BMJ Open

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

Howell Fu ,[1] Alex Novak,[2] Dennis Robert,[3] Shamie Kumar,[3] Swetha Tanamala,[3] Jason Oke,[4] Kanika Bhatia,[1] Ruchir Shah,[1] Andrea Romsauerova,[1] Tilak Das,[5] Abdalá Espinosa,[2] Mariusz Tadeusz Grzeda,[6] Mariapaola Narbone,[7] Rahul Dharmadhikari,[8] Mark Harrison,[9] Kavitha Vimalesvaran ,[10] Jane Gooch,[11] Nicholas Woznitza ,[12,13] Nabeeha Salik,[14] Alan Campbell,[12] Farhaan Khan,[1] David J Lowe ,[15] Haris Shuaib,[10] Sarim Ather[1]

To cite: 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. doi:10.1136/bmjopen-2023-079824

► Prepublication history for this paper is available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2023-079824).

HF and AN are joint first authors.

Received 13 September 2023
Accepted 28 January 2024

For numbered affiliations see end of article.

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

ABSTRACT

Introduction A non-contrast CT head scan (NCCTH) is the most common cross-sectional imaging investigation requested in the emergency department. Advances in computer vision have led to development of several artificial intelligence (AI) tools to detect abnormalities on NCCTH. These tools are intended to provide clinical decision support for clinicians, rather than stand-alone diagnostic devices. However, validation studies mostly compare AI performance against radiologists, and there is relative paucity of evidence on the impact of AI assistance on other healthcare staff who review NCCTH in their daily clinical practice.

Methods and analysis A retrospective data set of 150 NCCTH will be compiled, to include 60 control cases and 90 cases with intracranial haemorrhage, hypodensities suggestive of infarct, midline shift, mass effect or skull fracture. The intracranial haemorrhage cases will be subclassified into extradural, subdural, subarachnoid, intraparenchymal and intraventricular. 30 readers will be recruited across four National Health Service (NHS) trusts including 10 general radiologists, 15 emergency medicine clinicians and 5 CT radiographers of varying experience. Readers will interpret each scan first without, then with, the assistance of the qER EU 2.0 AI tool, with an intervening 2-week washout period. Using a panel of neuroradiologists as ground truth, the stand-alone performance of qER will be assessed, and its impact on the readers' performance will be analysed as change in accuracy (area under the curve), median review time per scan and self-reported diagnostic confidence. Subgroup analyses will be performed by reader professional group, reader seniority, pathological finding, and neuroradiologist-rated difficulty.

Ethics and dissemination The study has been approved by the UK Healthcare Research Authority (IRAS 310995, approved 13 December 2022). The use of anonymised retrospective NCCTH has been authorised by Oxford University Hospitals. The results will be presented at relevant conferences and published in a peer-reviewed journal.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This study will evaluate 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 will be the first UK-based multicentre validation of an AI for non-contrast CT head scan trained on a large data set (300 000 head CTs).
⇒ It includes non-radiologists (emergency medicine clinicians and radiographers) among the healthcare professionals that may benefit from AI assistance.
⇒ The prevalence of pathologies in the selected scans will be 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 will limit 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 will need to be excluded from the study.

Trial registration number NCT06018545.

INTRODUCTION

Diagnostic imaging plays a critical role in the timely and appropriate management of emergency department (ED) patients.[1 2] Emergency medicine (EM) clinicians routinely interpret and act on their findings for plain radiography, but are generally dependent on radiologists for more complex modalities such as CT. Reporting turnaround time already has a significant impact on ED workflow,[3] and demand will continue to increase relative to the radiologist workforce.[4 5] The

advent of artificial intelligence (AI)-assisted image interpretation offers a potential way to mitigate the impact of this shortfall, by improving radiologist reporting efficiency in terms of reporting time or worklist prioritisation or by elevating the interpretation accuracy of EM clinicians and radiographers to support diagnostic and treatment decision-making prior to the availability of a full radiology report.

Non-contrast CT head (NCCTH) is the most common cross-sectional imaging investigation performed in ED, comprising up to half of ED CT scans,[6] and performed in up to 9% of all ED encounters.[7] Recently developed AI tools are capable of identifying critical pathologies, mapping their location for clinician review and flagging abnormal scans for urgent attention.[8–10] These tools have demonstrated sensitivity and specificity comparable to neuroradiologists for important findings such as intracranial haemorrhage, cerebral infarct, and skull fracture,[8 11–14] and can reduce turnaround times for scans with urgent findings.[9 14 15]

Most AI solutions are currently designed as decision support tools rather than stand-alone diagnostic devices, and in the foreseeable future clinicians are likely to retain responsibility for interpretations, diagnosis and subsequent treatment decisions.[16 17] Despite this, most published work evaluates the stand-alone performance of AI tools and compares them to that of neuroradiologists. Few studies have directly evaluated the impact of AI assistance in improving the accuracy of general radiologists in NCCTH interpretation,[12 18 19] and fewer have evaluated its impact on reporter speed and confidence.[18] Other professional groups who may use AI include EM clinicians, who commonly review CT scans prior to radiology reports becoming available, in order to expedite patient management.[20 21] To our knowledge, only one previous study has evaluated the impact of AI assistance with other groups of healthcare professionals who regularly review or act on NCCTH interpretations, such as EM clinicians and radiographers.[22]

