OPEN ACCESS

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

Alex Novak ,[1] Sarim Ather,[2] Avneet Gill,[3] Peter Aylward,[4] Giles Maskell,[5] Gordon W Cowell,[6] Abdala Trinidad Espinosa Morgado ,[1] Tom Duggan,[7] Melissa Keevill,[3] Olivia Gamble,[3] Osama Akrama,[8] Elizabeth Belcher,[3] Rhona Taberham,[3] Rob Hallifax,[3] Jasdeep Bahra ,[3] Abhishek Banerji,[7] Jon Bailey ,[3] Antonia James,[3] Ali Ansaripour,[3] Nathan Spence,[3] John Wrightson,[3] Waqas Jarral,[9] Steven Barry,[3] Saher Bhatti,[9] Kerry Astley,[3] Amied Shadmaan,[10] Sharon Ghelman,[11] Alec Baenen,[12] Jason Oke,[13] Claire Bloomfield,[14] Hilal Johnson,[14] Mark Beggs,[14] Fergus Gleeson[2]

Handling editor Shammi L Ramlakhan

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/emermed-2023-213620).

For numbered affiliations see end of article.

Correspondence to
Dr Alex Novak;
alex.novak@ouh.nhs.uk

Received 15 September 2023
Accepted 10 June 2024

Check for updates

© Author(s) (or their employer(s)) 2024. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Novak A, Ather S, Gill A, *et al*. *Emerg Med J* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/emermed-2023-213620

ABSTRACT

Background Artificial intelligence (AI)-assisted image interpretation is a fast-developing area of clinical innovation. Most research to date has focused on the performance of AI-assisted algorithms in comparison with that of radiologists rather than evaluating the algorithms' impact on the clinicians who often undertake initial image interpretation in routine clinical practice. This study assessed the impact of AI-assisted image interpretation on the diagnostic performance of frontline acute care clinicians for the detection of pneumothoraces (PTX).

Methods A multicentre blinded multi-case multi-reader study was conducted between October 2021 and January 2022. The online study recruited 18 clinician readers from six different clinical specialties, with differing levels of seniority, across four English hospitals. The study included 395 plain CXR images, 189 positive for PTX and 206 negative. The reference standard was the consensus opinion of two thoracic radiologists with a third acting as arbitrator. General Electric Healthcare Critical Care Suite (GEHC CCS) PTX algorithm was applied to the final dataset. Readers individually interpreted the dataset without AI assistance, recording the presence or absence of a PTX and a confidence rating. Following a 'washout' period, this process was repeated including the AI output.

Results Analysis of the performance of the algorithm for detecting or ruling out a PTX revealed an overall AUROC of 0.939. Overall reader sensitivity increased by 11.4% (95% CI 4.8, 18.0, p=0.002) from 66.8% (95% CI 57.3, 76.2) unaided to 78.1% aided (95% CI 72.2, 84.0, p=0.002), specificity 93.9% (95% CI 90.9, 97.0) without AI to 95.8% (95% CI 93.7, 97.9, p=0.247). The junior reader subgroup showed the largest improvement at 21.7% (95% CI 10.9, 32.6), increasing from 56.0% (95% CI 37.7, 74.3) to 77.7% (95% CI 65.8, 89.7, p<0.01).

Conclusion The study indicates that AI-assisted image interpretation significantly enhances the diagnostic accuracy of clinicians in detecting PTX, particularly benefiting less experienced practitioners. While overall interpretation time remained unchanged, the use of AI improved diagnostic confidence and sensitivity, especially among junior clinicians. These findings underscore the potential of AI to support less skilled clinicians in acute care settings.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Artificial intelligence (AI)-assisted image interpretation algorithms can be used to accurately detect the presence of pathological findings on retrospective image datasets and improve radiologist performance, but their impact on frontline clinicians is unclear.

WHAT THIS STUDY ADDS

⇒ The diagnostic accuracy and confidence of clinicians in detecting pneumothorax (PTX) on plain CXR significantly improved with the use of AI-assisted image interpretation.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study provides evidence that AI-assisted image interpretation may be used to improve the diagnostic performance of clinicians, especially those more junior, in detecting pathological findings such as PTX.

INTRODUCTION

Artificial intelligence (AI) has transformative potential for healthcare. By far the most promising

and mature use-case for AI in clinical practice is in AI-assisted image interpretation, which has accounted for a significant majority of academic publications and FDA certifications,[1 2] and remains a rapidly developing area of innovation.[3 4] Much AI imaging research has focused on the development, validation and evaluation of algorithms as measured against the diagnostic performance of senior radiologist readers.[5–9] However, in the acute healthcare setting, many diagnostic images are typically reviewed and acted on by non-radiologists with varying degrees of clinical experience and expertise.[10] The potential impact of AI-assisted image interpretation on the diagnostic accuracy of clinicians who are directly involved in interpreting images and delivering care to patients based on their findings in routine clinical practice therefore remains an important research question, and studies have begun to demonstrate potential benefits in this regard in an emergency medicine context.[11]

Recent guidance from NICE[12] and AI-specific reporting guidelines have emphasised the importance of conducting evaluations in the clinical context in which they are likely to be cited, including feedback on usability and confidence directly from the intended users.[13–17]

Aims

To measure the diagnostic accuracy of the pneumothorax (PTX) detection facility of GEHC's CCS application against an independent reference standard and assess its impact on the reporting performance of clinician groups routinely involved in the diagnosis and management of PTXs.

METHODS

Design

A multicentre cohort multi-case multi-reader study was conducted between 6 October 2021 and 27 January 2022 (figure 1).

