BMJ Digital Health & AI

Artificial intelligence-assisted reader evaluation in acute CT head interpretation (AI-REACT): a multireader multicase study

Alex Novak,[1,2] Ruchir Shah,[1,3] Abdala T Espinosa Morgado,[1,3] Dennis Robert,[4] Shamie Kumar,[5] Jason Oke,[6] Kanika Bhatia,[7] Andrea Romsauerova,[7] Tilak Das,[8] The AI-REACT Reader Study Group,[7] Mariapaola Narbone,[9] Rahul Dharmadhikari,[10] Mark Harrison,[11] Kavitha Vimalesvaran,[12] Jane Gooch,[13] Nick Woznitza,[14,15] David Lowe,[16,17] Haris Shuaib,[9] Sarim Ather[1,3]

To cite: Novak A, Shah R, Espinosa Morgado AT, *et al*. Artificial intelligence-assisted reader evaluation in acute CT head interpretation (AI-REACT): a multireader multicase study. *BMJ Digit Health* 2026;**2**:f000071. doi:10.1136/bmjdh-2026-000071

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjdh-2026-000071).

Received 30 June 2025
Accepted 3 February 2026

For numbered affiliations see end of article.

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

ABSTRACT

Objective To assess whether an artificial intelligence (AI) tool improves the accuracy, speed and confidence of general radiologists, emergency clinicians and radiographers in detecting critical non-contrast CT head (NCCTH) abnormalities and to evaluate its stand-alone performance and factors influencing diagnostic accuracy.

Methods and analysis A retrospective dataset of 150 NCCTH (52 normal and 98 with critical abnormalities) was reviewed by 30 readers (10 radiologists, 15 emergency clinicians and 5 radiographers) from four National Health Service trusts. Each interpreted scan is performed unaided and then with the qER EU 2.0 AI tool, separated by a 2-week washout period. Ground truth was established by two neuroradiologists. We measured the AI's stand-alone performance and its effect on reader accuracy, confidence and speed.

Results The qER algorithm showed strong diagnostic performance (area under the receiver operator curve 0.821–0.976). With AI, pooled reader sensitivity for critical abnormalities increased from 82.8% to 89.7% (+6.9%, p<0.001) and for intracranial haemorrhage from 84.6% to 91.6% (+7.0%, p<0.001), while specificity decreased from 84.5% to 78.9% (–5.5%, p=0.046). Reader confidence did not change significantly. Emergency department (ED) clinicians with AI achieved sensitivity similar to unaided radiologists.

Conclusion AI assistance increased sensitivity for detecting critical abnormalities on NCCTH but reduced specificity. AI-enabled ED clinicians to achieve diagnostic sensitivity comparable to radiologists, supporting its potential to enhance non-radiologist performance. Further studies are needed to confirm these findings in clinical practice.

Trial registration number NCT06018545.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Artificial intelligence (AI) algorithms for detecting acute findings on non-contrast CT head scans have shown strong diagnostic performance in retrospective datasets. Prior studies demonstrated that AI assistance can improve radiologists' interpretation accuracy under controlled conditions. However, its potential to support non-radiologist clinicians, such as emergency physicians or radiographers, in real-world settings has not yet been evaluated.

WHAT THIS STUDY ADDS

⇒ This multicase, multireader study shows that AI-assisted interpretation significantly improves emergency physicians' diagnostic accuracy for CT head scans, reaching performance levels comparable to general radiologists.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ These findings suggest that AI-assisted image interpretation could support safer and more accurate CT head scan assessment by non-radiologist clinicians. This may help streamline decision-making in emergency settings, though prospective clinical evaluations are needed before implementation in practice.

INTRODUCTION

CT of the head is the most common cross-sectional imaging modality performed in the emergency department (ED), with over 1 million scans requested annually in the UK.[1] Increased ED attendances combined with improved CT availability and a lower clinical threshold for scanning have resulted in a significantly greater number of patients with lower acuity receiving non-contrast CT head (NCCTH) scans, most of whom will be ultimately discharged from the ED without admission.[2] Typically, however, ED clinicians are dependent on the NCCTH report from radiologists before making a clinical decision based on their findings.[3] This requirement

places a huge demand on radiology services, which are often unable to provide timely reports, leading to delays in treatment, referral and discharge decisions, ultimately impacting patient care and departmental efficiency and reducing patient flow through the ED.[4]

A number of AI-assisted image interpretation algorithms have been developed to assist clinicians in the identification of pathological findings on NCCTH scans.[5 6] Several of these are CE-marked and Food and Drug Administration-approved for clinical use and are already being deployed in hospitals worldwide. Retrospective studies have indicated high accuracy in some cases, and reader studies have demonstrated the potential for AI-assisted NCCTH interpretation to improve radiologist accuracy.[7] To date, however, evidence of efficacy and impact has been focused primarily on the performance of radiologists, and few studies have explored the potential for AI-assisted image interpretation in evaluating the impact of AI assistance with other groups of healthcare professionals who regularly review or act on NCCTH interpretations, such as ED clinicians and radiographers.[8 9]

qER 2.0 EU is a CE (European Conformity) Class IIb AI tool for the interpretation of NCCTH, which was developed using 300 000 retrospectively collected and labelled scans from 31 imaging centres in India and one of the largest teleradiology centres in the USA, including scans obtained from both in-hospital and outpatient radiology settings.[10] It can detect, classify and localise intracranial haemorrhage, hypodensities suggestive of infarct, mass effect, midline shift, atrophy and skull fractures in NCCTH. If any of the target abnormalities is detected by the software, the tool provides the user with a single summary listing all the target abnormalities found by qER on the CT, followed by all slices in the scan with the overlay highlighting the location of the abnormalities (figure 1a). Alternatively, if none of the target abnormalities are detected, the output will indicate that the software has analysed the image and identified no target abnormalities.[11]