Although some AI tools have been validated in the UK population, most have been in a single-centre setting with relatively small data sets.[23–25] The AI tools which have been developed using the largest data sets (above 100 000 scans) have originated from China and India[11 12] and have not been validated in the UK. The geographical setting is significant because early AI models' performance degraded sharply when tested on data from different populations, even from different hospitals within the same city, due to differences in patient characteristics, prevalence of abnormalities and imaging hardware.[17 26 27] More advanced models with larger and more diverse data sets show improved performance in this regard.[11 12]

qER 2.0 EU is a US Food and Drug Administration cleared and CE (European Conformity) marked AI tool for interpretation of NCCTH, which was developed using 300 000 retrospectively collected 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. It can detect, classify and localise intracranial haemorrhage, hypodensities suggestive of infarct, mass effect, midline shift, atrophy and skull fractures in NCCTH.[11] 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 1). Alternatively, if none of the target abnormalities are detected, the output

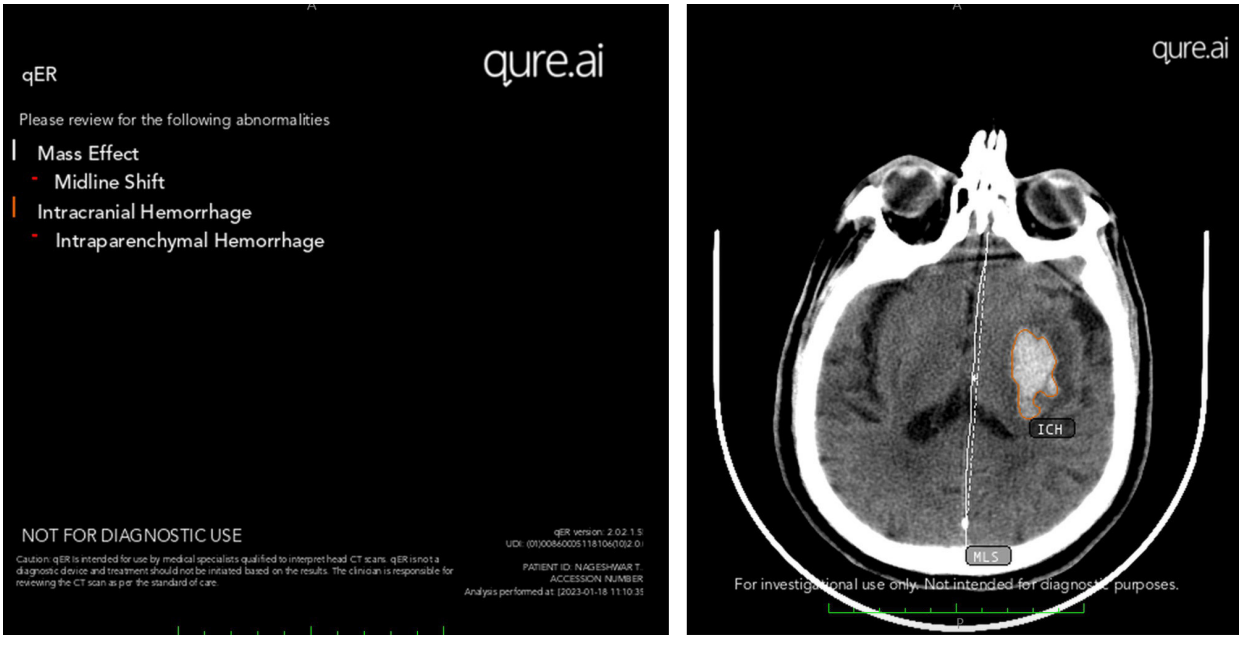

Figure 1 qER presents a summary of all abnormalities identified and the slice images containing those abnormalities with a localisation overlay. ICH, intracranial haemorrhage.

will indicate that the software has analysed the image and identified no target abnormalities.

Previous literature provides baseline expectations for the performance of qER. In qER's first publication, the AI was tested on an external validation data set of 491 NCCTH scans from patients in India. The Area Under the Receiver Operating Characteristics Curve (AUC) for ICH, skull fracture, midline shift and mass effect was 94.19% (95% CI: 91.87% to 96.51%), 96.24% (95% CI: 92.04% to 100.00%), 96.97% (95% CI: 94.03% to 99.91%) and 92.16% (95% CI: 88.83% to 95.48%), respectively.[11] In a Swedish stroke registry study, qER was found to have about 97% sensitivity in detecting non-traumatic ICH, and 95% of the false-negative ICHs were <1 mm in diameter.[28] In an unpublished study conducted for the purpose of regulatory submission, using 1320 NCCTH scans from multiple sites in the USA, the AUC was reported as over 97% for ICH, skull fracture, mass effect and midline shift. The sensitivity and specificity in detecting any one of the four abnormalities (ICH, skull fracture, mass effect, midline shift) were reported to be 98.53% (95% CI: 97.45% to 99.24%) and 91.22% (95% CI: 88.39% to 93.55%), respectively.[29]

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. As yet, its potential impact on the diagnostic accuracy of radiologists, radiographers and EM clinicians has not been fully evaluated.

To our knowledge, the current study will be the first UK-based multicentre validation of an AI for NCCTH trained on a large data set. It will assess the impact of AI assistance on clinician accuracy, and furthermore will include its impact on EM clinicians and radiographers, both of which are important areas with a paucity of published research.

Study aims

1. To determine the improvement in NCCTH image interpretation accuracy of general radiologists, EM 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, in the UK ED population, the standalone accuracy of qER at detecting 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, presence of multiple abnormalities, clinician seniority and professional group (secondary).

