

AI Assisted Reader Evaluation in Acute CT Head Interpretation (AI-REACT): A Multi-Reader Multi-Case Study

Howell Fu ^{1#}	Howell.Fu@ouh.nhs.uk ORCID: 0000-0002-0582-9518
Alex Novak ^{2#}	alex.novak@ouh.nhs.uk ORCID: 0000-0002-5880-8235
Dennis Robert ³	Dennis.Robert@qure.ai ORCID: 0000-0002-8539-6408
Shamie Kumar ³	Shamie.Kumar@qure.ai
Swetha Tanamala ³	Swetha.Tanamala@qure.ai
Jason Oke ⁴	Jason.oke@phc.ox.ac.uk
Kanika Bhatia ¹	Kanika.Bhatia@ouh.nhs.uk
Ruchir Shah ¹	Ruchir.shah@ouh.nhs.uk
Andrea Romsauerova ¹	Andrea.Romsauerova@ouh.nhs.uk ORCID: 0000-0002-8539-6408
Abdala Espinosa ²	Abdala.espinosa@ouh.nhs.uk ORCID: 0000-0003-0967-3554
Mariusz Tadeusz Grzeda ⁵	mt.grzeda@gmail.com
Mariapaola Narbone ⁶	mariapaola.narbone@gstt.nhs.uk ORCID: 0009-0003-9600-6801
Rahul Dharmadhikari ⁷	rahul.dharmadhikari@northumbria-healthcare.nhs.uk
Mark Harrison ⁸	mark.harrison@northumbria-healthcare.nhs.uk ORCID: 0000-0002-5537-5147
Kavitha Vimalasvaran ⁹	kavitha.1.vimalasvaran@kcl.ac.uk ORCID: 0000-0003-2236-7279
Jane Gooch ¹⁰	j.gooch@derby.ac.uk ORCID: 0000-0003-0136-4116
Nick Woznitza ^{11, 12}	nicholas.woznitza@nhs.net ORCID: 0000-0001-9598-189X
Nabeeha Salik ¹³	nabeeha@raiqc.com
Alan Campbell ¹¹	alan.campbell3@nhs.net ORCID: 0000-0002-5956-8560
Farhaan Khan ¹	Farhaan.khan@ouh.nhs.uk
David J Lowe ¹⁴	david.lowe@ggc.scot.nhs.uk ORCID: 0000-0003-4866-2049
Haris Shuaib ⁶	haris.shuaib@kcl.ac.uk ORCID: 0000-0001-6975-5960
Sarim Ather ¹	sarim.ather@ouh.nhs.uk ORCID: 0000-0001-9614-5033

¹ Oxford University Hospitals NHS Foundation Trust, Oxford

² Emergency Medicine Research Oxford (EMROx), Oxford University Hospitals NHS FT, Oxford

³ Qure.AI

⁴ Department of Primary Health Care Sciences, University of Oxford

⁵ School of Biomedical Science, King's College London

⁶ Guy's & St Thomas' NHS Foundation Trust, London

⁷ Northumbria Healthcare NHS Foundation Trust, Northumberland

⁸ Emergency Department, Northumbria Specialist Emergency Care Hospital, Cramlington

⁹ Clinical Scientific Computing, Guy's and St Thomas' NHS Foundation Trust, London

¹⁰ College of Health, Psychology & Social Care, University of Derby, Derby

¹¹ Radiology Department, University College London Hospitals

¹²School of Allied and Public Health Professions, Canterbury Christ Church University

¹³RAICQ

¹⁴ NHS Greater Glasgow and Clyde, Glasgow

#Joint first authors.

Corresponding Author:

Professor Alex Novak

Director of Emergency Medicine Research Oxford (EMROx)

Consultant in Emergency Medicine and Ambulatory Care

Oxford University Hospitals NHS Foundation Trust

Oxford

Abstract:

Introduction:

A non-contrast CT head scan (NCCTH) is the most common cross-sectional imaging investigation requested in the Emergency Department (ED). 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 dataset 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 sub-classified into extradural, subdural, subarachnoid, intraparenchymal, and intraventricular.

30 readers will be recruited across four NHS trusts including 10 general radiologists, 15 Emergency Medicine clinicians, and five 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 13/12/2022). 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.

Keywords:

Artificial intelligence, computed tomography, brain, clinical decision support systems

Article Summary

Strengths and limitations of this study

- This study will evaluate 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 will be the first UK-based multicentre validation of an AI for NCCTH trained on a large dataset (300,000 head CTs).
- It includes non-radiologists (EM 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.

(Word count: 2408 words)

Main text

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 upon their findings for plain radiography, but are generally dependent on radiologists for more complex modalities such as Computed Tomography (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 upon 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 datasets [23–25]. The AI tools which have been developed using the largest datasets (above 100,000 scans) have originated from China and India [11,12] and have not been validated in the UK. The geographic 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 datasets show improved performance in this regard [11,12].

qER 2.0 EU is an FDA cleared and CE 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 US, 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 will indicate that the software has analysed the image and identified no target abnormalities.