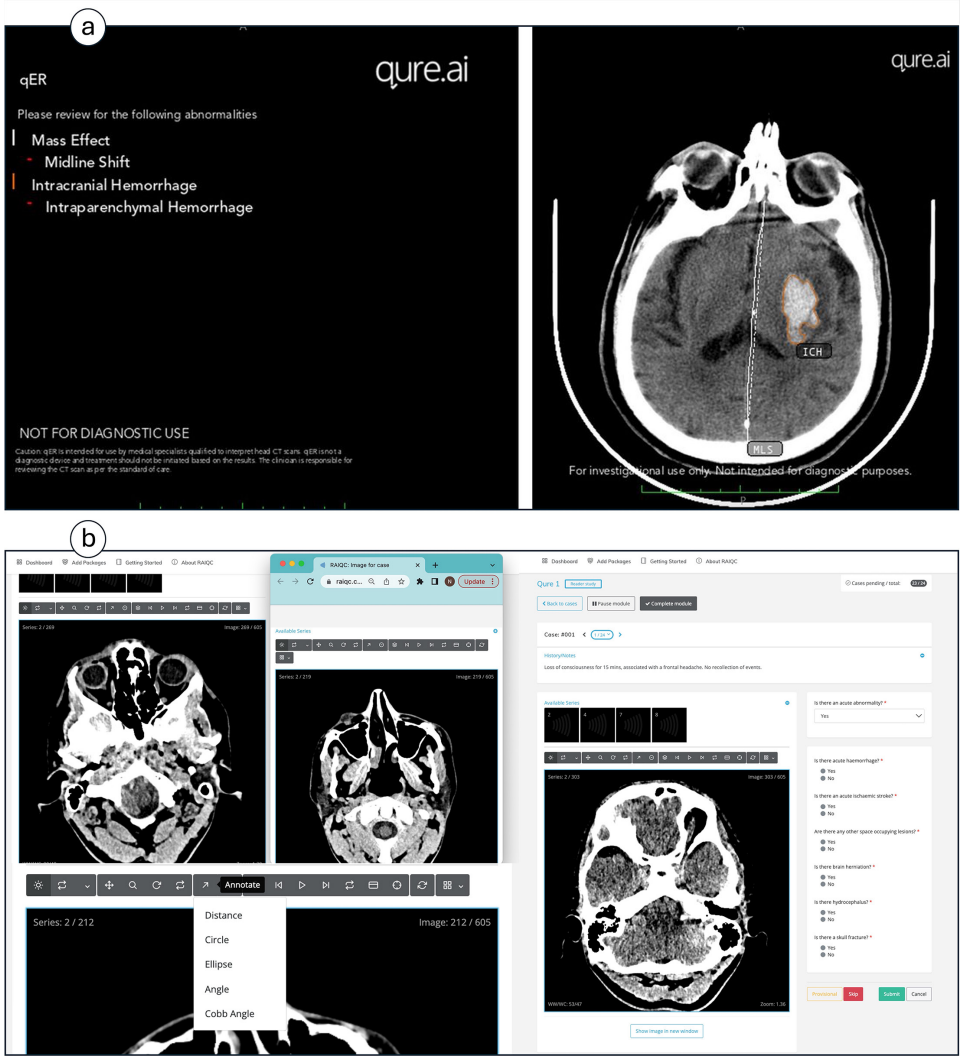

Figure 1 (a) qER presents a summary of all abnormalities identified and the slice images containing those abnormalities with a localisation overlay. (b) Screenshots illustrating the reader study online module interface used by clinical participants. These images depict how readers recorded their findings and confidence levels for each identification. ICH, intracranial haemorrhage

qER is intended to support certified radiologists and/or licensed medical practitioners for clinical decision-making. It is a support tool which, when used with original scans, may assist the clinician to improve efficiency, accuracy and turnaround time in reading NCCTH. Multiple prior studies have reported the standalone diagnostic accuracy of qER; however, as yet, its potential impact on the diagnostic accuracy of general radiologists, radiographers and ED clinicians has not been fully tested.[10 12–14] Here we evaluate a deep-learning model designed to assist clinicians in the interpretation of NCCTH images, using clinicians and radiologists of varying seniority and experience to evaluate an image dataset which includes a broad range of pathologies commonly encountered in routine emergency care.

Study aims

1. To determine the improvement in NCCTH image interpretation accuracy of general radiologists, ED clinicians and radiographers in detecting critical abnormalities (any one or more of intracranial haemorrhage, midline shift, mass effect, skull fracture or hypodensity suggestive of infarct) with the assistance of the qER AI tool (primary).
2. To determine the stand-alone accuracy of qER for the detection of intracranial haemorrhage, hypodensity suggestive of infarct, midline shift, mass effect and skull fractures (secondary).
3. To measure the time taken by the above clinicians to evaluate scan images and their diagnostic confidence, with and without the AI tool (secondary).
4. To explore which imaging factors influence clinicians' reporting accuracy and efficiency and algorithm performance, for example, category of abnormality, difficulty of image interpretation, clinician seniority and professional group (secondary).

METHODS

Study design and participants

We undertook a fully crossed paired multireader multicase (MRMC) study as per our previously published study protocol.[15] 150 NCCTH scans of ED patients aged 18 years or above were retrospectively identified by the clinical and Picture Archiving and Communication System/Information Technology team by searching the Radiology Information System at Oxford University Hospitals National Health Service (NHS) Foundation Trust (see online supplemental figure S1 in supplementary material). The dataset was selected using stratified consecutive sampling of existing clinical radiology reports to contain 60 control scans and 90 abnormal scans, including a minimum of 10 scans containing each of the following nine defined 'critical abnormalities' as defined by the AI tool in question: extradural haemorrhage, subdural haemorrhage, subarachnoid haemorrhage, intraparenchymal haemorrhage, intraventricular haemorrhage, hypodensity suggestive of infarct, midline shift, mass effect and skull fractures.

A summary of the process used to create the dataset is presented in online supplemental figure S2. For the purposes of case selection, the existing clinical radiology reports were used to determine whether a given scan contains an abnormality of interest. To reduce selection bias, consecutive scans were reviewed, and all scans which fittedthat fit the inclusion and exclusion criteria were included until the specific case number totals were reached. Each case was derived from a different patient. In addition to the prespecified pathology subgroups, image inclusion criteria are summarised as follows:

Inclusion criteria

- Individuals undergoing NCCTH in the ED.
- Age ≥18 years.
- Non-contrast axial CT scan series with consistently spaced axial slices.
- Soft reconstruction kernel covering the complete brain.
- Maximum slice thickness of 6 mm.

