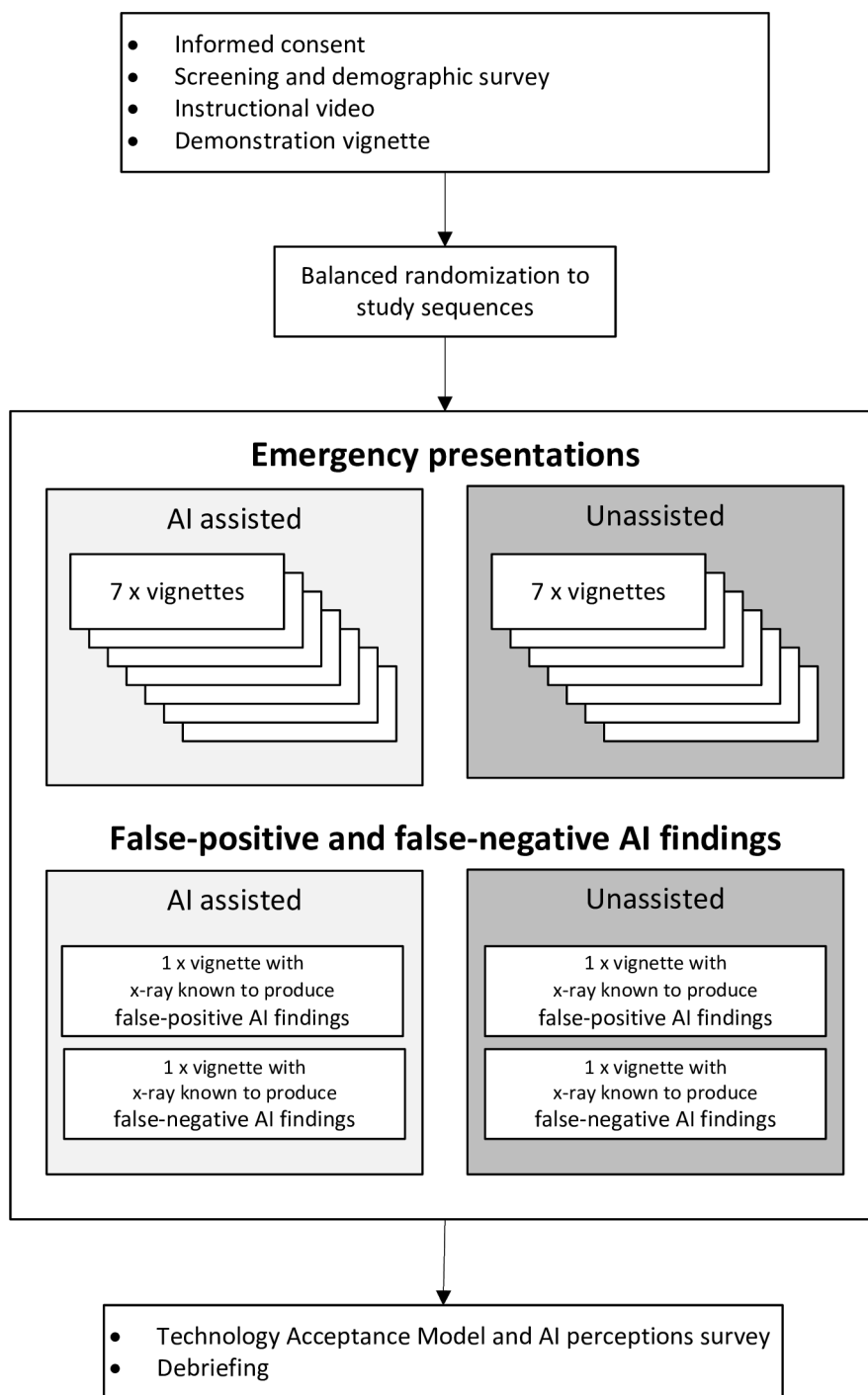


Figure 1 Study design.

assist doctors with chest X-ray interpretation and has been approved for use as a class 1 medical device in Australia,¹⁵ and a class IIb medical device in the EU and the UK. Unlike other AI tools, which are restricted to a small number of findings for specific conditions, the Annalise Enterprise CXR detects 124 radiological findings and provides localisation information for certain findings. Further details on model training and testing are reported by Seah *et al.*¹¹

Our study set-up mimicked a likely real-world implementation where the AI tool processes X-rays from the hospital picture archiving and communication system (PACS) and its findings are saved back to the PACS and can be viewed alongside the X-ray within the hospital's radiological information system. The

processing of X-rays to generate the AI findings was undertaken following vignette creation enabling validation by the expert panel.