METHODS AND ANALYSIS

Study design

The study will employ a fully-crossed paired multireader multicase (MRMC) design. Recruitment will begin in August 2023, with estimated completion of data capture and analysis by March 2024.

Case selection

150 NCCTH of ED patients aged 18 years or above will be retrospectively identified by the clinical and Picture Archiving and Communication System (PACS)/Information Technology (IT) team by searching the Radiology Information System at Oxford University Hospitals NHS Foundation Trust (figure 2). The case mix will include 60 control scans and 90 abnormal scans, including a minimum of 10 scans containing each of the following 9 critical abnormalities (for definitions, see Appendix):

1. Extradural haemorrhage.
2. Subdural haemorrhage.
3. Subarachnoid haemorrhage.
4. Intraparenchymal haemorrhage.
5. Intraventricular haemorrhage.
6. Hypodensity suggestive of infarct.
7. Midline shift.
8. Mass effect.
9. Skull fractures.

For the purposes of case selection, the existing clinical radiology reports will be used to determine whether a given scan contains an abnormality of interest. Consecutive scans will be reviewed and all scans fitting the inclusion and exclusion criteria will be included until the case number requirements have been met. A subset of images may demonstrate multiple of the above abnormalities. The control cases may include scans with other abnormalities than the nine of interest listed above, as well as normal scans. Every case will be from a distinct patient, and no patient will have more than one scan included in the study.

Inclusion criteria for cases:

- 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 for cases will consist of the following features which 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.

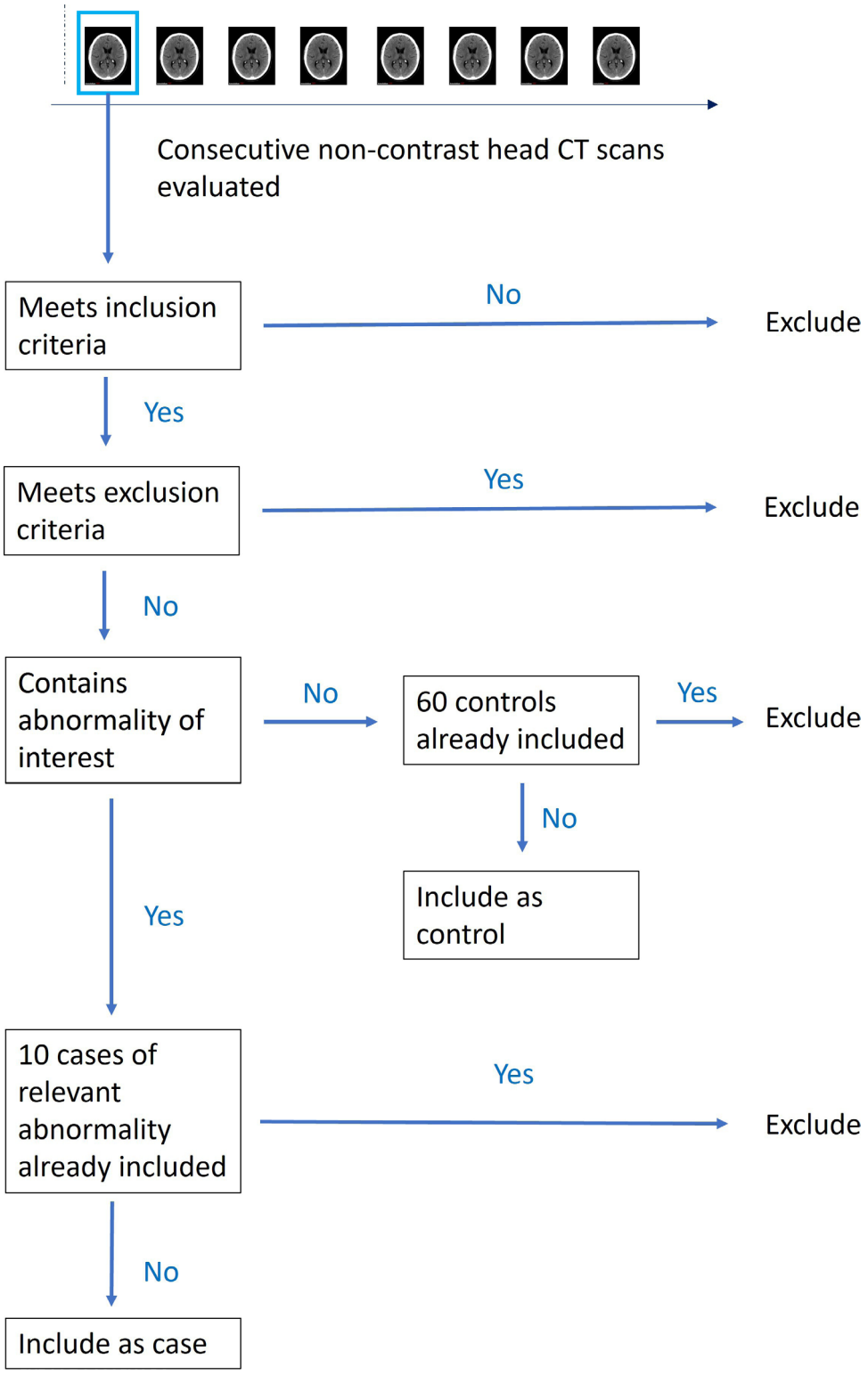

Figure 2 Case selection flow diagram.