Exclusion criteria

The following features are known to cause inaccurate outputs from the qER AI:

- Scans with obvious postoperative defects, or from patients who previously underwent brain surgery.
- Scans with artefacts such as burr holes, shunts or clips.
- Scans containing metal artefacts.

Establishing a reference standard

To establish a reference standard, two consultant neuroradiologists independently reviewed the CT images in the dataset and recorded the presence or absence of the nine target abnormalities for each scan. In the case of disagreement, a third senior neuroradiologist's opinion was sought for arbitration. In line with previous similar studies, a difficulty score was assigned to each pathological finding by the two ground truthers using a five-point Likert scale, and where there was disagreement, the mean score was taken.[16 17]

Image inferencing

All images were inferred by the qER algorithm, and the resulting secondary capture images were stored as a separate dataset. The qER output was presented as an additional series with a notification to suggest the presence or absence of a target abnormality as the first image of the series, and the segmentation of the abnormal areas identified was overlaid on the scan images.

Reader participants

30 readers were recruited from the following four hospital trusts: Guy's and St Thomas NHS Foundation Trust, Northumbria Healthcare NHS Foundation Trust, NHS Greater Glasgow and Clyde and Oxford University Hospitals NHS Foundation Trust. The composition and inclusion/exclusion criteria for the readers are summarised in table 1.

Table 1 Reader characteristics

Professional group	Seniority/subgroup	Number of readers
Emergency medicine	Consultants	5
	Registrars (ST3-6)	5
	Juniors (F1-ST2)	5
General radiologists	Consultants	5
	Registrars (ST3-6)	5
Radiographers	CT Radiographers	5
Total		30

Reader phases 1 and 2

Reader recruitment was undertaken by each principal investigator for their own site via email and in person. Five modules, each containing 30 cases, were created and uploaded to a secure online Digital Imaging and Communications in Medicine (DICOM) viewer (www.RAIQC.com). All 30 readers undertook a brief training module, including five practice cases to familiarise themselves with the platform and study requirements, then proceeded to review all 150 cases over a 4-week period using a laptop or personal computer. For each scan, the readers recorded whether they identified any of the nine critical abnormalities as being present, providing a confidence rating for each of their findings on a 10-point Likert scale. The time taken for each scan interpretation was automatically recorded. The order of the cases was randomised for each reader at each phase, and readers were blinded to the number of abnormal and normal cases in the study.

In the first phase, all readers reviewed all 150 scans, blinded to the ground truth and without AI assistance. Following a 2-week washout period to mitigate recall bias, readers undertook the second phase, where they reviewed all scans again in a randomised order, remaining blinded to the ground truth, but this time with access to the results from the qER tool. Screenshots from the RAIQC modules, as presented to the readers, are included in figure 1b. The complete study workflow is summarised in figure 2.

Outcome measures

The primary outcome measure was the diagnostic accuracy for the detection of critical findings at the case level, quantified by the difference in the area under the receiver operator curve (AUC) of readers in identifying a scan as containing one or more critical findings (intracranial haemorrhage subcategorised into five subtypes, hypodensity suggestive of infarct, with and without AI assistance). Secondary outcome measures included differences in reader sensitivity, specificity and confidence with and without AI assistance, the stand-alone diagnostic performance of the qER AI versus ground truth for each target abnormality, and median scan interpretation time.

Sample size and power calculation

Using the Hillis and Berbaum method for multireader multicase (MRMC) power analysis, a sample of 30 readers and a minimum of 135 scans (82 with the presence of critical findings and 53 with no critical findings) was estimated to have a minimum 80% power at a type I error rate of 5% to detect a minimum difference in readers' AUC of 5%, assuming a large inter-reader and intrareader variability of 0.3 and 0.05, respectively; a 0.35 conservative correlation between readers, and an anticipated average readers' AUC of 0.75, guided by previous literature.[18 19]

Statistical analyses

The stand-alone performance of the qER algorithm was compared with the ground truth generated by the neuroradiologists, using the continuous probability score from the algorithm for the AUC analyses and binary classification results for the evaluation of sensitivity, specificity, positive predictive value and negative predictive value.

The difference in AUC of readers with and without AI was tested based on the Obuchowski-Rockette model for MRMC analysis, which models the data using a two-way mixed effects analysis of variance model treating readers and cases (images) as random effects and the effect of AI as a fixed effect with the recommended adjustment to df by Hillis *et al*. Sensitivity and specificity were analysed as part of this model. The main analysis was performed as a single pool including all groups and sites. Prespecified subgroup analyses were performed for the following variables: professional group (radiologist vs ED clinician vs radiographer), postgraduation experience level (junior <5 years, middle grade 5–10 years and senior >10 years), pathological finding and difficulty of image.[18–20]

The median review time per scan with versus without AI was compared using a non-parametric Wilcoxon signed-rank test. Statistical analyses were all performed using R software (V.4.0.2; R Foundation for Statistical Computing). The significance threshold was set at a two-sided 5% (p=0.05) for all secondary analyses.

Patient and public involvement

This study was presented to the Oxford ACUTECare PPI group, which supported the study and its aims and influenced design, data management and dissemination strategies.

Ethics and dissemination

The study is registered at Clinicaltrials.gov (NCT06018545) and the ISRCTN registry (ISRCTN17560291).