Procedure

Participants self-enrolled in the study online. After consenting, eligibility screening and providing demographic information, they watched a 4.5-minute instructional video explaining the study task, how to view and interpret AI findings and how to access definitions for all 124 X-ray findings. The instructions provided information from the Annalise Enterprise CXR manual, including indications for use, intended users, intended

patient populations, contraindications and warnings. Participants had the opportunity to explore the user interface and AI findings for a demonstration vignette.

Each vignette was presented in four phases: (1) The presenting complaint with relevant symptoms and observations was displayed. Participants were asked for their provisional diagnosis; (2) an X-ray was shown, and participants were asked about the pathology demonstrated by the X-ray and to provide their diagnosis. For AI-assisted vignettes, the AI findings were presented alongside the X-ray (figure 2b); (3) four patient management options were presented, and participants were asked to select the most appropriate; and (4) finally, participants were asked to rate their confidence in the diagnosis, cognitive load and the extent to which they considered AI findings when making decisions. X-ray interpretation time was automatically recorded. The study was deployed using Gorilla.sc,¹⁶ a platform for conducting online experiments, which was customised to display X-rays within a medical imaging viewer.¹⁷

Diagnosis was recorded as a free-text response to simulate documentation in an electronic medical record. This approach was intended to engage participants in the diagnostic process, including generation of differential diagnoses based on vignettes and X-ray information. However, due to the wide range of acceptable patient management options, free-text responses were less practical, and a multiple-choice single best answer response format was adopted.

Outcome measures

Primary outcomes

1. Correct diagnosis: Responses had to identify the most relevant diagnosis for emergency medicine. Participants' free-text responses were independently scored as correct or incorrect against the gold standard responses for vignettes that had been validated by the expert panel. All but four vignettes had a single diagnosis participants needed to identify. The remaining four required participants to identify both the condition and an iatrogenic injury present in the vignette (eg, pneumothorax and misplaced intercostal catheter; online supplemental appendix B). Responses were scored by two doctors (one a FACEM) who were blinded to whether vignettes were AI assisted. Scoring was calibrated on 540 responses (15% of sample), and inter-rater agreement was very good (Cohen's $\kappa = 0.88$) for the remaining 3060 responses (85%). Disagreements were resolved by consensus.
2. Correct patient management: Responses were multiple choice and automatically scored as correct or incorrect against the validated gold standard response.

Secondary outcomes

1. X-ray interpretation time: Measured from presentation of an X-ray until the participant submitted their responses about the pathology demonstrated and their diagnosis (ie, time to complete the step in figure 2).
2. Confidence in diagnosis: After completing each vignette, participants rated confidence in their diagnosis on a 7-point scale from (1) very unconfident to (7) very confident.
3. Perceptions about the AI tool: Measured in the post-study questionnaire using the Technology Acceptance Model which is intended to predict intentions to use the tool.¹⁸ Four additional questions elicited participants' perceptions about the AI for ED diagnosis (online supplemental appendix E).

We also measured participants' cognitive load and the extent to which AI findings were considered; these analyses will be reported separately.

Analysis

Data from the 200 participants who completed the study and acknowledged the debriefing information were retrieved and included in the analysis, which compared AI-assisted and unassisted vignettes within participants. Effects of AI assistance on correct diagnosis and patient management were examined by comparing the number of correct responses between the seven unassisted and seven AI-assisted vignettes. X-ray interpretation time and confidence in diagnosis were averaged across all seven unassisted vignettes and compared with the seven AI-assisted vignettes.

Correctly diagnosed and managed vignettes, and confidence in diagnoses were evaluated using paired two-tailed t-tests. While these are discrete outcome measures, the robustness of t-tests means they were reasonable for this analysis.

X-ray interpretation times were positively skewed and analysed using Wilcoxon signed-rank test, with differences reported using median and IQR for central tendency. Differences for the false-positive and false-negative AI findings were evaluated within participants with McNemar tests. The effect of AI assistance by seniority of doctors was explored with post hoc two-tailed paired t-tests and evaluated against a Bonferroni corrected alpha of 0.0125.

Data from two vignettes (0.06% of total) belonging to two participants were excluded from the analysis of the seven AI-assisted and unassisted vignettes as responses indicated the X-ray was not displayed due to a technical issue and were thus treated as missing data. Seventy-three observations of X-ray interpretation time (2% of total) exceeding 8 min were considered as outliers and removed from the analysis (online supplemental appendix F).

RESULTS

Participants

A total of 200 emergency doctors completed the study between May and November 2023 (online supplemental appendix G). The median age of the participants was 32 years, 40% identified as female, and they had a median of 3 years of experience in emergency medicine (table 1). 19% of participants (n=46 of 246) who started the study tasks did not complete all vignettes and were excluded from the analysis.