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 dataset. 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 stand-alone 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, e.g. 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 multi-reader multi-case (MRMC) design.

Case selection:

150 NCCTH of ED patients aged 18 years or above will be retrospectively identified by the clinical and PACS / 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 9 of interest listed above, as well as normal scans.

Inclusion criteria for cases:

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

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 artifacts such as burr holes, shunts, or clips
- Scans containing metal artifacts

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))
- 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

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 9 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 vs 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). In addition to the original scan, the qER output will include a notification to suggest the presence/absence of a target abnormality and segmentation of the abnormal areas identified. 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 dataset) 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 Area Under the Receiver Operating Characteristic Curve (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 vs without AI assistance, will also be derived from the reader classifications. The stand-alone diagnostic performance of qER AI will be evaluated by estimating sensitivity and specificity of qER in classifying a scan as critical vs 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 vs 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, 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 [28]. Since we made few assumptions, a higher sample size of 150 NCCTH (90 with critical findings, 60 without any critical findings) will be used for the study.

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 2-way mixed effects ANOVA model treating readers and cases (images) as random effects and effect of AI as a fixed effect [29] with recommended adjustment to degrees of freedom by Hillis et. al. [30]. 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 vs 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.

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 utilised for the AUC analyses, while binary classification results will be utilised for evaluation of sensitivity, specificity, positive predictive value, and negative predictive value.

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 (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 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:

All authors contributed to the writing of the protocol and reviewed the manuscript, with AN and SA leading the overall design. HF, AN, DR, and SA wrote the manuscript. KB and AR are the ground-truthers.

Patient and public involvement

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. One of the study authors, JG, serves as PPI lead.

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 Swetha Tanamala are employees of Qure AI.

Nick Woznitza declares consultancy fees from InHealth and SM Radiology not related to the current submission.

References

- 1 Juszczuk 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. doi:10.1111/ans.15285
- 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. doi:10.1016/j.crad.2019.10.025
- 4 Richards M. Diagnostics: Recovery and Renewal – Report of the Independent Review of Diagnostic Services for NHS England. NHS England 2022. <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. Royal College of Radiologists 2020. <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. doi:10.1016/j.carj.2018.10.004
- 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. doi:10.1016/j.amjmed.2011.06.023
- 8 Lin E, Yuh EL. Computational Approaches for Acute Traumatic Brain Injury Image Recognition. *Front Neurol* 2022;**13**:791816. doi:10.3389/fneur.2022.791816
- 9 Sheth SA, Giancardo L, Colasurdo M, *et al.* Machine learning and acute stroke imaging. *J Neurointerv Surg* 2023;**15**:195–9. doi:10.1136/neurintsurg-2021-018142
- 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. doi:10.1136/neurintsurg-2020-017099
- 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. doi:10.1016/S0140-6736(18)31645-3
- 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. doi:10.1016/S2589-7500(22)00090-5
- 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. doi:10.1038/s41598-020-77441-z

- 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. doi:10.1038/s41746-017-0015-z
- 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. doi:10.1067/j.cpradiol.2020.10.007
- 16 National Institute for Health and Care Excellence. Artificial intelligence for analysing CT brain scans. 2020. <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. doi:10.1161/STROKEAHA.121.036204
- 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 (Basel)* 2022;**12**:452. doi:10.3390/diagnostics12020452
- 19 Warman R, Warman A, Warman P, *et al.* Deep Learning System Boosts Radiologist Detection of Intracranial Hemorrhage. *Cureus*;14:e30264. doi:10.7759/cureus.30264
- 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. doi:10.1159/000339116
- 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. doi:10.5812/traumamon.12023
- 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* Published Online First: 2022. doi:10.2139/ssrn.4062896
- 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. doi:10.1007/s00234-021-02826-4
- 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. doi:10.1016/j.jstrokecerebrovasdis.2022.106702
- 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. doi:10.1016/j.crad.2022.10.017
- 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 Medicine* 2018;**15**:e1002683. doi:10.1371/journal.pmed.1002683

- 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**:1–16. doi:10.1038/s41746-023-00811-0
- 28 Obuchowski NA. Sample size tables for receiver operating characteristic studies. *AJR Am J Roentgenol* 2000;**175**:603–8. doi:10.2214/ajr.175.3.1750603
- 29 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. doi:10.1080/03610919508813243
- 30 Hillis SL, Obuchowski NA, Scharz KM, *et al.* A comparison of the Dorfman–Berbaum–Metz and Obuchowski–Rockette methods for receiver operating characteristic (ROC) data. *Statist Med* 2005;**24**:1579–607. doi:10.1002/sim.2024

Figure legends

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

Figure 2: Case selection flow diagram.

Figure 3: Reader study flowchart.

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 1-10. B) Ground truthers will be asked to rate the difficulty of each diagnosis on a Likert scale from 1-5.

Figure 5: data flow diagram.

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.