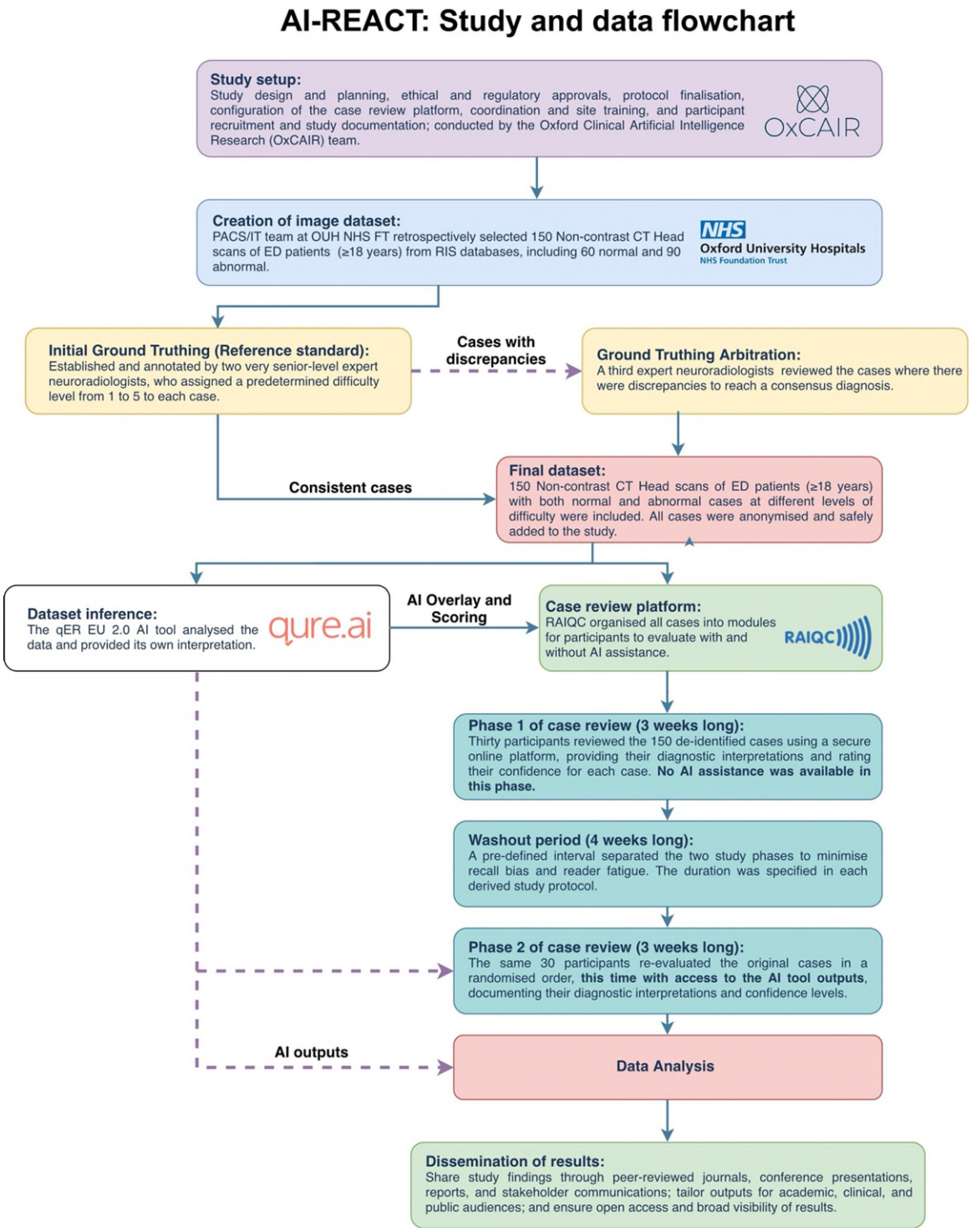
Artificial Intelligence-assisted reader evaluation in acute CT head interpretation (AI-REACT): a multireader multicase study.



Results:

Baseline characteristics

Baseline characteristics are summarised in Table S1 in supplementary material. 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) with in the prespecified seniority classifications each interpreted all 150 cases both with and without AI, totalling 9000 individual interpretations for analysis.