The Critical Care Suite, an FDA-approved and CE-marked suite of AI-based applications from GE, includes an algorithm for detecting and localising PTX on frontal CXRs. The output is presented to the clinician in the form of a 'heatmap' localising the area on the CXR image with a detected pneumothorax and a corresponding confidence score. This is accompanied by a variable cut-off threshold for the confidence score that determines the level at which the algorithm identifies an image as positive for the presence of PTX (online supplemental figure 1).[18]

We evaluated the impact of this AI-assisted image interpretation algorithm on the diagnostic performance and confidence of a range of clinicians involved in identifying and acting on PTXs found on plain chest radiography in routine clinical practice. In lieu of awaited AI-specific guidelines, this manuscript follows the STARD reporting guidelines for studies evaluating the performance of diagnostic tests in clinical practice.[19 20]

Image dataset

A total of 400 retrospectively collected and de-identified CXR images of patients aged 18 years or older were identified by a research radiographer by searching the radiology database from Oxford University Hospitals and subsequently curated into the project via the National Consortium of Intelligent Medical Imaging (NCIMI) databank (online supplemental table 1). 200 images positive for PTX were selected via radiology reports, with 50 of the following subtypes identified from note review and/or radiology request:

1. Primary PTX (no underlying lung disease).
2. Secondary PTX (underlying lung disease, eg, COPD).
3. Iatrogenic PTX (eg, cardiothoracic surgery).
4. Traumatic PTX.

Difficulty scoring

All images with positive PTXs were assigned a 'Difficulty Score' by a consultant radiologist and a senior radiographer (online supplemental table 2), considering four contributing factors, specifically:

- Size of PTX (large is considered >0.5 cm at any point).
- Patient-specific factors that is, kyphosis/obesity.
- Image quality, specifically exposure factors and image penetration.
- Presence of foreign bodies, artefacts and other pathological findings on the image.

Images without PTXs were also allocated a 'Difficulty Score', excluding the 'PTX Size' parameter.

Ground truth

Readings by senior consultant thoracic radiologists from Royal Cornwall, Greater Glasgow and Clyde and Oxford University Hospitals determined ground truth (gold standard equivalent).

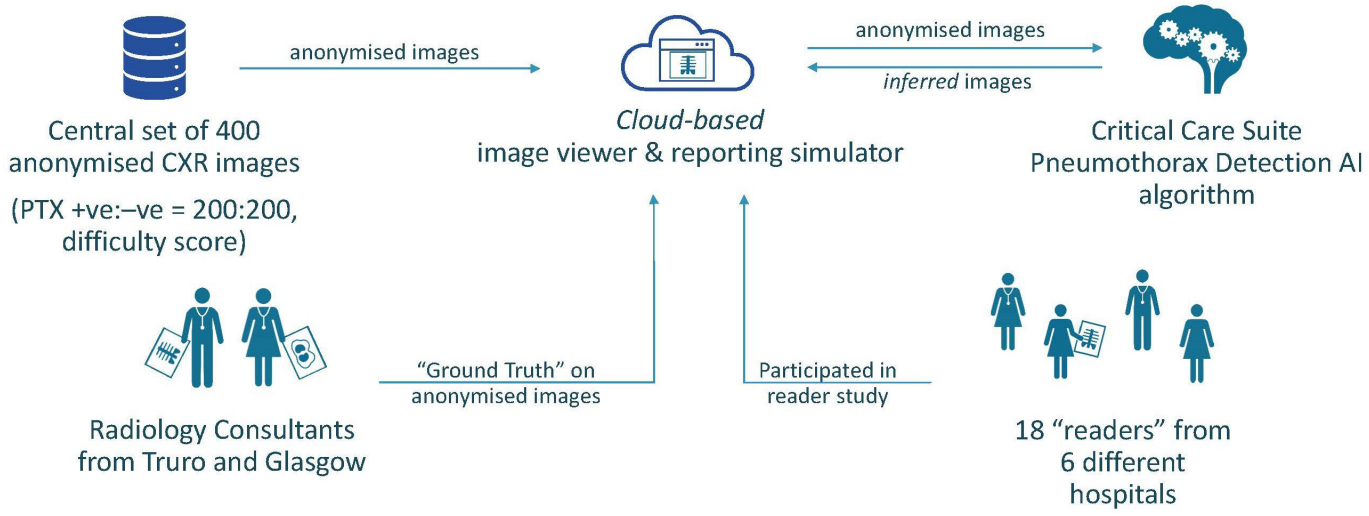


Figure 1 PTX reader study activity flowchart. PTX, pneumothorax.

All CXR images were independently reviewed first by two radiologists for the presence/absence of PTX using a web-based DICOM viewer (www.raiqc.com). Images were annotated based on the presence or absence of a PTX, and if deemed present, a region of interest (ROI) was applied covering the entire PTX. In cases where there was discordance between the two radiologists' interpretation, arbitration was undertaken by the third radiologist.

Participants

To evaluate the impact of the algorithm on human diagnostic performance, an observer performance test was performed by a reader panel comprised of 18 clinicians at three levels of seniority: consultant/senior (>7 years' experience), middle grade/registrar (4–7 years), junior (<4 years) equally derived from six different clinical specialties (Emergency Medicine, Acute General/Internal Medicine, Respiratory Medicine, Intensive Care, Cardiothoracic Surgery and Radiography), who were working across four NCIMI hospital sites via the Thames Valley Emergency Medicine Research Network (www.TaVERNresearch.org).