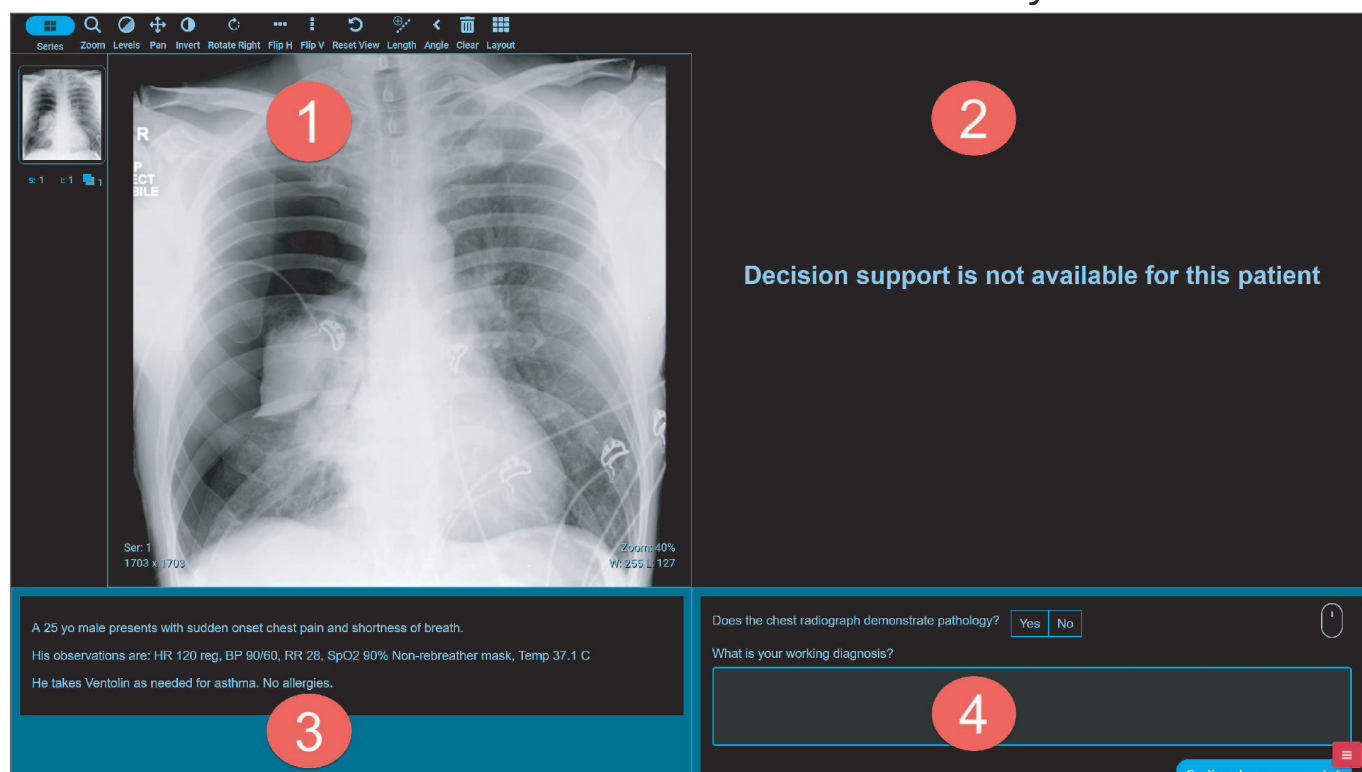
Primary outcomes

The effect of AI assistance on emergency doctors' decision-making was examined by comparing correct diagnosis and patient management between clinical vignettes that were completed with and without AI assistance (table 2, online supplemental appendix H). We found that AI assistance for chest X-ray interpretation significantly increased correct diagnoses by 5.9% (95%CI 2.7% to 9.2%; $p<0.001$) and patient management by 3.2% (95% CI 0.1% to 6.4%; $p=0.04$) compared with vignettes completed without AI assistance.

Secondary outcomes: X-ray interpretation time, confidence in diagnosis and perceptions about the AI tool

Effects on decision-making were also assessed by examining related outcomes such as X-ray interpretation time, confidence in diagnosis and participant perceptions about the AI tool. We found that AI assistance for chest X-ray interpretation increased