Setting

Readers will be recruited from the following four hospital Trusts:

- Guy's & St Thomas NHS Foundation Trust.
- Northumbria Healthcare NHS Foundation Trust.
- NHS Greater Glasgow and Clyde.
- Oxford University Hospitals NHS Foundation Trust.

Participants

30 volunteer participant readers will be selected from the following groups:

- Emergency medicine consultants and registrars (5 consultant, 5 registrar (ST3-6), 5 junior (F1-ST2).
- General radiologist consultants and registrars (5 consultant, 5 registrar (ST3-6)).

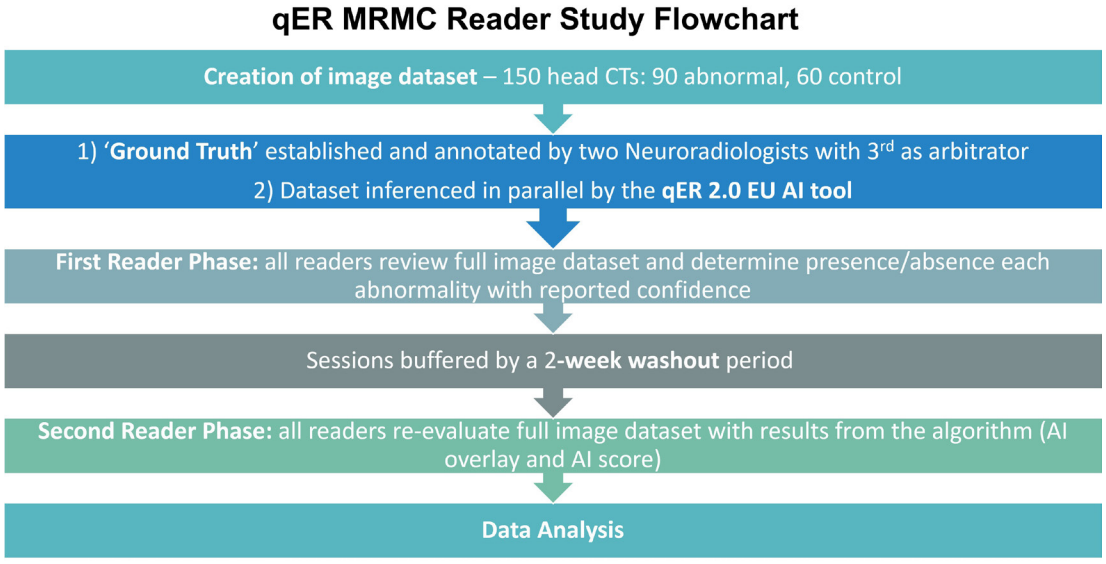

Figure 3 Reader study flowchart. AI, artificial intelligence; MRMC, multireader multicase.

- 5 CT radiographers.

Inclusion criteria:

- Radiologists/radiographers/EM clinicians who review NCCTH as part of their clinical practice.

Exclusion criteria:

- Neuroradiologists.
- (Non-radiologist groups) Clinicians with previous formal postgraduate CT reporting training.
- (Emergency medicine group) Clinicians with previous career in radiology/neurosurgery to registrar level.

Reader recruitment will be carried out by each principal investigator for their own site. The call for volunteers will be disseminated to all radiologists, radiographers and EM staff via email and in person, and the principal investigator will check all volunteers against the inclusion and exclusion criteria until the required reader numbers are met.

CT interpretation

All 30 readers will review all 150 cases, in each of two study phases (figure 3). For each scan, the readers will provide their opinion on whether any critical abnormalities are present, and if so, on the presence or absence of each of the nine abnormalities listed above. They will also be asked to provide a confidence for each of their diagnoses on a 10-point quasi-continuous scale (QCS) (figure 4). The time taken for each scan (not including confidence QCS) will be automatically recorded. The order of the cases will be randomised for each reader at each phase, and readers will be unaware of the number of cases versus controls.

Phase 1: All readers review all scans, blinded to the ground truth and without AI assistance.

Washout period: 2 weeks, to mitigate recall bias.

Phase 2: All readers review all scans again, in a randomised order, remaining blinded to the ground truth, but with access to the results from the qER tool (figure 3). The qER output will be presented as an additional series, available in the same window by clicking on it. In addition to the original scan, the qER output will include a notification to suggest the presence/absence of a target abnormality as the first image of the series (figure 4C) and segmentation of the abnormal areas identified overlaid on the scan images (figure 4D). Readers will rate their confidence as they did in Phase 1. If readers disagree with the algorithm output, they will be asked to state their reasons using a free-text box (time taken to do this will not be included for the purposes of the speed-of-interpretation analysis).

The reads will be performed using a secure web-based DICOM viewer (www.raiqc.com). Prior to commencing each phase of the study, the readers will be asked to review 5 practice cases (not part of the 150 case data set) to familiarise themselves with the use of the study platform and the output of the qER tool.

Ground truthing

Two consultant neuroradiologists will independently review the images to establish the 'ground truth' findings on the CT scans which will be used as the reference standard. In the case of disagreement, a third senior neuroradiologist's opinion will be sought for arbitration. A difficulty score will be assigned to each scan by the two ground truthers using a 5-point Likert scale (figure 4), and where there is disagreement the mean score will be taken.

Outcome measures

The primary outcome measure will be the difference in AUC of readers in classifying a scan as critical or non-critical, with vs without AI assistance.