Reader phases

The study included two reader phases. After enrolment, readers were given access to a short online training module with a series of five test cases to familiarise them with the study and Report and Image Quality Control (RAIQC) platform. They were then given access to four dedicated modules on the RAIQC platform (first reader phase) and asked to interpret the entire dataset over a period of 3 weeks, recording the perceived presence/absence of a PTX on each image and their degree of confidence on a four-point Likert scale. Readers were asked to localise any PTX identified by clicking on the ROI—the identification of a PTX would only be marked as 'correct' if they clicked within the area contoured during the ground truth process. Readers were given access to the clinical indication as entered on the original X-ray request form for each image but were blinded to the ground truth for each image, to the overall performance of the algorithm against ground truth and to the overall prevalence of PTX in the dataset (online supplemental figure 2).

The first phase was followed by a minimum 5 week 'washout' period, following which all readers re-interpreted the images with access to an extra version of the image depicting the AI output (online supplemental figure 3). Qualitative surveys based on the perceived utility and value of AI-assisted image interpretation were completed by all participants before and after each reader phase.

Readings for each phase were undertaken remotely online on a laptop or PC in a location of the participants' choosing, in single or multiple sessions as preferred. The dataset was divided into four equally sized modules, and the sequence of the CXRs was randomised within each module for each individual reader during both phases.

Automated time measurements for the completion of each case were undertaken through a dedicated function in the RAIQC platform.

Statistical analysis

The primary outcome was the difference in diagnostic characteristics (sensitivity, specificity, accuracy) of readers without and with AI augmentation. Secondary outcomes included comparative analysis of performance by subgroups by medical profession, level of seniority and image difficulty score, including type, location and size of PTX, and time taken to complete the reads.

The stand-alone algorithm performance set at varying sensitivity thresholds for labelling an image as positive for PTX was assessed by calculating the area under the curve of the receiver operating characteristic (ROC) and free-response receiver operating characteristic (FROC) curves plotted with their variance. To account for correlated errors arising from readers interpreting the same cases with and without AI, the Obuchowski and Rockette, Dorfman-Berbaum-Metz (OR-DBM) procedure, a modality-by-reader random effects analysis of variance model was used for estimation.[21–23]

Sample size

We used the tool 'Multi-Reader Sample Size Program for Diagnostic Studies'[24] to estimate power for the number of readers cases in our study (https://perception.lab.uiowa.edu/power-sample-size-estimation). For 18 readers, reading 400 cases yields 80% power to detect a difference in the accuracy of 10% with a type I error of 5%.[24]

Statistical analyses were all performed using R software (V.4.0.2; R Foundation for Statistical Computing). The significance threshold was set at two-sided 5% (p=0.05) for all secondary analyses.

Role of the funding sources

This study was funded by Innovate UK and by GE Healthcare. Research activity, including data analysis, and decision to publish, was conducted independently of the funders with the exception that AI inferencing of the CXR images was undertaken by GEHC.

RESULTS

Dataset characteristics

Of the 200 cases positive for pneumothorax, five cases failed the pre-inferencing quality check and were removed from the image dataset, leaving a total of 395 radiographs. Six PTX-positive images were reclassified as negative following the radiologist review and labelling (ground truth), leaving 189 positive for PTX and 206 negative.

Algorithm performance

138 cases out of 189 positive cases were correctly identified by the AI algorithm as positive for pneumothorax, with 50 false negatives. 198 cases out of 206 negative cases were classified correctly (online supplemental figure 4). Analysis of the performance of the algorithm alone at default settings compared with the ground truth (thoracic radiologists) demonstrated a sensitivity of 0.73 (95% CI 0.66, 0.80) and a specificity of 0.96 (95% CI 0.92, 0.98), with a positive predictive value of 0.95 (95% CI 0.89, 0.98), and a negative predictive value of 0.79 (95% CI 0.74, 0.84), with an overall AUROC of 0.94 based on a variable sensitivity threshold (table 1) (online supplemental table 3, online supplemental figure 5).

Reader performance

All participants interpreted all images with and without AI, generating a total of 14 220 image reads. Pooled analysis demonstrated an overall sensitivity of 66.8% (95% CI 57.3, 76.2) without AI increasing to 78.1% (95% CI 72.2, 84.0, p=0.002) with AI, and a non-statistically significant increase in specificity from 93.9% (95% CI 90.9, 97.0) without AI to 95.8% (95% CI 93.7, 97.9, p=0.247) with AI. Overall diagnostic performance characteristics for readers compared with ground truth with and without AI are summarised in table 2 and figure 2. Intraclass

Table 1 Diagnostic performance of algorithm performance vs ground truth

Performance metric	Estimate	Lower 95% CI	Upper 95% CI
True positives	47.8%	42.8%	52.9%
Sensitivity	73.0%	66.1%	79.2%
Specificity	96.1%	92.5%	98.3%
Accuracy	85.1%	81.2%	88.4%
PPV	94.5%	89.5%	97.6%
NPV	79.5%	74.0%	84.4%
LR+	18.8	9.5	37.3
LR–	0.28	0.22	0.36

LR+, likelihood ratio for positive test result; LR–, likelihood ratio for negative test result; NPV, negative predictive value; PPV, positive predictive value.

correlation (online supplemental table 4) improved significantly from 0.575 (95% CI 0.536, 0.615) unaided to 0.802 (95% CI 0.778, 0.825) aided. There was a marked and consistent improvement in pooled reader accuracy between phases 1 and 2 (online supplemental figure 6). Accuracy increased throughout the aided phase compared with the unaided phase, with decreased variation and an overall inclined gradient in the aided phase, reflected in the improved sensitivity scores in modules 3 and 4 (online supplemental table 5).