a Unassisted examination of chest x-rays



b AI assisted examination of chest x-rays

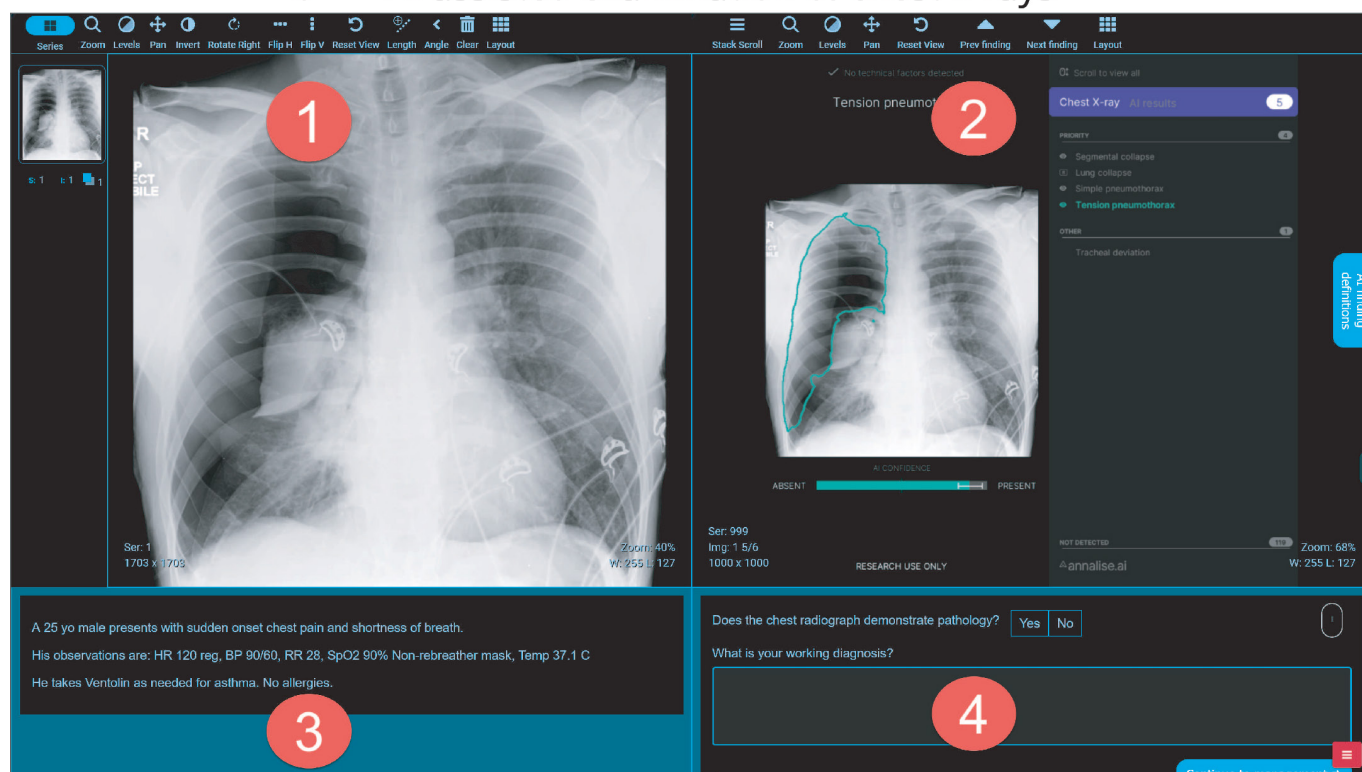


Figure 2 Examination of chest X-rays and reporting diagnosis. The online user interface for the study was divided into four panes: (1) chest X-ray viewer; (2) AI findings viewer; (3) description of the presenting complaint, relevant symptoms and observations; and (4) response panel for recording diagnosis. Panel (A) shows how unassisted vignettes were presented, and panel (B) shows how the same vignette appears with AI assistance. See online supplemental appendix C for a complete example of all AI findings.

Table 1 Sample characteristics of emergency doctors who participated in the study (n=200)

Seniority	n (%)	Experience in emergency medicine Median of years (IQR)	Median age (IQR)*	% female†
Emergency consultants Fellows/board certified	41 (20.5%)	12 (10–20)	41 (37–47)	32
Emergency registrars Emergency trainee	53 (26.5%)	5 (3.3–7)	32.5 (30–35)	40
Senior resident medical officers	58 (29%)	2 (1–4)	32 (28–39)	40
Junior medical officers	48 (24%)	1 (0.5–1)	27 (25–28)	48
Total	200	3 (1–8)	32 (28–38)	40

*Thirty-three participants elected not to report their age.
†Two participants elected not to report their gender.

the time taken by emergency doctors to complete clinical vignettes by 4.9s compared with vignettes without AI assistance; however, this difference was not statistically significant ($p=0.08$).

Participants indicated that AI assistance significantly increased their confidence in diagnosis ($p<0.001$; table 2). They agreed (median=4) with statements from the Technology Acceptance Model measuring perceived usefulness, perceived ease of use, actual use and intention to use which are positive predictors of system use. Participants indicated AI assistance: was helpful (median=4); made interpreting chest X-rays easier (median=4); sometimes drew attention to findings that might otherwise have been missed (median=3); and neither trusted nor distrusted its accuracy (median=3; see online supplemental appendix E.)