RESULTS

Baseline characteristics

Baseline characteristics are summarised in online supplemental table S1. Of the 150 images in the dataset, 98 were defined by the radiologist panel as containing one or more critical abnormalities. 30 readers (10 ED clinicians, 10 general radiologists and 5 radiographers) within the

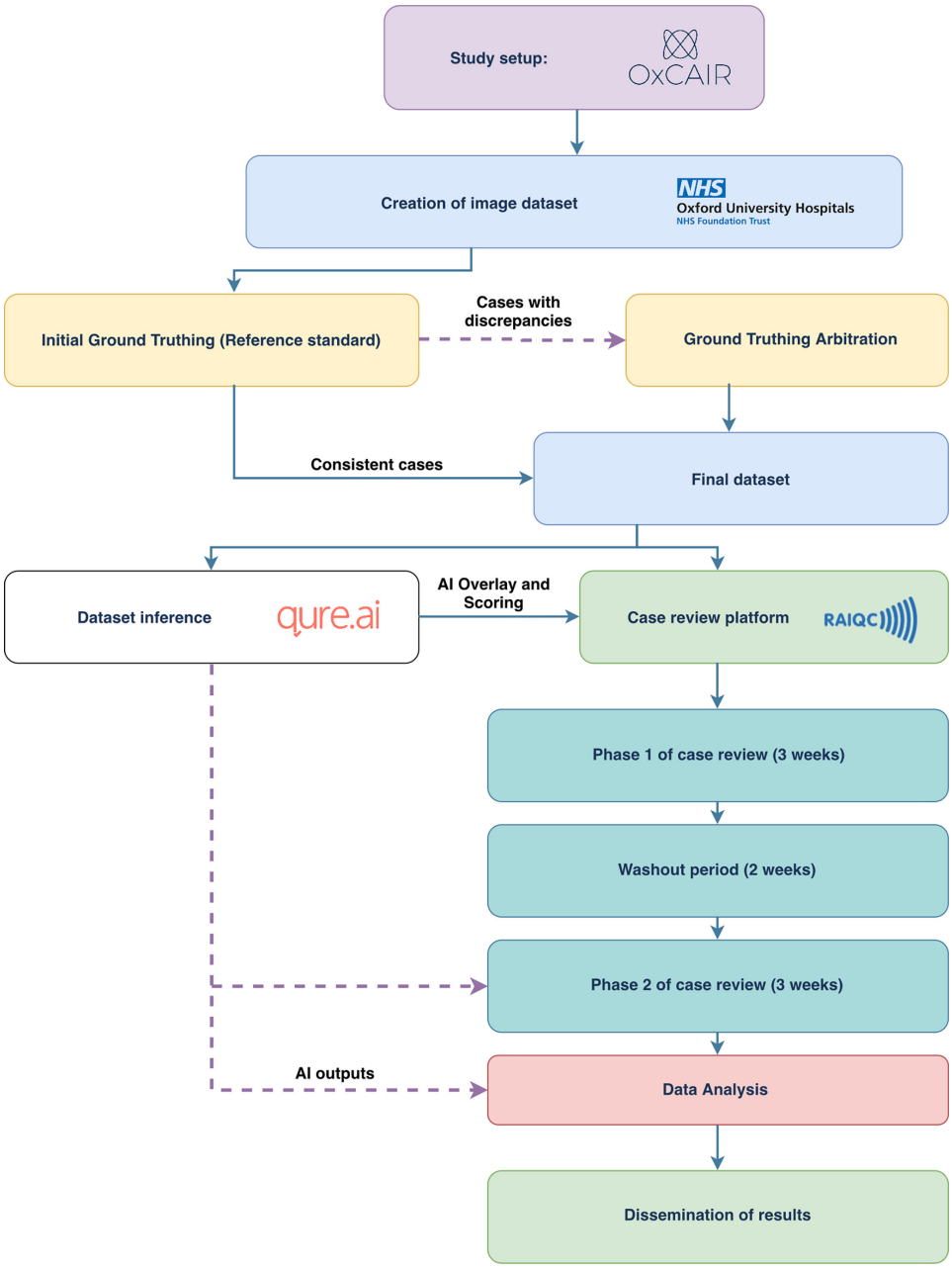

Figure 2 Study and data flowchart.

prespecified seniority classifications each interpreted all 150 cases both with and without AI, totalling 9000 individual interpretations for analysis. No cases were rejected due to artefacts, and all were processed successfully by the algorithm.

Algorithm versus ground truth

Retrospective analysis of the diagnostic performance of the qER algorithm versus ground truth is presented in online supplemental table S2 and figure 3. The algorithm showed strong overall diagnostic performance for all abnormality subgroups with AUCs ranging from 0.821 (95% CI 0.740 to 0.903) to 0.986 (95% CI 0.969 to 1.000), with the exception of mass effect, for which it showed poor discriminative ability in this study with an AUC of 0.604 (95% CI 0.435 to 0.774). Lower sensitivities (<0.80) were seen for extradural haemorrhage (0.692, 95% CI 0.613 to 0.766), intraparenchymal haemorrhage (0.576, 95% CI 0.490 to 0.654), intraventricular haemorrhage (0.650, 95% CI 0.571 to 0.729), mass effect (0.286, 95% CI 0.216 to 0.366) and fracture (0.727, 95% CI 0.648 to 0.796); lower specificities (<0.80) were seen for infarct (0.699, 95% CI 0.620 to 0.772) and mass effect (0.727, 95% CI 0.648 to 0.796).

Overall reader performance

Changes in overall (pooled) reader performance for different pathological subgroups are presented in online supplemental table S3 and figure 4. The primary outcome measure of diagnostic performance for the detection of critical abnormality, measured by AUC, showed no statistically significant change; however, a