Individual reader performance

A summary of individual reader sensitivity with and without AI is presented in online supplemental table 6. 11 out of 18 readers showed improvement in sensitivity with AI, with 3 readers showing a small, non-statistically significant increase. Three senior readers (#12, 16 and 17) showed a non-statistically significant decrease in sensitivity, and one senior reader (#5) showed a small but statistically significant decrease in sensitivity with AI.

One junior reader (#9) did not adhere correctly to the technical specifications of the study, using hardware incompatible with the RAIQC platform. Due to this protocol deviation, they were asked to repeat both reading phases and the subsequent results were included in the final analysis. Post hoc sensitivity analysis demonstrated no change in the overall reader performance as a result of this update. The highest improvements in sensitivity were seen in those readers with the lowest unaided sensitivity (figure 3).

Subgroup analyses

Across specialty groups, the highest performance and smallest increase in sensitivity was seen in the cardiothoracic surgery subgroup, with ITU clinicians showing the lowest performance and largest improvement; however, no specialty subgroup difference reached statistical significance (online supplemental table 7). A very large statistically significant increase in sensitivity 21.7% (95% CI 10.9 to 32.6), p<0.01) was seen in the aided junior reader subgroup compared with that seen in the middle 6.3% (95% CI −4.8 to 17.5, p=0.21) and senior 6.0% (95% CI −6.0 to 18.0, p=0.26) groups (figure 4). Aided junior grade readers increased in sensitivity to a comparable level with that of aided middle and senior grades and higher than that of unaided senior grades with a significant increase in intraclass correlation in all three groups (online supplemental table 8). Statistically significant increases in sensitivity with AI were seen across all types of PTX, and all PTX locations except medial, which showed a non-statistically significant increase (5.1%, 95% CI −3.7 to 13.8, p=0.25). Statistically significant increases in sensitivity were shown for X-rays of all levels of difficulty, though increases were non-statistically significant in the higher obesity and smaller PTX subcategories.

Reader reporting time

The effect of the use of the CCS PTX AI tool on reader analysis and reporting time is shown in online supplemental figure 7. In the absence of the AI tool, the mean reporting time by all 18 readers was 30.2 s per image (95% CI 24.4 to 37.4 s). while with the AI tool, reporting time took a mean of 26.9 s per image (95% CI 21.8 to 33.4 s; p>0.05 for the effect of AI tool). Extreme outlying results presumed to be due to technical failure (eg, leaving cases open on the browser between read sessions) were winsorised to prevent skewing of the data.

Self-reported confidence

Confidence in correctly interpreted images increased in the aided reader phase compared with the unaided reader phase. The proportion of 'certain' and 'high' confidence interpretations in the correct interpretation category (ie, true positives/true negatives) increased in the aided reader phase (online supplemental figure 8).

DISCUSSION

This study evaluated the impact of an AI-assisted image interpretation on the diagnostic performance of a range of clinicians routinely involved in identifying PTX. Evaluation of the algorithm's performance against a 'ground truth' reference standard of senior radiologist reporting showed high specificity (96%) and moderate sensitivity (73%), with an AUROC of 0.94. Overall group sensitivity improved by 11.4% from 66.8% to 78.1% (p=0.002), with specificity showing a non-significant improvement. Overall accuracy significantly improved between aided and unaided readers with increased intraclass correlation and an overall increase in the magnitude of accuracy improvements throughout the course of the aided reader phase. Clinicians demonstrated no overall increase in reporting time associated with the use of AI. Overall self-reported confidence in an accurate diagnosis improved, while confidence in an inaccurate diagnosis decreased. Subgroup analyses demonstrated increases in sensitivity across all reader and image subgroups, with a marked

Table 2 Overall sensitivity and specificity for readers compared with ground truth without and with AI

Metric	W/o AI (95% CI)	With AI (95% CI)	Diff: with AI – w/o AI (95% CI)	P value
Presence of pneumothorax				
Sensitivity	66.8 (57.3 to 76.2)	78.1 (72.2 to 84.0)	11.4 (4.8 to 18.0)	0.002
False negative fraction (1 – sens)	33.2 (23.8 to 42.7)	21.9 (16.0 to 27.8)	–11.4 (–18.0 to –4.8)	0.002
Specificity	93.9 (90.9 to 97.0)	95.8 (93.7 to 97.9)	1.8 (–1.4 to 5.0)	0.247
False positive fraction (1 – specificity)	6.1 (3.0 to 9.1)	4.2 (2.1 to 6.3)	–1.8 (–5.0 to 1.4)	0.247

AI, artificial intelligence; Diff, difference; Sens, sensitivity; W/O, without.