Subgroup analysis: effect of AI assistance by seniority

The greatest effect of AI assistance on decision-making was observed in senior resident medical officers who were in their third or higher year of postgraduate medical practice. Correct diagnoses significantly increased by 11.8% (95% CI 5.2% to 18.3%; $p<0.001$) compared with vignettes completed without AI assistance (table 3). While improvements were observed in all other levels of seniority, these were smaller and not statistically significant.

Effect of incorrect AI findings on emergency doctors' decision-making

The effect of incorrect AI findings was evaluated by asking participants to complete clinical vignettes involving X-rays that were known to produce false-positive or false-negative AI findings. The four vignettes known to produce incorrect AI findings were

randomised to ensure each was evenly allocated to AI assistance. We found false-negative AI findings non-significantly decreased correct diagnosis by 9% compared with unassisted vignettes ($p=0.1$; table 4). All other differences in correct diagnosis and patient management were negligible ($\leq 2.5\%$ decrease with AI assistance) and not statistically significant.

DISCUSSION

Overall, AI assistance with chest X-ray interpretation significantly increased correct diagnosis by 5.9% (95% CI 2.7% to 9.2%), or for every 17.3 patients where doctors use AI, one misdiagnosis would be prevented (Number Needed to Treat; 95% CI 11 to 43). The greatest improvement was observed among senior resident medical officers at 11.8% (95% CI 5.2% to 18.3%; NNT=8.5 (95% CI 6 to 19)), with smaller, non-significant improvements of 3% to 4.1% observed for other levels of seniority. For patient management, AI assistance led to a smaller but significant improvement of 3.2% (95% CI 0.1 to 6.4%), equivalent to preventing one inappropriate decision for every 33.3 patients where AI is used (95% CI NNT 16 – ∞), although the CI is wide.

Participants' attitudes towards and perceptions about AI assistance were favourable, and their confidence in diagnosis increased when assisted by the AI tool. In the context of favourable technology acceptance, the small and non-significant increase in X-ray interpretation time is unlikely to impede adoption in ED.

Although the improvement in diagnosis observed in this study was relatively small, it is likely to have significant clinical implications given the volume of ED patients requiring chest X-rays, which is increasing relative to all presentations over time.¹⁹ The

Table 2 Effects of AI assistance on the diagnostic and patient management decisions of emergency doctors interpreting chest X-rays (n=198)

	Unassisted	AI assisted	Difference between clinical vignettes completed with and without AI assistance		Paired two-tailed t-test	Effect size Cohen's d*
	Mean (SD)	Mean (SD)	Mean (95% CI)	Percentage (95% CI)		
Correct diagnosis (no. out of 7 vignettes)	4.5 (1.38)	4.9 (1.17)	0.4 (0.19 to 0.64)	5.9% (2.7% to 9.2%)	t(197)=3.608, $p<0.001$	0.26
Correct patient management (no. out of 7 vignettes)	4.9 (1.28)	5.1 (1.10)	0.2 (0.01 to 0.45)	3.2% (0.1% to 6.4%)	t(197)=2.056, $p=0.04$	0.15
Confidence in diagnosis†	4.6 (0.78)	4.9 (0.69)	0.3 (0.20 to 0.40)	5.0% (3.4% to 6.6%)	t(197)=6.139, $p<0.001$	0.44
X-ray interpretation time (s)	Median 87.5 (IQR 63–136)	Median 92.4 (IQR 69–134)			z=−1.743, n=198, $p=0.08$ ‡	–

*Cohen's d: 0.2=small effect; 0.5=medium effect; and 0.8=large effect size.¹⁰

†Measured on a 7-point scale from 1=very unconfident, 4=neither confident nor unconfident, 5=slightly confident, 6=confident, to 7=very confident.

‡X-ray interpretation time analysed using Wilcoxon signed-rank test.

AI, artificial intelligence.

Table 3 Effect of AI assistance on the diagnostic decisions of emergency doctors interpreting chest X-rays by seniority (n=198)

Seniority	n	Unassisted	AI assisted	Difference between clinical vignettes completed with and without AI assistance			Effect size Cohen's d*
		Mean no. of correct diagnoses out of 7 (SD)	Mean no. of correct diagnoses out of 7 (SD)	Mean (95% CI)	Percentage (95% CI)	Paired two-tailed t-test	
Emergency consultants	41	5.2 (1.19)	5.4 (1.16)	0.2 (−0.27 to 0.76)	3.5% (−3.9% to 10.9%)	t(40)=0.952, p=0.35	–
Emergency registrars	52†	5.0 (1.10)	5.3 (0.90)	0.3 (−0.09 to 0.66)	4.1% (−1.2% to 9.5%)	t(51)=1.543, p=0.129	–
Senior resident medical officers	57†	4.0 (1.49)	4.8 (1.29)	0.8 (0.37 to 1.28)	11.8% (5.2% to 18.3%)	t(56)=3.613, p<0.001‡	0.48
Junior medical officers	48	4.1 (1.29)	4.3 (0.99)	0.2 (−0.28 to 0.70)	3.0% (−4.0% to 10.0%)	t(47)=0.855, p=0.4	–

*Cohen's d: 0.2=small effect; 0.5=medium effect; and 0.8=large effect size.¹⁰

†One registrar and one senior resident were affected by missing data.