Secondary outcome measures: The differences in reader sensitivity and specificity, with versus without AI assistance, will also be derived from the reader classifications. The stand-alone diagnostic performance of the qER AI will be evaluated by estimating sensitivity and specificity of qER in classifying a scan as critical versus

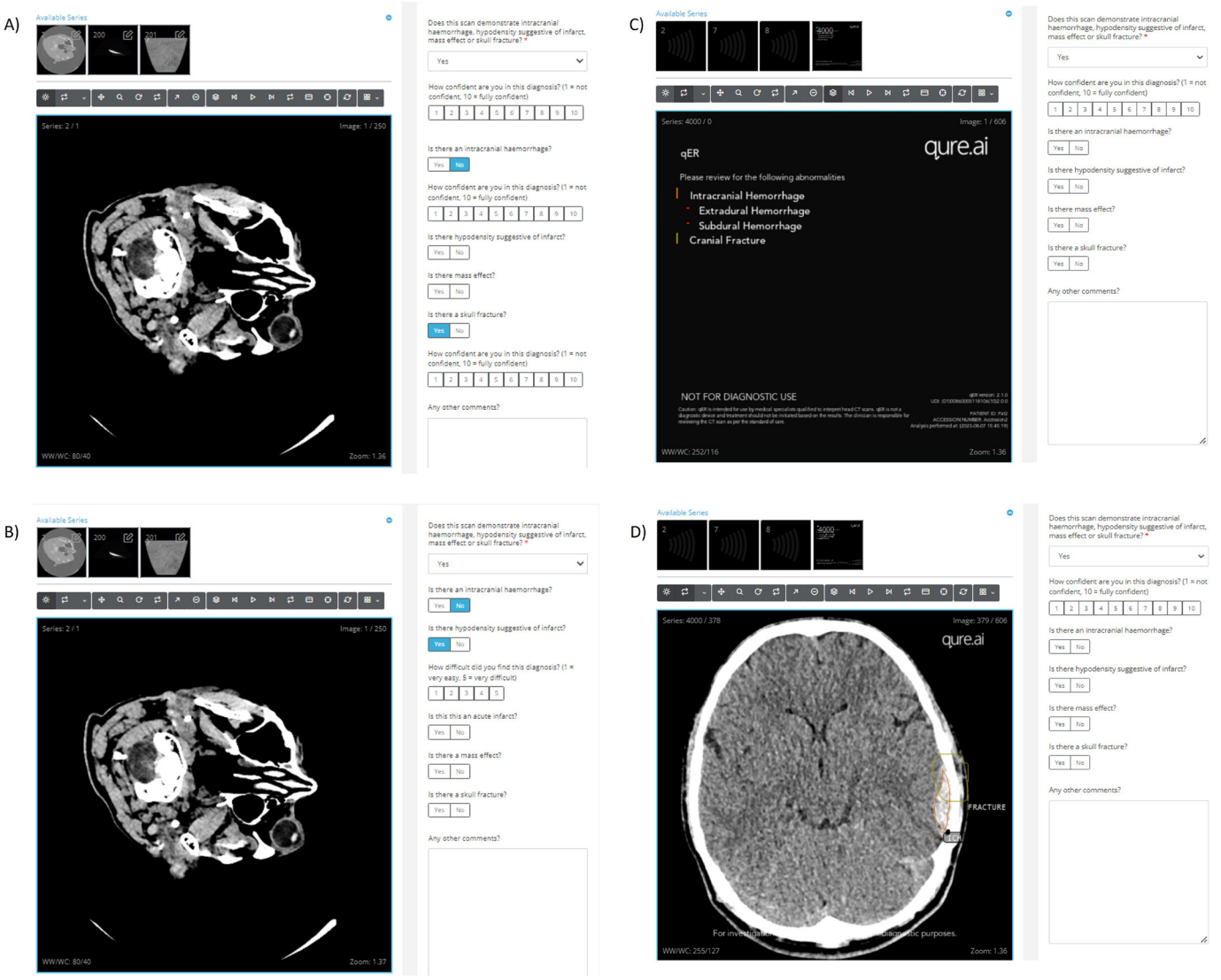

Figure 4 Sample reader questions and responses. (A) Readers will be asked to state their confidence in each positive diagnosis they make, on a quasi-continuous scale from 1 to 10. (B) Ground truthers will be asked to rate the difficulty of each diagnosis on a Likert scale from 1 to 5. (C) The artificial intelligence output will be presented as an additional available series with a notification of the abnormalities detected and (D) segmentation of the abnormalities overlaid on the scan images.

non-critical and by AUC, sensitivity and specificity of qER for detecting each target abnormalities. The difference in reader speed will be evaluated as the median review time per scan with versus without AI assistance.

Data de-identification and management

Scans selected for the study will be anonymised in accordance with Oxford University Hospitals NHS Foundation Trust information governance protocol using the Insignia Insight Anonymisation tool and uploaded to the secure image viewing platform (www.raiqc.com). Access to the scans will be controlled via the study platform using separate user accounts for each reader. The anonymised images will be sent securely to Qure servers, where the AI analysis will run, and its outputs will be transferred back to RAIQC (figure 5).

All study data will be entered into a password-protected and secure database. Individual reader accuracy scores will be anonymised, and the study team will not have access to the identifying link between the readers' personal details and the data. Data about the readers' seniority level and professional group will be retained to allow group comparisons.