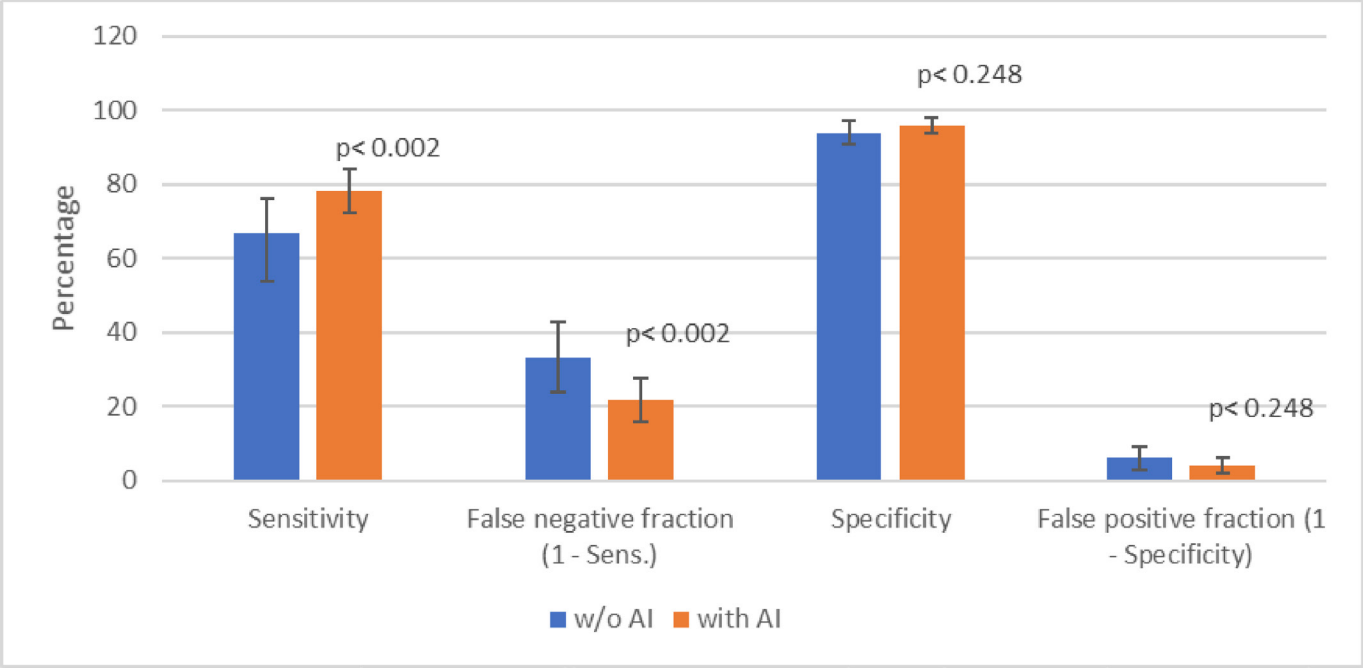


Figure 2 Overall unaided and aided reader diagnostic performance. AI, artificial intelligence.

increase in sensitivity in the junior reader group (21.7%), which improved to a level (77.7%) comparable with that of the aided middle (77.2%) and senior (79.5%) groups, and greater than that of unaided senior groups (73.5%).

There are few existing studies that directly evaluate the impact of AI on the diagnostic performance of clinicians as opposed to radiologists, and none that evaluate the impact of AI-assisted PTX detection on frontline UK clinicians of varying seniority and specialty.[5 9 25] A recent study evaluated the impact of an algorithm which detected four abnormalities on the reporting performance of radiologists of varying experience level. The algorithm in that study showed a 100% sensitivity for pneumothorax, although on a significantly smaller dataset (80 images). Readers showed a higher baseline sensitivity than in our study as might be expected given their skillset, though with a similarly large improvement in aided performance.[5] It is unclear whether the dataset in that study was comparable in terms of complexity to the dataset used in this study. Other studies evaluating the impact of AI on fracture detection on plain radiographs and CT by clinicians have found comparable results to ours in terms of sensitivity and overall improvement.[7 11 26]

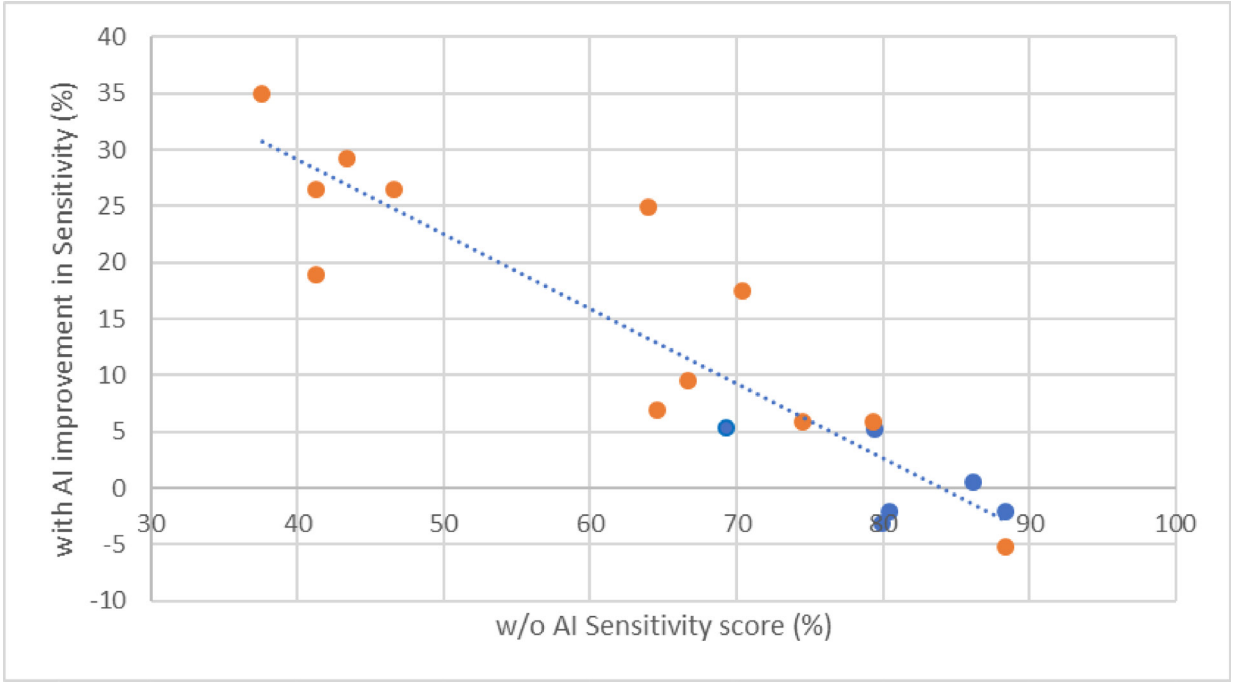


Figure 3 Improvement in sensitivity of individual readers conferred by AI assistance vs initial unaided sensitivity. AI, artificial intelligence.