‡Statistically significant with Bonferroni adjusted alpha of 0.0125.

AI, artificial intelligence.

number needed to treat for benefit associated with AI assistance was around 17, which in the clinical context of a busy ED, is a relatively small number. Clinical management often relies on correct interpretation of X-rays which can be challenging in time-critical cases requiring resuscitation where expert interpretation by a radiologist may not be immediately available and clinical decisions need to be made rapidly. Misinterpretation of chest X-ray findings can result in misdiagnosis or delays in management, resulting in poorer patient outcomes.²⁰ For instance, delayed antibiotics for pneumonia increase the risk of adverse patient outcomes.²¹ Technologies, such as the evaluated chest X-ray AI tool, that can improve the rate of correct diagnosis may therefore assist in reducing diagnostic and management errors in the ED.

Accordingly, these findings suggest potential utility for AI assistance with chest X-ray interpretation to inform diagnostic and patient management decisions, meriting further research in a live clinical environment to ascertain whether improvements observed with clinical vignettes translate to real-world emergency settings.

Doctor seniority and experience

AI-assisted improvements in correctly diagnosed vignettes varied by seniority, with the largest improvement seen among senior resident medical officers having 1–4 years of emergency experience. The reduced benefit for the more experienced consultants and registrars, who were emergency specialists or undertaking specialist training, is unsurprising, given the inverse relationship between expertise and the benefit of decision support.^{22 23} These findings are consistent with Novak *et al*⁸ Yet, that benefit

did not extend to junior medical officers, with less experience compared with senior residents. This may suggest a hypothesis to be explored in future work, that junior medical officers who practise under supervision are still developing the clinical experience necessary to effectively integrate AI findings.

The effects of false-positive and false-negative AI findings

The risk of incorrect decision support biasing doctors and potentially leading to misdiagnoses, a phenomenon known as automation bias, has long been of concern^{24 25} and has been observed in computer-aided interpretation of electrocardiograms²⁶ and screening mammograms.²⁷

In our study, there was a 9% decrease in correct diagnoses when participants were provided with false-negative AI findings relevant to the true diagnosis. While not statistically significant, the magnitude suggests the sample may have been underpowered to detect the observed difference and confirmation with a larger sample is needed. While the risk from false-negative or missed X-ray AI findings is not definitively established, clinical studies would be well advised to test for, monitor and evaluate effects of false negatives.

Strengths and limitations

The 200 doctors were a good sample equivalent to approximately 7.6% of the 2616 Australian doctors who reported working in emergency medicine in 2022.²⁸ Likewise, there was good representation of doctors at each level of seniority across the full range of clinical experience and specialist training of doctors working in ED. Indeed, the sample is substantially larger than the median of four participants (IQR 3–8) in the studies reviewed by Vasey *et al*²⁹ to assess the effect of AI assistance on clinician diagnostic performance.

While the online study facilitated the obtained sample, there was an attrition rate of 37% of all eligible participants (online supplemental appendix G). The largest portion of attrition (23%; 72 of 318) occurred during the pre-experiment questionnaire and instructions, which provided a realistic preview of the study to ensure alignment between participant expectations and study requirements. We did not analyse the characteristics of non-completing participants, as we could not determine whether early exits indicated withdrawal of consent. Nor could we determine whether any participants who exited during the instructions later returned to complete the study.

Table 4 Effect of incorrect AI findings on the diagnostic and patient management decisions of emergency doctors (n=200)

	Unassisted	AI assisted	McNemar test
	n (%)	n (%)	
Correct diagnosis			
False-positive AI finding	83 (42%)	81 (41%)	McNemar, n=200, p=0.9
False-negative AI finding	88 (44%)	70 (35%)	McNemar, n=200, p=0.1
Correct patient management			
False-positive AI finding	139 (70%)	138 (69%)	McNemar, n=200, p=1
False-negative AI finding	112 (56%)	107 (54%)	McNemar, n=200, p=0.7

AI, artificial intelligence.

Though the sample was likely underpowered to detect effects of false-negative AI findings on correct diagnosis using McNemar tests, the number of vignettes with incorrect AI findings was intentionally limited. This was to avoid confounding, given the known causal relationship between the perceived reliability of automation and susceptibility to bias.³⁰ People tend to be less influenced by decision support they perceive as inaccurate or unreliable. Including more vignettes with incorrect AI artificially increases the AI error rate and, in turn, influences participants' perceptions of AI accuracy.