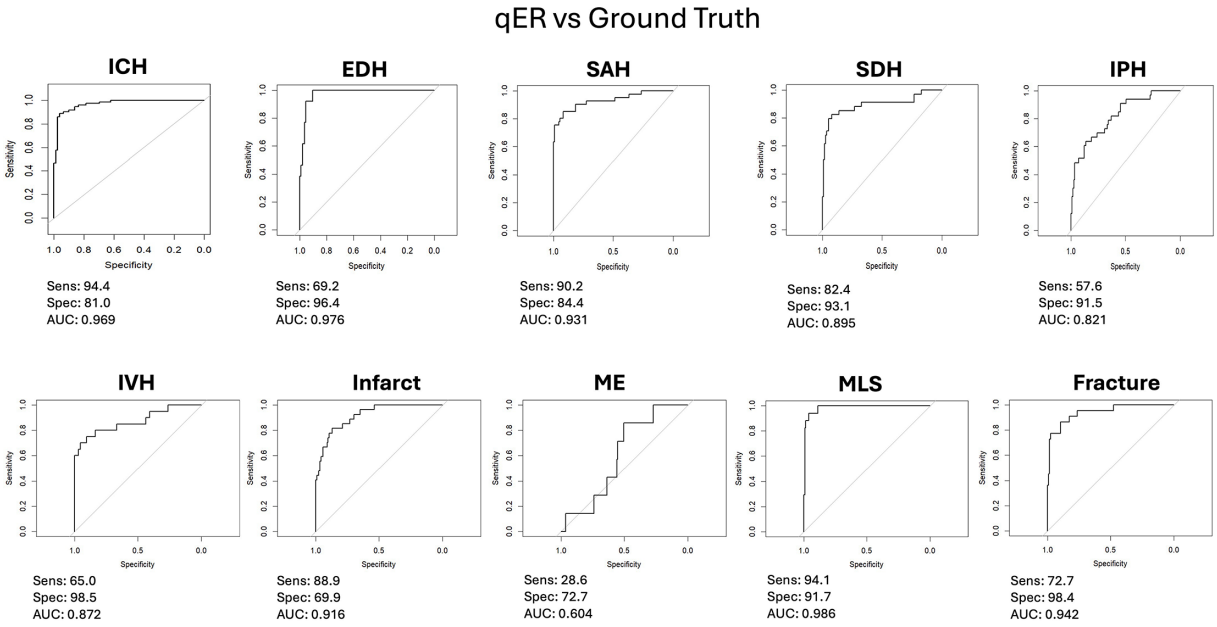

Figure 3 Diagnostic performance of the algorithm versus ground truth depicted as AUC for different pathological subgroups, including ICH, EDH, SAH, SDH, IPH, IVH, infarct, ME, MLS and fracture. Sens and Spec are given as per the default threshold settings for the algorithm for each pathological subgroup. AUC, area under the receiver operator curve; EDH, extradural haemorrhage; ICH, intracranial haemorrhage; IPH, intraparenchymal haemorrhage; IVH, intraventricular haemorrhage; ME, mass effect; MLS, midline shift; SAH, subarachnoid haemorrhage; SDH, subdural haemorrhage; Sens, sensitivity; Spec, specificity.

statistically significant increase in pooled sensitivity for critical abnormality was observed from 82.8% to 89.8% (difference +7.0%, 95% CI 3.4% to 10.6%, p<0.001). This was accompanied, however, by a corresponding decrease in specificity from 84.5% to 78.9% (difference −5.5%, 95% CI −0.09% to 11.0%, p=0.046). Statistically significant increases were demonstrated for sensitivity across all main pathology subgroups and for AUC in intracranial haemorrhage (0.853–0.956, difference +0.104, 95% CI 0.0602 to 0.147, p<0.001), infarct (0.782–0.846, difference +0.0636, 95% CI 0.0211 to 0.106, p=0.0035) and fracture (0.845–0.902, difference +0.0571, 95% CI 0.0162 to 0.098, p=0.007) subgroups, with a decrease in specificity seen in the midline shift subgroup from 97.4% to 94.2% (difference −3.2%, 95% CI −0.94% to 5.5%, p=0.001) (see online supplemental tables S5–S11).

Subgroup analyses

Relative performance for different reader subgroups in detecting critically abnormal scans is summarised in

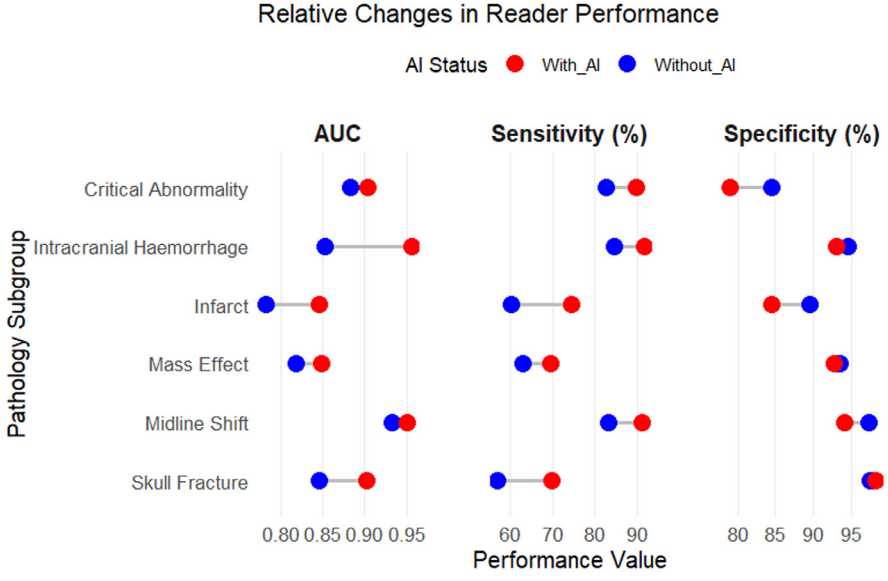

Figure 4 Relative changes in pooled reader performance measured in terms of AUC, sensitivity and specificity for different pathology subgroups. AI, artificial intelligence; AUC, area under the receiver operator curve.

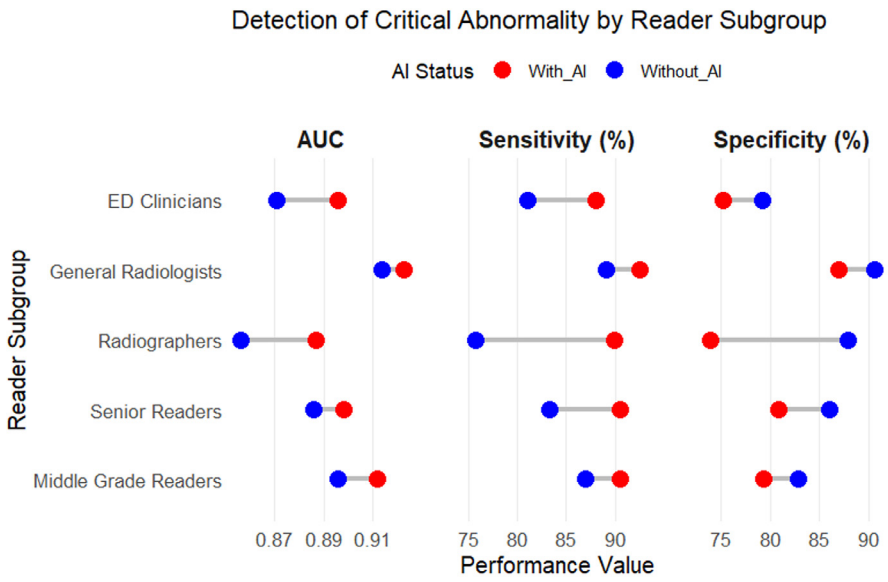

Figure 5 Relative reader subgroup performance for abnormality and ICH, infarct, mass effect and fracture. AI, artificial intelligence; AUC, area under the receiver operator curve; ED, emergency department.