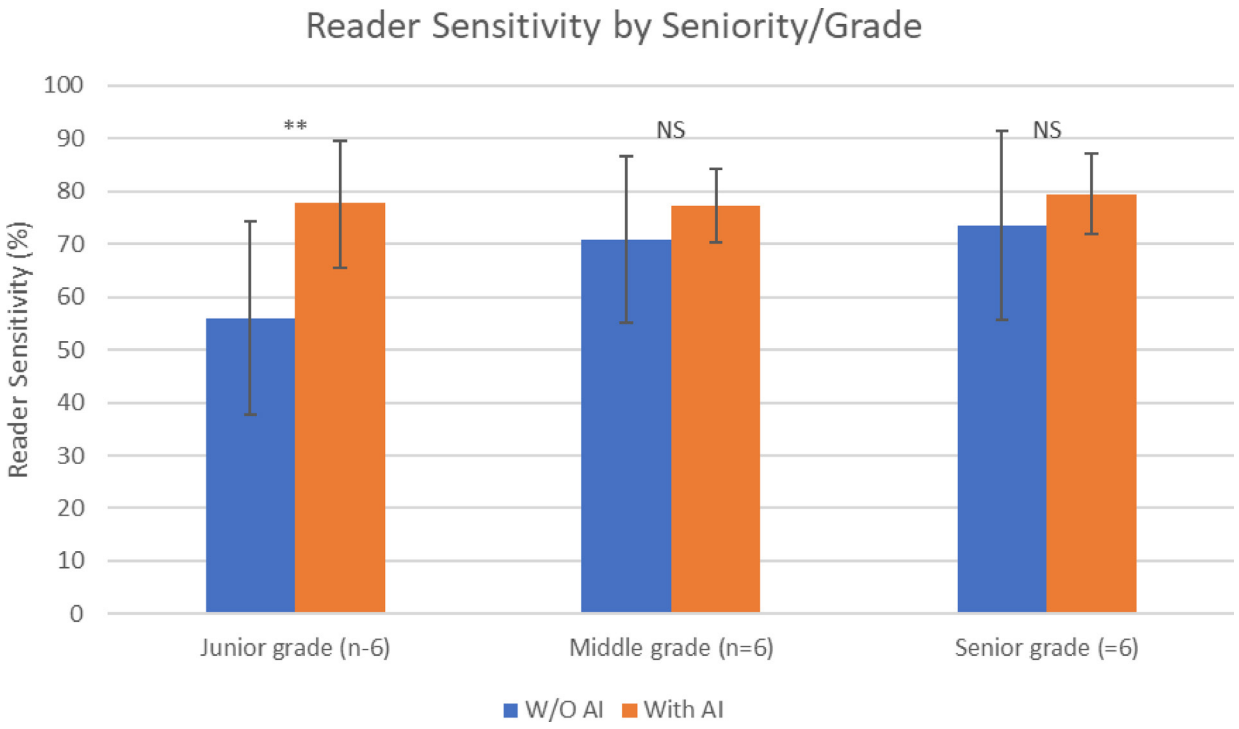


Figure 4 Reader sensitivity by seniority/grade. AI, artificial intelligence.

The large statistically significant improvement seen here appears to have been driven by marked increases in the diagnostic performance of less proficient and less experienced clinicians in the group. The AI tool increased the sensitivity of readers with lower AI-unassisted scores as opposed to those with higher AI-unassisted scores. Indeed, certain high-performing individuals whose unaided sensitivity was higher than that of the algorithm alone did not show a similar increase in accuracy, with some even decreasing their performance in the aided reader phase. This suggests both an important potential use case for AI-enhanced imaging in supporting the performance of less skilled clinicians and equally the need for caution and education of clinicians in terms of implementation. Overall sensitivity and all individual reader sensitivities, with one exception, were higher than the performance of the algorithm alone (73%), suggesting that readers overall appear to have been incorporating AI findings into their diagnostic reasoning rather than blindly following the algorithm, which addresses an important concern regarding the introduction of AI into clinical practice.[27 28] The increase in confidence scores for correct interpretations may indicate that practitioners are likely to appropriately act on their findings when supported by AI. This would however need to be fully evaluated in a prospective study.

The introduction of new AI-based diagnostic technologies which can detect abnormalities in medical imaging may raise concerns regarding the potential for existing clinicians/radiologists to gradually lose existing interpreting skills or to rely solely on the findings of the algorithm.[27 29] Our results however indicate the presence of a potential learning effect associated with access to AI-enhanced imaging. This may have significant implications for potential use cases and implementation strategies for similar technologies. Productivity was unaffected by the algorithm, with the time taken to report on each image being unaffected. Other studies have demonstrated similar findings.[5 7]

Strengths

This prospective study was based on a large combined dataset and reader group in contrast to many other reader studies, with a resultant large number of individual readings for analysis. Comparisons were based on a ground truth of three independent senior radiologist reports which is the current gold standard for similar studies.[7] The study used a challenging dataset reflective of real-world complexity which was detailed and characterised, and readings were undertaken by a broad range of clinicians and radiographers of varying seniority from multiple hospital sites.