Although participants were instructed to approach the study as if treating a real patient, the clinical vignettes were simulated, therefore any diagnostic or management errors had no consequences and the resulting decisions made in the absence of patients more theoretical.

We evaluated an AI tool that detects up to 124 radiological findings; therefore, our findings may not generalise to other tools detecting fewer findings. Nevertheless, the broad applicability of the tool enabled testing across the spectrum of conditions presenting to ED.

CONCLUSION

AI assistance with chest X-ray interpretation has potential to improve decision-making in ED and merits further evaluation to ascertain whether the improvements observed with clinical vignettes can be translated to improved patient outcomes in real-world emergency settings.

Acknowledgements The authors wish to acknowledge the following contributions to the conduct of the study: the emergency doctors who gave their time to participate in the study; the Fellows of the Australasian College for Emergency Medicine and Royal Australian and New Zealand College of Radiologists, who sat on the expert panel; Jade Barton, who developed the task interface for the study; and Satya Vedantam, who developed the underlying study platform; Dr Peter Petocz, who advised on and reviewed the statistical analyses; Kalissa Brooke-Cowden and Denise Tsiros, who contributed research and administrative support; Dr Ying Wang, who contributed machine learning expertise. We wish to specifically acknowledge the contributions of those involved in the recruitment of emergency doctors: the principal investigators at Royal Prince Alfred, Canterbury, Concord Repatriation General and Westmead Hospitals; the Agency for Clinical Innovation (ACI) Emergency Care Institute (ECI); My Emergency Doctor and the Australian Medical Association. Finally, the contributions of all project partners, the Digital Health CRC, Sydney Local Health District and Annalise.ai, and especially the contributions of Dr Nalan Ektas, Dr Mel Ryan, Dr Mark Phillips, Professor Catherine Jones, Dr Hassan Ahmad, Dr John Lambert and Georgie Bottrell from Annalise.ai who participated in the project steering committee or contributed technical expertise on the chest X-ray AI.

Contributors DL, FM, MD, MG and EC conceived this study. DL and FM designed and conducted the study, with RS, MG and MD assisting with participant recruitment. NA, BAC, MD and MG developed the clinical vignettes. ES and APS scored the participant responses. DL and FM conducted the analysis, with the discussion contributed by DL, FM, MD, MG and EC, with additional input from ES and APS. DL, MD and FM drafted the manuscript with input from all authors. All authors provided revisions for intellectual content. All authors have approved the final manuscript. The authors were not precluded from accessing the study data, and they accept responsibility to submit for publication. FM is the guarantor.

Funding The study was supported by Digital Health CRC Limited ("DHCRC"). DHCRC is funded under the Australian Commonwealth's Cooperative Research Centres (CRC) programme, an Australian Government initiative for collaboration between university researchers and industry. The industry partner is Annalise.ai Pty Ltd, who developed and marketed the Annalise Enterprise CXR tool evaluated in the study.

Competing interests Annalise.ai had employee representatives on the project steering committee, but were not involved in data collection, analysis or interpretation, nor were they involved with the selection, design or scoring of clinical vignettes. Their involvement in recruitment was limited to forwarding the study invitation to potential participants and was only one of many channels for recruitment. The authors have no professional or personal affiliation with Annalise.ai.

Patient consent for publication Not applicable.

Ethics approval Ethics approval was granted by the Northern Sydney Local Health District Human Research Ethics Committee (2021/ETH11147) and participants were offered a AUD\$100 gift voucher. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information. Participating doctors consented to the use of study results for aggregate reporting only as a measure to protect their identity and confidentiality.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

David Lyell <https://orcid.org/0000-0002-2695-0368>
Anindya Pradipta Susanto <https://orcid.org/0000-0001-5155-6904>
Bashir Antoine Chakar <https://orcid.org/0009-0000-0571-5222>
Radhika V Seimon <https://orcid.org/0000-0002-3903-4801>
Farah Magrabi <https://orcid.org/0000-0002-8426-5588>