online supplemental table S4 and figure 5. Subgroup analyses for emergency medicine doctors demonstrated no significant change in AUC but a large increase in sensitivity for the detection of critically abnormal scans (81%–88%, difference +6.9%, 95% CI 0.028% to 11.1%, p=0.001), with increases in sensitivity seen across all pathology subgroups (see online supplemental materials) and increases in AUC for ICH, infarct and fracture subgroups; however, a decrease in specificity was observed in the infarct subgroup. General radiologists showed no significant change in ability to detect critical abnormality with AI but demonstrated a large increase in AUC and small increase in sensitivity for intracranial haemorrhage from 84.2% (95% CI 75.1% to 93.2%) to 97.8% (95% CI 96.0% to 99.6%, p=0.007) and 93.9% (95% CI 89.9 to 98.0) to 95.4% (95% CI 91.7% to 99.0%, p=0.014) respectively (see online supplemental table S6). Radiographers demonstrated increases in the AUC for intracranial haemorrhage, infarct and mass effect subgroups, with the latter two also showing an increase in sensitivity. Senior readers demonstrated a significant increase in sensitivity for critical abnormality (83.3%–90.5%, difference +7.2, 95% CI 2.3% to 12.1%); otherwise, no statistically significant difference in critical abnormality detection was seen across seniority subgroups.

Interpretation time

Mean interpretation time was found to be 192s per case in phase 1 (without AI) and reduced to 173s per case in phase 2 (with AI) (p<0.001). Relevant data were winsorised at a 95% threshold to exclude extreme outliers.

DISCUSSION

This study evaluated the impact of an AI-assisted image interpretation on the diagnostic performance of a group of radiologists, radiographers and emergency medicine clinicians routinely involved in the care of patients undergoing NCCTH. Key findings included (1) a strong overall diagnostic performance of the algorithm measured as AUC against an enhanced 'ground truth' reference standard of senior neuroradiologist reporting for 9 out of 10 pathological subgroups, (2) a significant increase in pooled reader sensitivity for the detection of critical abnormality with AI assistance coupled with a comparable decrease in specificity; however, no statistically significant change in AUC and (3) a marked increase in the sensitivity of emergency medicine clinicians for the detection of critical abnormality with AI, comparable to the unaided performance of general radiologists. No difference in effect was seen across different seniority subgroups with AI assistance.

The qER algorithm has previously been retrospectively tested on an external validation data set of 491 NCCTH scans from patients in India. In that study, the AUC for ICH, skull fracture, midline shift and mass effect was determined as 0.941 (95% CI 0.919 to 0.965), 0.962 (95% CI 0.92 to 1.0), 0.970 (95% CI 0.94 to 0.999) and 0.922 (95% CI 0.888 to 0.955), respectively. In a subsequent Swedish stroke registry study, qER was found to have 97% sensitivity in detecting non-traumatic ICH. This study demonstrated similar performance characteristics in most respects; however, low sensitivities of qER were seen for intraparenchymal haemorrhage (0.58), mass effect (0.29) and intraventricular haemorrhage (0.65) and low specificities (<0.80) were seen for infarct (0.70) and mass effect (0.73). This serves as an important reminder of variability in the performance of AI-assisted image interpretation algorithms on a case-by-case basis. Readers in this study showed no improvement in sensitivity for detection of intraparenchymal or intraventricular haemorrhage (see online supplemental table S5), a

decrease in sensitivity for midline shift (from a high baseline) and a significant decrease in specificity for infarct detection with AI assistance, suggesting that limitations in algorithm performance can potentially translate to an adverse effect on the performance of human readers in some use cases, which is of critical importance in considering options for real-world deployment and the need to inform readers regarding the reliability and relative 'confidence' of the algorithm output for a given finding.

There are no prior reports on the impact of qER assistance on readers; hence, the AI-assisted reader evaluation in acute CT head interpretation study is the first to investigate this effect. Few clinical reader studies have been undertaken for any NCCTH interpretation algorithm.[21] Of note, a 2023 paper by Buchlak *et al* evaluated the impact of a deep learning algorithm for 22 pathological findings (containing 192 subcategory findings) for NCCTH on the reporting accuracy of 32 radiologists. Those findings, which demonstrated an AUC >0.80 (n=144, average AUC 0.93), were then incorporated into an MRMC, in which assisted and unassisted radiologists demonstrated an average AUC of 0.79 and 0.73 across 22 grouped parent findings and 0.72 and 0.68 across 189 child findings, respectively. When assisted by the model, radiologist AUC was significantly improved for 91 findings, and reading time was significantly reduced. The average algorithm AUC obtained in that analysis was comparable with those measured for qER in this study; however, the markedly lower AUC for skilled radiologist readers both with and without AI assistance implies significant differences between both the reader group and datasets, which in the Buchlak study were randomly selected rather than consecutively derived from clinical practice, and hence caution should therefore be exercised in comparing the results.

In our study, AI assistance significantly increased reader sensitivity (82.8%–89.8%, p<0.001) while the AUC remained unchanged (p=0.076), indicating a shift in the readers' operating threshold rather than an improvement in overall discriminative ability. In signal detection terms, readers detected more true positives at the cost of more false positives, effectively moving their operating point along the same receiver operator curve rather than shifting the curve upwards. This implies that clinicians may have used the AI overlay primarily as a 'high-sensitivity alert', prompting them to re-scrutinise areas they might have missed, thereby catching more subtle pathologies. However, they were unable to sufficiently dismiss false-positive AI marks, leading to the concurrent drop in specificity.