Limitations

This study used an artificially enriched dataset in terms of a high prevalence of PTX; this approach is commensurate with other similar studies.[6 30 31] It was based on the detection of a single pathology, which limits the degree to which findings may be generalised to the interpretation of chest radiographs in the acute setting, though other pathologies were included in this study in order to attempt to distract and confound the readers and algorithm in an attempt to reflect 'real-world' practice. Each seniority level had only one representative from each specialty but the clinician pool was large, allowing pooled analysis results to be more generalisable than prior similar studies. It is also feasible that some degree of improvement in performance during the study may have been due to the experience from reporting large numbers of chest radiographs; however, there was a clear separation between the phases in terms of performance, as well as in an apparent increase in performance throughout the second phase not seen in the first. The participants in this study undertook the reader phases in artificial conditions, free from the distractions associated with a busy clinical shopfloor, and were informed that they were specifically looking for the presence of PTX—this may have resulted in their unaided accuracy being higher than might be encountered in real-world practice, potentially reducing the effect of the intervention in this study. Readers undertook the

study using personal or institutional laptops and workstations and were not limited to PACS quality devices which may have reduced their accuracy. Many small PTX may be successfully managed conservatively and therefore possible that the detection of these may not translate directly into meaningful clinical impact. Equally, a false positive finding for a PTX may lead to unnecessary intervention. However, improved identification of PTX through the reduction of false negative results is likely to reduce re-presentation or the need for recall, along with the concomitant clinical governance burden and should ultimately improve patient and clinician experience and potentially reduce healthcare presentations—these aspects should be evaluated in future prospective clinical and health economic studies.

Implications

The findings suggest that AI-assisted image interpretation may significantly improve diagnostic performance and confidence in identifying PTX on chest radiographs, especially for junior or inexperienced clinicians with lower unassisted performance. However, it may adversely affect highly skilled senior clinicians, highlighting the need for appropriate user training before implementation to highlight the operational strengths and weaknesses of the algorithm. These findings may potentially generalise to other pathologies and imaging modalities if supported by comparably efficacious algorithms. The study also raised the possibility of a learning effect from repeated AI-assisted interpretation, which warrants further exploration.

CONCLUSIONS

This study demonstrates that AI-assisted image interpretation can potentially improve clinicians' performance in detecting pathologies like PTX, with the most marked improvement seen in less proficient individuals, without increasing interpretation time. It provides evidence that AI may support junior practitioners in non-specialist acute settings by improving diagnostic performance. However, definitive evidence and the magnitude of potential clinical and health economic benefits require further study.

Author affiliations
[1]Emergency Department, Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[2]Radiology Department, Oxford University Hospitals, Oxford, UK
[3]Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[4]Report and Image Quality Control (RAIQC), London, UK, UK
[5]Royal Cornwall Hospitals NHS Trust, Truro, UK
[6]Queen Elizabeth University Hospital Campus, Glasgow, UK
[7]Buckinghamshire Healthcare NHS Trust, Amersham, UK
[8]Emergency Department, Royal Berkshire NHS Foundation Trust, Reading, UK
[9]Frimley Health NHS Foundation Trust, Frimley, UK
[10]GE Healthcare Diagnostic Imaging, Little Chalfont, UK
[11]GE Healthcare, Chicago, Illinois, USA
[12]GE Healthcare Ltd, Chicago, Illinois, USA
[13]University of Oxford Greyfriars, Oxford, UK
[14]University of Oxford, Oxford, UK

Correction notice Since this paper first published the author Hilal Johnson has been added to the author names. Oliva Gordon has been updated to Olivia Gamble and Melissa Keevil is now listed as Melissa Keevill.

X Alex Novak @AlexTNovak and Jasdeep Bahra @DrJazBahra

Collaborators Co-Authors - The Critical Care Suite Pneumothorax Reader Study Group: Dr Tom Duggan, Dr Melissa Keevil, Dr Oliva Gordon, Dr Osama Akram, Ms Elizabeth Belcher, Ms Rhona Taberham, Dr Rob Hallifax, Dr Jasdeep Bahra, Dr Abhishek Banerji, Dr Jon Bailey, Dr Antonia James, Dr Ali Ansaripour, Dr Nathan Spence, Dr John Wrightson, Dr Waqas Jarral, Steven Barry, Saher Bhatti, Ms Kerry Astley, Peter Aylward, Sharon Ghelman, Alec Baenen.

Contributors AN was responsible for the overall conduct of the project including analysis of results, independent write up and publication. Independent statistical analysis was performed by JO. GM, GWC and FG conducted ground truth reads, while the clinical reader group comprised TD, MK, OG, OA, EB, RT, RH, JB, AB, JB, AJ, AA, NS, JW, SB, SB, KA and WJ. CB was involved in study design, contract completion and steering group inputs. FG was involved in study design, ground truthing and steering group inputs. MB was involved in contract completion, data analysis, report generation and steering group inputs. AN, Emergency Medicine Research Oxford (EMROx), Oxford University Hospitals NHS Foundation Trust, is the guarantor.

Funding This work was supported by by the National Consortium of Intelligent Medical Imaging through the Industrial Strategy Challenge Fund (Innovate UK Grant 104688) and GE HealthCare.