REFERENCES

- Grant K, McParland A, Mehta S, *et al*. Artificial Intelligence in Emergency Medicine: Surmountable Barriers With Revolutionary Potential. *Ann Emerg Med* 2020;75:721–6.
- Garg N. Artificial Intelligence in Emergency Medicine: A Case for More. *Ann Emerg Med* 2024;84:154–6.
- Petrella RJ. The AI Future of Emergency Medicine. *Ann Emerg Med* 2024;84:139–53.
- Susanto AP, Lyell D, Widyantoro B, *et al*. Effects of machine learning-based clinical decision support systems on decision-making, care delivery, and patient outcomes: a scoping review. *J Am Med Inform Assoc* 2023;30:2050–63.
- Magrabi F, Lyell D, Coiera E. Automation in Contemporary Clinical Information Systems: a Survey of AI in Healthcare Settings. *Yearb Med Inform* 2023;32:115–26.
- Han R, Acosta JN, Shakeri Z, *et al*. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. *Lancet Digit Health* 2024;6:e367–73.
- Rudolph J, Huemmer C, Preuhs A, *et al*. Nonradiology Health Care Professionals Significantly Benefit From AI Assistance in Emergency-Related Chest Radiography Interpretation. *Chest* 2024;166:157–70.
- Novak A, Ather S, Gill A, *et al*. Evaluation of the impact of artificial intelligence-assisted image interpretation on the diagnostic performance of clinicians in identifying pneumothoraces on plain chest X-ray: a multi-case multi-reader study. *Emerg Med J* 2024;41:e213620:602–9.
- Coiera E. Assessing technology success and failure using information value chain theory. In: Scott P, Keizer N, Georgiou A, eds. *Applied Interdisciplinary Theory in Health Informatics*. IOS Press, 2019: 35–48.
- Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd edn. Hillsdale, N.J.: L. Erlbaum Associates, 1988.
- Seah JCY, Tang CHM, Buchlak QD, *et al*. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *Lancet Digit Health* 2021;3:e496–506.
- Obuchowski NA, Gallas BD, Hillis SL. Multi-reader ROC studies with split-plot designs: a comparison of statistical methods. *Acad Radiol* 2012;19:1508–17.
- Wickens CD, Xu X. Automation trust, reliability and attention: technical report ahfd-02-14/maad-02-2: aviation human factors division. 2002.
- Australasian College for Emergency Medicine. Curriculum 2022: fellowship of the australasian college for emergency medicine. secondary curriculum 2022: fellowship of the australasian college for emergency medicine. 2023. Available: <https://acem.org.au/getmedia/9af41df8-677f-44ed-b245-440164155f56/FACEM-Curriculum>
- Therapeutic Goods Administration. Australian Register of Therapeutic Goods: Annalise-AI Pty Ltd - Radiology DICOM image processing application software (343577). Secondary Australian Register of Therapeutic Goods: Annalise-AI Pty Ltd - Radiology DICOM image processing application software (343577), 2020. Available: <https://www.tga.gov.au/resources/artg/343577>

- 16 Anwyl-Irvine A, Dalmaijer ES, Hodges N, *et al.* Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behav Res Methods* 2021;53:1407–25.
- 17 Ziegler E, Urban T, Brown D, *et al.* Open Health Imaging Foundation Viewer: An Extensible Open-Source Framework for Building Web-Based Imaging Applications to Support Cancer Research. *JCO Clin Cancer Inform* 2020;4:336–45.
- 18 Holden RJ, Karsh B-T. The technology acceptance model: its past and its future in health care. *J Biomed Inform* 2010;43:159–72.
- 19 Poyiadji N, Beauchamp N 3rd, Myers DT, *et al.* Diagnostic Imaging Utilization in the Emergency Department: Recent Trends in Volume and Radiology Work Relative Value Units. *J Am Coll Radiol* 2023;20:1207–14.
- 20 Institute of Medicine National Academies of Sciences E, Medicine. Improving Diagnosis in Health Care. Washington, DC: The National Academies Press, 2015.
- 21 Zasowski EJ, Bassetti M, Blasi F, *et al.* A Systematic Review of the Effect of Delayed Appropriate Antibiotic Treatment on the Outcomes of Patients With Severe Bacterial Infections. *Chest* 2020;158:929–38.
- 22 Povyakalo AA, Alberdi E, Strigini L, *et al.* How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography. *Med Decis Making* 2013;33:98–107.
- 23 Riley V. Operator reliance on automation: theory and data. Automation and human performance: CRC Press; 2018:19–35.
- 24 Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc* 2012;19:121–7.
- 25 Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc* 2017;24:423–31.
- 26 Bogun F, Anh D, Kalahasty G, *et al.* Misdiagnosis of atrial fibrillation and its clinical consequences. *Am J Med* 2004;117:636–42.
- 27 Dratsch T, Chen X, Rezazade Mehrizi M, *et al.* Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. *Radiology* 2023;307:e222176.
- 28 Australian government department of health and aged care. Health workforce data. secondary health workforce data. 2023. Available: <https://hwd.health.gov.au>
- 29 Vasey B, Ursprung S, Beddoe B, *et al.* Association of Clinician Diagnostic Performance With Machine Learning-Based Decision Support Systems: A Systematic Review. *JAMA Netw Open* 2021;4:e211276.
- 30 Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. *Hum Factors* 2010;52:381–410.