Sample size and power calculation

A sample of 30 readers and minimum 135 scans (82 with presence of critical findings and 53 with no critical findings) will have 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 intra-reader variability of 0.3 and 0.05, respectively, a 0.35 conservative correlation between readers, and anticipated average readers' AUC of 0.75, guided by previous literature.[30–32] Since we made few assumptions, a higher sample size of 150 NCCTH (90 with critical findings and 60 without any critical findings) will be used for the study.

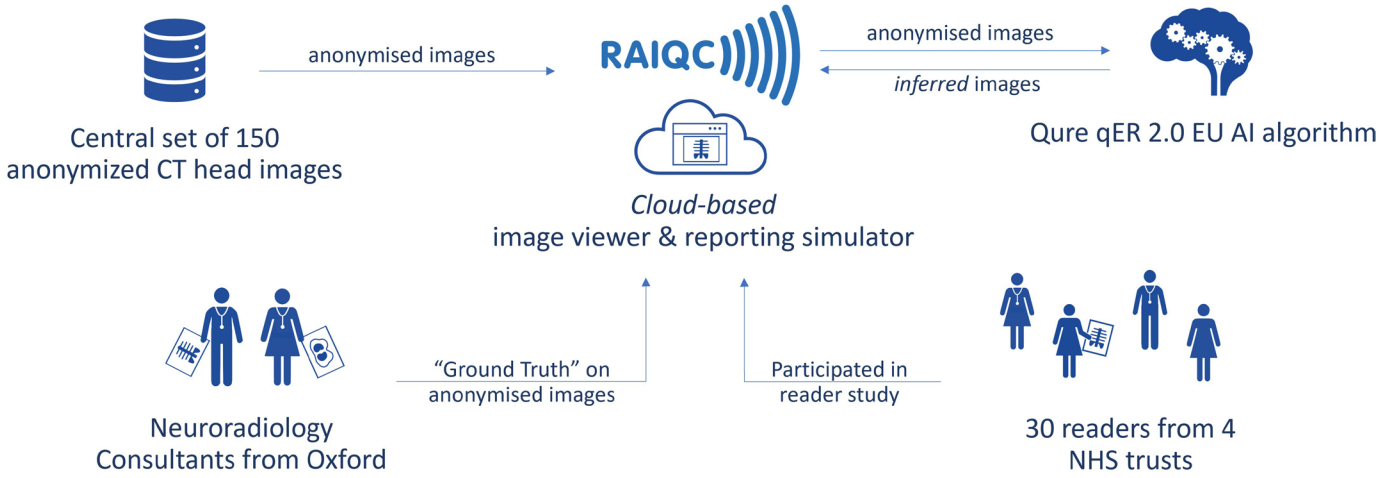

Figure 5 Data flow diagram. AI, artificial intelligence; NHS, National Health Service.

Statistical analyses

The difference in AUC of readers with and without AI will be tested based on the Obuchowski-Rockette model for MRMC analysis which will model the data using a two-way mixed effects analysis of variance model treating readers and cases (images) as random effects and effect of AI as a fixed effect[31] with recommended adjustment to df by Hillis *et al*.[33] Sensitivity and specificity will be analysed as part of this model. The main analysis will be performed as a single pool including all groups and sites. Subgroup analyses will be performed for the following:

- Professional group (radiologist vs EM clinician vs radiographer).
- Senior versus junior.
- Pathological finding
- Difficulty of image.

The median review time per scan with vs without AI will be compared using a non-parametric Wilcoxon sign-rank test.

To gain insight into how AI may bias reader judgement, a comparison will be made between the rate at which readers agree with the AI when it is correct vs when it is incorrect.

The stand-alone performance of qER algorithm will be compared with the ground truth generated by the neuroradiologists. The continuous probability score from the algorithm will be used for the AUC analyses, while binary classification results will be used for evaluation of sensitivity, specificity, positive predictive value and negative predictive value. The false-negative rate (number of missed pathologies) by the AI and the human readers will be reported for each of the nine studied abnormalities.

Patient and public involvement

JG is a co-investigator for the study. As a brain aneurysm survivor, she has experience of the diagnosis and management of neurological emergencies and recognises the importance that rapid diagnosis has on improving treatment outcomes. JG is a patient advisor to the National Institute for Health and Care Excellence head injury guideline review and an active contributor to patient and carer fora. JG has reviewed the initial study priorities and research question; her feedback emphasised the need to measure the time to report for normal examinations as this is also important to patients. JG also suggested inclusion of patient focus groups at key stages of the study (design, preliminary and final results) and this has been incorporated into the trial. JG has reviewed and edited the lay abstract, helped with the research question design and study protocol development. This protocol has been reviewed by the Oxford ACUTECare PPI group. They supported the study and its aims, influenced design and data management and dissemination strategies. The recruited study participants will be healthcare professionals rather than patients, and therefore patients will not be involved in recruitment. The study results will be disseminated to these participants via email.

DISCUSSION

AI tools such as qER are a potential way to mitigate the pressures of increasing imaging demand and radiologist workforce shortage in the National Health Service. They have the potential to improve the efficiency and accuracy of radiologist reporting and empower non-radiologists to optimise patient workflow and further improve patient care. Decisions regarding their implementation ought to be informed by evidence confirming their diagnostic performance in the UK context, and assessing their efficacy in improving the accuracy, confidence, and speed of clinicians who use them.