While several evaluations of AI-assisted image interpretation have indicated a positive impact on radiologist performance, in this study, AI assistance showed limited potential to improve even non-specialist radiologists when applied to a dataset which used consecutive cases and was therefore derived more closely from routine clinical practice. This should temper expectations and assumptions regarding the impact of AI-assisted image interpretation in this context, that is, skilled radiologist reporting, though this finding may reflect a lower number of 'difficult' cases in the dataset used in this study and the removal of factors such as distraction and fatigue, which may impair radiologist performance in real-world settings. Furthermore, the decreases in pooled reader specificity for the detection of critical abnormality and certain pathology subgroups indicate the potential for AI assistance to adversely affect reader performance in some circumstances. Equally, algorithms optimised for sensitivity may support non-specialist readers, but bias may lead to 'overcall', which needs to be taken into account when considering potential roles for AI assistance. AI did not significantly improve non-specialist radiographer accuracy to the same levels as the other specialty groups, indicating that a priori reader skill remains important in assisted accuracy and that AI assistance alone is unlikely to replace experience in this context. Conversely, the lack of difference between seniority subgroups suggests that clinical experience is not necessarily an indicator of skill in interpreting NCCTH images in the context of evaluations such as this. Assisted image interpretation AI improved the diagnostic performance of ED clinicians to levels comparable with that of unaided general radiologists, who represent a pragmatic benchmark for current radiological clinical practice. This suggests that it may be possible to identify subgroups of patients on a clinical basis for whom the interpretation of ED clinicians may be safe and effective enough to allow clinical actions to be taken prior to the availability of a radiological report. Future studies should evaluate this potential on a prospective clinical basis and should explore the optimisation of algorithm threshold and calibration to increase negative predictive value and facilitate the reliable identification of 'normal' scans to facilitate early ED discharge.

This study was designed to reflect current trends in the methodology of assessing clinicians and to facilitate comparison with other studies reporting similar evaluations, and as such, chose AUC as a primary outcome measure of reader accuracy, using self-reported reader confidence as a variable performance metric. This is useful in demonstrating changes in performance between paired unassisted and assisted groups, though it can be misleading to use this for cross-comparison between different reader subgroups or between readers and the algorithm. While AUC is useful to understand the potential of an AI algorithm to accurately identify pathologies and to determine the optimum operating thresholds for reporting pathological findings as present or absent, in clinical practice, the need for a specific predetermined threshold renders this metric misleading in terms of clinical impact, as algorithms which demonstrate a high AUC overall may still have relatively low sensitivity or specificity at default operating thresholds, which may limit clinical applicability; hence, the need to report and consider all performance metrics in evaluating these technologies.

Strengths

This study evaluated the impact of the AI tool on diagnostic accuracy, speed and confidence in its most realistic use case, as an assistant to healthcare professionals rather than in isolation. It represents the first UK-based multicentre validation of an AI for NCCTH scans trained on a large data set (300 000 head CTs). The dataset constructed used a systematic approach to collect consecutive cases derived from routine clinical datasets, increasing the validity and generalisability of results compared with other studies, which have used more subjective and less transparent approaches to creating image datasets.

The reader group itself is large (n=30) compared with other multicase multireader studies.[22] Five readers represent a typical minimum group size for such studies, so each reader specialty subgroup in our study included at least five readers, allowing for independent subgroup analyses.[19] Nevertheless, variation in reader performance occurs on an individual basis, which may limit the generalisability of findings. The reader group includes non-radiologists (emergency medicine clinicians and radiographers) among the healthcare professionals who may benefit from AI assistance. This allows the potential utility of AI-assisted NCCTH interpretation to be explored in use cases other than supporting the diagnostic performance of radiologists.

Limitations

This was an online study using a curated dataset with an artificially high prevalence of abnormal images in the selected scans, which was enriched in order to achieve statistical power to detect the impact of AI assistance. Although necessary to facilitate an important evaluation of diagnostic accuracy, this limits the immediate generalisability of results to real-life clinical performance. Scans with postoperative changes and significant artefacts (eg, patient movement) fall outside the AI's scope of training and were excluded from the study. Clinical data was limited to the clinical vignette on the request form—this reflects real-world practice for radiologists, but clinician readers would not have been able to judge who was high risk for the presence of pathology, for example, evident clinically significant injury/presentation, which may have informed their diagnostic decisions when interpreting the scans in a real-life clinical context.

CONCLUSION

Use of AI-assisted image interpretation for NCCTH significantly increased the pooled sensitivity of a group of radiologist and clinician readers in detecting critical abnormalities; however, this was accompanied by a comparable decrease in specificity. Subgroup analysis showed limited benefit to skilled radiologist readers but demonstrated a significant increase in the sensitivity of ED clinicians in detecting abnormality to a level comparable to that of unaided radiologists. These findings should be fully explored prospectively to validate these results and identify potential use cases for this application.

Author affiliations
[1]Oxford Clinical Artificial Intelligence Research (OxCAIR), Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[2]Emergency Medicine Research Oxford, Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[3]Oxford Clinical Artificial Intelligence Research (OxCAIR), Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[4]Qure.ai, Bangalore, India
[5]Qure.ai Technologies Limited, London, UK
[6]Department of Primary Health Care Sciences, University of Oxford, Oxford, UK
[7]Oxford University Hospitals NHS Foundation Trust, Oxford, England, UK
[8]Department of Clinical Radiology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
[9]Guy's and St Thomas' NHS Foundation Trust, London, UK
[10]Northumbria Healthcare NHS Foundation Trust, North Shields, UK
[11]Emergency Department, Northumbria Specialist Emergency Care Hospital, Cramlington, UK
[12]Guy's and St Thomas' NHS Foundation Trust, London, UK
[13]College of Health, Psychology & Social Care, University of Derby, Derby, UK
[14]University College London NHS Foundation Trust, London, UK
[15]School of Allied and Public Health Professions, Canterbury Christ Church University, Canterbury, UK
[16]Digital Health Validation Lab, University of Glasgow, Glasgow, UK
[17]Emergency Department, NHS Greater Glasgow & Clyde, Glasgow, UK

Collaborators The AI-REACT Reader Study Group: Satish Golla, Irfan Ullah Akbarkhan, Laura Hunter, Nazreen Kaneez, Ana Nicolescu, Emma Kelliher, Kyle Stephenson, Shair Ali, Danielle Benson, Fiona Hunter, Ross Hunter, Benjamin Scally, Rhys Worgan, Louise Hartley, Ryan Grech, Martine Walker, Neil Mitchell, Ravi Shashikala, Alice Gibson, Hélène Matte, Shubhendu Kulshrestha, Hannah Yang, Radoslaw Rippel, Roland Amoah, Zahi Qamhawi, Thomas Millard, Avneet Gill, Michael Thompson, Josh Beck, Harsh Merchant, Ben Lockwood and Nabeeha Salik.