Algorithm versus ground truth

Retrospective analysis of the diagnostic performance of the qER algorithm versus ground truth is presented in Table S2 in supplementary material, and Figure 3. The algorithm showed strong overall diagnostic performance for all abnormality subgroups with AUCs ranging from 0.821 (95% CI 0.740 - 0.903) to 0.986 (95% CI 0.969 - 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 - 0.774). Lower sensitivities (<0.80) were seen for extradural haemorrhage (0.692, 95% CI 0.613 - 0.766), intraparenchymal haemorrhage (0.576, 95% CI 0.490 - 0.654), intraventricular haemorrhage 0.650 (95% CI 0.571 - 0.729), mass effect (0.286, 95% CI 0.216 - 0.366) and fracture (0.727, 95% CI 0.648 - 0.796); lower specificities (<0.80) were seen for infarct (0.699, 95% CI 0.620 - 0.772) and mass effect (0.727, 95% CI 0.648 - 0.796).

Overall reader performance

Changes in overall (pooled) reader performance for different pathological subgroups are presented in table S3 in supplementary material, and figure 4. Diagnostic performance for the detection of critical abnormality measured by area under the curve showed no statistically significant change, however a 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 to 0.956, difference +0.104, 95%CI 0.0602 to 0.147, p<0.001), infarct (0.782 to 0.846, difference +0.0636, 95%CI 0.0211 to 0.106, p=0.0035) and fracture (0.845 to 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 Supplementary Materials Tables S5 to S11).

Subgroup analyses

Relative performance for different reader subgroups in detecting critically abnormal scans are summarised in table S4 in supplementary material, 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 (81.0% to 88.0%, difference +6.9%, 95% CI 0.02.8% to 11.1%) p = 0.001), with increases in sensitivity seen across all pathology subgroups (see 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 Supplemental Materials 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% to 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 192 seconds per case in phase 1 (without AI) and reduced to 173 seconds per case in phase 2 (with AI) ($p < 0.001$). Relevant data were winsorized 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: i) 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, ii) 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 iii) 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 of 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 supplementary material 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 re 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-REACT study is the first to investigate this effect. Few clinical reader studies have been undertaken for any CT Head interpretation algorithm.[19] 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 CT head 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 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.

Whilst 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, i.e. 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 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 CT head 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 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, utilising 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. Whilst AUC is useful to understand the potential of an AI algorithm to accurately identify pathologies and to determine the optimum operating thresholds for reporting a pathological findings as present or absent, in clinical practice need

for a specific pre-determined 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 artificial intelligence (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 non-contrast CT head 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 validity and generalisability of results compared to other studies which have used more subjective and less transparent approaches to creating image datasets.

The reader group itself is large (n=30) compared to other multicase multireader studies.[20] 5 readers represents a typical minimum group size for such studies, so each reader speciality subgroup in our study included at least five readers, allowing for independent subgroup analyses.[17] 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 that may benefit from AI assistance. This allows the potential utility of AI-assisted CT Head 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 request form – this reflects real-world practice for radiologists, but clinician readers would have not been able to judge who was high risk for the presence of pathology, e.g. 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.

Ethics and dissemination

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.

The study is registered at Clinicaltrials.gov (NCT06018545), and the ISRCTN registry (ISRCTN17560291). The results of the study will be presented at relevant conferences and published in peer-reviewed journals. The detailed study protocol will be freely available upon request to the corresponding author. Further dissemination strategy will be strongly guided by our PPIE activities. This will be based on co-productions between patient partners and academics and will involve media pieces (mainstream and social media) as well as communication through charity partners.

Authors' contributions

AN and SA led the design of the project, with contributions from RS, ATEM, DR, SK, SG, JO, KB, AR, TD, MTG, MN, RD, MH, KV, JG, NW, NS, AC, FK, DL, and HS. SA, RS, FK and NS and AC and reviewed the image dataset. KB and AR are the primary ground-truthers, with arbitration from TD. NS manages the online CT reading platform and assisted in data collection and management. ATEM registered the study and coordinates reader recruitment and data collection. AN, and SA wrote the manuscript. NW and JG led the PPI activities.

Funding statement

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

Competing interests statement

Dennis Robert, Shamie Kumar, and Satish Goya are employees of Qure AI. Nick Woznitza declares consultancy fees from InHealth and SM Radiology not related to the current submission. Mark Harrison declares consultancy fees from Qure AI not related to the current submission.