The qER AI tool used in this study has previously reported good performance, as described above. However, it does not detect all possible pathologies in an NCCTH, and is limited to the nine target abnormalities listed above. Its use is therefore limited to assist readers only with the abnormalities it is capable of detecting. Furthermore, the version of the AI tool used in this study does not differentiate between acute and chronic infarcts, and this decision will remain dependent on the readers' expertise and clinical judgement. Finally, the AI tool is unable to interpret scans with postoperative changes and implants, such

as burr holes, shunts and clips, or scans with significant artefact, for example, patient movement and metal artefact. These scans form a small but important minority of real-world clinical caseload, but they cannot be included in the current study.

The main strengths of the current study are that the data will be derived from a UK ED population, that the multicentre design allows representation of a broad range of readers from across the UK and that the design is specifically tailored to address the question of how the AI tool functions as a decision aid and impacts the performance of its human users, which is the use case for which it is marketed and currently being adopted.

In order to ensure that the study is powered to detect these effects, the study will use an enriched data set with a higher proportion of abnormalities than would be observed in normal clinical work. The chosen proportions ensure that the study is powered to detect a difference in average readers' AUC of 5% between aided and unaided reads. Since AUC, sensitivity and specificity are considered to be inherent measures of accuracy unlike positive and negative predictive values, the ratio of abnormal to normal cases will not affect the estimations of these metrics assuming the spectrum of the disease in the sample is not largely different from the population. If results from the study are encouraging, they may inform larger follow-up studies using data sets representative of real-life ED caseload.

ETHICS AND DISSEMINATION

The study has been approved by the UK Health Research Authority (IRAS number 310995, approved 13 December 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, and the ISRCTN registry (approval pending). 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 on request to the corresponding author. Further dissemination strategy will be strongly guided by our Patient and Public Involvement 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.

APPENDIX: DEFINITIONS OF CRITICAL ABNORMALITIES

Intracranial haemorrhage (ICH) (including subtype): Any type of bleeding within the brain and cranial vault. It encompasses five broad types of haemorrhage: epidural haemorrhage, subdural haemorrhage, subarachnoid haemorrhage, intraventricular haemorrhage and intraparenchymal haemorrhage.

Midline shift (MS): A horizontal shift of the brain past its centre line, a subset of mass effect.

Mass effect (ME): A visible compression or displacement of adjacent structures of the brain parenchyma, sulci or ventricles and can cause midline shift as results of an underlying pathology.

Cranial fracture: One or more breaks in the cranial bone. Often seen as lucencies and/or as discontinuities in the bone.

Hypodensity (suggestive of infarct): An area of necrosis in the brain tissue, resulting from obstruction of the local circulation by a thrombus or embolus. Early features of an infarct on CT scan include loss of grey-white matter differentiation, and cortical hypodensity with associated parenchymal swelling. This can be either an acute or a chronic infarct.

Cerebral atrophy: Loss of neurons and its connections due to conditions such as stroke, neurodegenerative diseases. Seen as brain parenchymal volume loss on a head CT scan.

Author affiliations
[1]Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[2]Emergency Medicine Research Oxford, Oxford University Hospitals NHS Foundation Trust, Oxford, UK
[3]Qure.AI, Bangalore, India
[4]Department of Primary Care Health Sciences, University of Oxford, Oxford, UK
[5]Department of Clinical Radiology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
[6]School of Biomedical Science, King's College London, London, UK
[7]Guy's and St Thomas' Hospitals NHS Trust, London, UK
[8]Northumbria Healthcare NHS Foundation Trust, Northumberland, UK
[9]Emergency Department, Northumbria Specialist Emergency Care Hospital, Cramlington, UK
[10]Clinical Scientific Computing, Guy's and St Thomas' NHS Foundation Trust, London, UK
[11]College of Health, Psychology & Social Care, University of Derby, Derby, UK
[12]Radiology Department, University College London Hospitals NHS Foundation Trust, London, UK
[13]School of Allied and Public Health Professions, Canterbury Christ Church University, Canterbury, UK
[14]RAIQC Ltd, Oxford, UK
[15]NHS Greater Glasgow and Clyde, Glasgow, UK

Twitter Dennis Robert @technOslerphile, Kavitha Vimalesvaran @kavitha_varan, Nicholas Woznitza @xray_nick and David J Lowe @djlmed

Contributors AN and SA led the design of the protocol, with contributions from DR, SK, ST, JO, RS, MTG, RD, MH, KV, JG, NW, NS, AC, FK, DJL and HS. NW and JG led the PPI activities. SA, RS, FK, NS and AC and reviewed the image data set. KB and AR are the primary ground-truthers, with arbitration from TD. NS manages the online CT reading platform and will be aiding in data collection and management. AE registered the study and coordinates reader recruitment and data collection. MTG performed the simulations estimating statistical power for the study. HF, AN, DR and SA wrote the manuscript.