Contributors AN is the guarantor. AN and SA led the design of the project, with contributions from RS, ATEM, DR, SK, SG, JO, KB, AR, TD, MN, RD, MH, KV, JG, NW, NS, DJL and HS. SA, RS and NS reviewed the image dataset. KB and AR were the primary ground-truthers, with arbitration from TD. NS managed the online CT reading platform and assisted in data collection and management. ATEM registered the study and coordinated reader recruitment and data collection. AN led writing of the manuscript with review from SA. NW and JG led the PPI activities. The AI-REACT Reader Study Group (group authorship) contributed by reviewing the cases for the study, as well as reviewing and providing feedback on the final manuscript.

Funding This work was supported by Qure.ai via the NHSX AI in Health and Care Award grant number AI_AWARD02354.

Competing interests DR, SK and SG are employees of Qure AI. NW declares consultancy fees from InHealth and SM Radiology not related to the current submission. MH declares consultancy fees from Qure AI not related to the current submission.

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. The study has been approved by the UK Health Research Authority (IRAS number 310995, approved 13/12/2022). The use of anonymised retrospective CT scans has been authorised by the Caldicott Guardian and information governance team at Oxford University Hospitals NHS Foundation Trust. Readers will provide written informed consent and will be able to withdraw at any time.

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.

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

REFERENCES

1 Ali UM, Withrow DR, Judge AD, *et al*. Temporal trends in the incidence of malignant and nonmalignant primary brain and central nervous system tumors by the method of diagnosis in England, 1993-2017. *Neuro Oncol* 2023;25:1177–92.
2 Hassan Z, Smith M, Littlewood S, *et al*. Head injuries: a study evaluating the impact of the NICE head injury guidelines. *Emerg Med J* 2005;22:845–9.
3 Arhami Dolatabadi A, Baratloo A, Rouhipour A, *et al*. Interpretation of Computed Tomography of the Head: Emergency Physicians versus Radiologists. *Trauma Mon* 2013;18:86–9.
4 Richards M. Diagnostics: recovery and renewal – report of the independent review of diagnostic services for NHS England. NHS England. Available: https://www.england.nhs.uk/publication/diagnostics-recovery-and-renewal-report-of-the-independent-review-of-diagnostic-services-for-nhs-england/ [Accessed 3 Feb 2024].
5 Davis MA, Rao B, Cedeno PA, *et al*. Machine Learning and Improved Quality Metrics in Acute Intracranial Hemorrhage by Noncontrast Computed Tomography. *Curr Probl Diagn Radiol* 2022;51:556–61.
6 Lee JY, Kim JS, Kim TY, *et al*. Detection and classification of intracranial haemorrhage on CT images using a novel deep-learning algorithm. *Sci Rep* 2020;10:20546.
7 Warman R, Warman A, Warman P, *et al*. Deep Learning System Boosts Radiologist Detection of Intracranial Hemorrhage. *Cureus* 2022;14:e30264.
8 Buchlak QD, Tang CHM, Seah JCY, *et al*. Effects of a comprehensive brain computed tomography deep learning model on radiologist detection accuracy. *Eur Radiol* 2024;34:810–22.
9 Hardy M, Harvey H. Artificial intelligence in diagnostic imaging: impact on the radiography profession. *Br J Radiol* 2020;93:20190840.
10 Chilamkurthy S, Ghosh R, Tanamala S, *et al*. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet* 2018;392:2388–96.
11 Vimalesvaran K, Robert D, Kumar S, *et al*. Assessing the effectiveness of artificial intelligence (AI) in prioritising CT head interpretation: study protocol for a stepped-wedge cluster randomised trial (ACCEPT-AI). *BMJ Open* 2024;14:e078227.
12 Hillal A, Sultani G, Ramgren B, *et al*. Accuracy of automated intracerebral hemorrhage volume measurement on non-contrast computed tomography: a Swedish Stroke Register cohort study. *Neuroradiology* 2023;65:479–88.
13 Chiramal JA, Johnson J, Webster J, *et al*. Artificial Intelligence-based automated CT brain interpretation to accelerate treatment for acute stroke in rural India: An interrupted time series study. *PLOS Glob Public Health* 2024;4:e0003351.
14 Pettet G, West J, Robert D, *et al*. A retrospective audit of an artificial intelligence software for the detection of intracranial haemorrhage used by a teleradiology company in the United Kingdom. *BJR Open* 2024;6:tzae033.
15 Fu H, Novak A, Robert D, *et al*. AI assisted reader evaluation in acute CT head interpretation (AI-REACT): protocol for a multireader multicase study. *BMJ Open* 2024;14:e079824.
16 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;41:602–9.
17 Novak A, Ather S, Morgado ATE, *et al*. Evaluation of the impact of artificial intelligence-assisted image interpretation on the diagnostic performance of clinicians in identifying endotracheal tube position on plain chest X-ray: a multi-case multi-reader study. *Crit Care* 2025;29:330.
18 Hillis SL, Schartz KM. Multireader sample size program for diagnostic studies: demonstration and methodology. *J Med Imaging (Bellingham)* 2018;5:045503.
19 Obuchowski NA, Bullen J. Multireader Diagnostic Accuracy Imaging Studies: Fundamentals of Design and Analysis. *Radiology* 2022;303:26–34.
20 Obuchowski NA, Beiden SV, Berbaum KS, *et al*. Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods. *Acad Radiol* 2004;11:980–95.
21 Fussell DA, Tang CC, Sternhagen J, *et al*. Artificial Intelligence Efficacy as a Function of Trainee Interpreter Proficiency: Lessons from a Randomized Controlled Trial. *AJNR Am J Neuroradiol* 2024;45:1647–54.
22 Hillis SL, Schartz KM. Multireader sample size program for diagnostic studies: demonstration and methodology. *J Med Imag* 2018;5:1.