Competing interests AS, SG and AB are employed by GE HealthCare, a key NCIMI stakeholder. AN and CB have undertaken paid consultancy work for GEHC. PA, SA and FG are employees of Report and Image Quality Control (www.raiqc.com), a spin-out company from Oxford University Hospitals NHS Foundation Trust.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants but as per Health Research Agency guidance (https://www.hra-decisiontools.org.uk/research/), this study constitutes Observational Research. Clinical data was used under existing ethical governance framework between National Consortium of Intelligent Medical Imaging and Oxford University Hospitals NHS Foundation Trust. The collection and aggregation of the retrospective dataset was undertaken under an approved data protection impact assessment undertaken by Oxford Hospitals NHS Foundation Trust. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. All datasets and documents related to this study currently reside securely in Oxford University Hospitals NHS Foundation Trust, and will be made available upon reasonable request to the corresponding author. The AI algorithm used in this research is a commercially available third-party product and as such the authors do not have sharing rights – enquires can be made via https://www.gehealthcare.com.

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
Alex Novak http://orcid.org/0000-0002-5880-8235
Abdala Trinidad Espinosa Morgado http://orcid.org/0000-0003-0967-3554
Jasdeep Bahra http://orcid.org/0000-0001-9503-5510
Jon Bailey http://orcid.org/0000-0001-9752-0712

REFERENCES

1 Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med* 2020;3:118.
2 van Leeuwen KG, Schalekamp S, Rutten MJCM, *et al*. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol* 2021;31:3797–804.
3 Oren O, Gersh BJ, Bhatt DL. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. *Lancet Digit Health* 2020;2:e486–8.
4 Kelly BS, Judge C, Bollard SM, *et al*. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). *Eur Radiol* 2022;32:7998–8007.
5 Ahn JS, Ebrahimian S, McDermott S, *et al*. Association of artificial intelligence-aided chest radiograph interpretation with reader performance and efficiency. *JAMA Netw Open* 2022;5:e2229289.

6 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.
7 Guermazi A, Tannoury C, Kompel AJ, *et al*. Improving radiographic fracture recognition performance and efficiency using artificial intelligence. *Radiology* 2022;302:627–36.
8 Jin KN, Kim EY, Kim YJ, *et al*. Diagnostic effect of artificial intelligence solution for referable thoracic abnormalities on chest radiography: a multicenter respiratory outpatient diagnostic cohort study. *Eur Radiol* 2022;32:3469–79.
9 Hillis JM, Bizzo BC, Mercaldo S, *et al*. Evaluation of an artificial intelligence model for detection of pneumothorax and tension pneumothorax in chest radiographs. *JAMA Netw Open* 2022;5:e2247172.
10 Nagendran M, Chen Y, Lovejoy CA, *et al*. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;368:m689.
11 Warman P, Warman A, Warman R, *et al*. Using an artificial intelligence software improves emergency medicine physician intracranial haemorrhage detection to radiologist levels. *Emerg Med J* 2024;41:298–303.
12 National Institute for Health and Care Excellence (NICE). Artificial intelligence (AI)-derived software to help clinical decision making in stroke (DG57), 2024. Available: https://www.nice.org.uk/guidance/dg57/resources/artificial-intelligence-aiderived-software-to-help-clinical-decision-making-in-stroke-pdf-1053876693445 [Accessed 14 May 2024].
13 Unsworth H, Wolfram V, Dillon B, *et al*. Building an evidence standards framework for artificial intelligence-enabled digital health technologies. *Lancet Digit Health* 2022;4:e216–7.
14 Ibrahim H, Liu X, Rivera SC, *et al*. Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines. *Trials* 2021;22:11.
15 Vasey B, Nagendran M, Campbell B, *et al*. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nat Med* 2022;28:924–33.
16 RCoE Medicine. RCEM position statement on artificial intelligence. 2022.
17 Ramlakhan SL, Saatchi R, Sabir L, *et al*. Building artificial intelligence and machine learning models: a primer for emergency physicians. *Emerg Med J* 2022;39:e1.
18 GE Health care. Available: https://www.gehealthcare.com/products/radiography/critical-care-suite [Accessed 28 Feb 2023].
19 Bossuyt PM, Reitsma JB, Bruns DE, *et al*. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015;351:h5527.
20 Sounderajah V, Ashrafian H, Golub RM, *et al*. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. *BMJ Open* 2021;11:e047709.
21 Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations. *Commun Stat Simul Comput* 1995;24:285–308.
22 Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the Jackknife method. *Invest Radiol* 1992;27:723–31.
23 Hillis SL, Obuchowski NA, Schartz KM, *et al*. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. *Stat Med* 2005;24:1579–607.
24 Hillis SL, Schartz KM. Multireader sample size program for diagnostic studies: demonstration and methodology. *J Med Imag* 2018;5:1.
25 Yoo H, Kim EY, Kim H, *et al*. Artificial intelligence-based identification of normal chest radiographs: a simulation study in a multicenter health screening cohort. *Korean J Radiol* 2022;23:1009–18.
26 Duron L, Ducarouge A, Gillibert A, *et al*. Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: a multicenter cross-sectional diagnostic study. *Radiology* 2021;300:120–9.
27 Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. *JAMA* 2017;318:517–8.
28 Challen R, Denny J, Pitt M, *et al*. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 2019;28:231–7.
29 Floridi L, Cowls J, Beltrametti M, *et al*. Ai4People-an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach (Dordr)* 2018;28:689–707.
30 Homayounieh F, Digumarthy S, Ebrahimian S, *et al*. An artificial intelligence-based chest X-ray model on human nodule detection accuracy from a multicenter study. *JAMA Netw Open* 2021;4:e2141096.
31 Dendumrongsup T, Plumb AA, Halligan S, *et al*. Multi-reader multi-case studies using the area under the receiver operator characteristic curve as a measure of diagnostic accuracy: systematic review with a focus on quality of data reporting. *PLoS ONE* 2014;9:e116018.