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

Competing interests DR, SK and ST 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. DJL declares an institution grant for additional research activity on a separate product unrelated to the current submission. AN declares another NHSX grant in collaboration with Qure AI for research unrelated to the current submission. SA declares grants from Qure AI for other research activity unrelated 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.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Howell Fu http://orcid.org/0000-0002-0582-9518
Kavitha Vimalesvaran http://orcid.org/0000-0003-2236-7279
Nicholas Woznitza http://orcid.org/0000-0001-9598-189X
David J Lowe http://orcid.org/0000-0003-4866-2049

REFERENCES

1 Juszczyk K, Ireland K, Thomas B, *et al*. Reduction in hospital admissions with an early computed tomography scan: results of an outpatient management protocol for uncomplicated acute diverticulitis. *ANZ J Surg* 2019;89:1085–90.
2 Chan J, Fan KS, Mak TLA, *et al*. Pre-operative imaging can reduce negative appendectomy rate in acute appendicitis. *Ulster Med J* 2020;89:25–8.
3 Greenhalgh R, Howlett DC, Drinkwater KJ. Royal College of Radiologists national audit evaluating the provision of imaging in the severely injured patient and compliance with national guidelines. *Clin Radiol* 2020;75:224–31.
4 Richards M. Diagnostics: recovery and renewal – report of the independent review of diagnostic services for NHS England [NHS England]. 2022. Available: https://www.england.nhs.uk/publication/diagnostics-recovery-and-renewal-report-of-the-independent-review-of-diagnostic-services-for-nhs-england/ [Accessed 6 May 2023].
5 Royal College of Radiologists. Clinical Radiology UK workforce census 2019 report. 2020. Available: https://www.rcr.ac.uk/publication/clinical-radiology-uk-workforce-census-2019-report [Accessed 6 May 2023].
6 Seidel J, Bissell MB, Vatturi S, *et al*. Retrospective analysis of emergency computed tomography imaging utilization at an academic centre: an analysis of clinical indications and outcomes. *Can Assoc Radiol J* 2019;70:13–22.
7 Prevedello LM, Raja AS, Zane RD, *et al*. Variation in use of head computed tomography by emergency physicians. *Am J Med* 2012;125:356–64.
8 Lin E, Yuh EL. Computational Approaches for Acute Traumatic Brain Injury Image Recognition. *Front Neurol* 2022;13:791816.
9 Sheth SA, Giancardo L, Colasurdo M, *et al*. Machine learning and acute stroke imaging. *J Neurointerv Surg* 2023;15:195–9.
10 Yeo M, Tahayori B, Kok HK, *et al*. Review of deep learning algorithms for the automatic detection of intracranial hemorrhages on computed tomography head imaging. *J Neurointerv Surg* 2021;13:369–78.
11 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.
12 Guo Y, He Y, Lyu J, *et al*. Deep learning with weak annotation from diagnosis reports for detection of multiple head disorders: a prospective, multicentre study. *Lancet Digit Health* 2022;4:e584–93.
13 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.
14 Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, *et al*. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *NPJ Digit Med* 2018;1:9.
15 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.
16 National Institute for Health and Care Excellence. Artificial intelligence for analysing CT brain scans. 2020. Available: https://www.nice.org.uk/advice/mib207 [Accessed 6 May 2023].
17 Wardlaw JM, Mair G, von Kummer R, *et al*. Accuracy of automated computer-aided diagnosis for stroke imaging: a critical evaluation of current evidence. *Stroke* 2022;53:2393–403.
18 Finck T, Moosbauer J, Probst M, *et al*. Faster and better: how anomaly detection can accelerate and improve reporting of head computed tomography. *Diagnostics*;12:452.
19 Warman R, Warman A, Warman P, *et al*. Deep learning system boosts radiologist detection of intracranial hemorrhage. *Cureus* 2022;14:e30264.
20 Ünlüer EE, Yaka E, Akhan G, *et al*. Ability of emergency physicians to detect early ischemic changes of acute ischemic stroke on cranial computed tomography. *Med Princ Pract* 2012;21:534–7.
21 Dolatabadi AA, Baratloo A, Rouhipour A, *et al*. Interpretation of computed tomography of the head: emergency physicians versus radiologists. *Trauma Mon* 2013;18:86–9.
22 Scavasine VC, Ferreti LA, Costa RT, *et al*. Artificial intelligence in the emergency room: how E-ASPECTS helps emergency physicians evaluate brain CT of patients with acute ischemic stroke. *SSRN Journal* 2022.
23 Dyer T, Chawda S, Alkilani R, *et al*. Validation of an artificial intelligence solution for acute triage and rule-out normal of non-contrast CT head scans. *Neuroradiology* 2022;64:735–43.
24 Mallon DH, Taylor EJR, Vittay OI, *et al*. Comparison of automated ASPECTS, large vessel occlusion detection and CTP analysis provided by Brainomix and RapidAI in patients with suspected ischaemic stroke. *J Stroke Cerebrovasc Dis* 2022;31:106702.
25 Andralojc LE, Kim DH, Edwards AJ. Diagnostic accuracy of a decision-support software for the detection of intracranial large-vessel occlusion in CT angiography. *Clin Radiol* 2023;78:e313–8.
26 Zech JR, Badgeley MA, Liu M, *et al*. Variable generalization performance of A deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Med* 2018;15:e1002683.
27 Huang S-C, Pareek A, Jensen M, *et al*. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *NPJ Digit Med* 2023;6:74:74.:.
28 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.
29 US Food and Drug Administration. FDA marketing approval K200921. 2020. Available: https://www.accessdata.fda.gov/cdrh_docs/pdf20/K200921.pdf [Accessed 1 Dec 2023].
30 Obuchowski NA. Sample size tables for receiver operating characteristic studies. *AJR Am J Roentgenol* 2000;175:603–8.
31 Obuchowski NA, Rockette HE. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an anova approach with dependent observations. *Communications in Statistics - Simulation and Computation* 1995;24:285–308.
32 Rockette HE, Campbell WL, Britton CA, *et al*. Empiric assessment of parameters that affect the design of multireader receiver operating characteristic studies. *Acad Radiol* 1999;6:723–9.
33 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.