Software with artificial intelligence-derived algorithms for analysing CT brain scans in people with a suspected acute stroke: a systematic review and cost-effectiveness analysis

Marie Westwood,[1*] Bram Ramaekers,[2] Sabine Grimm,[1] Nigel Armstrong,[1] Ben Wijnen,[1] Charlotte Ahmadu,[1] Shelley de Kock,[1] Caro Noake[1] and Manuela Joore[2]

[1]Kleijnen Systematic Reviews (KSR) Ltd, York, UK
[2]Department of Clinical Epidemiology and Medical Technology Assessment, Maastricht University Medical Centre (MUMC), Maastricht, Netherlands

[*]Corresponding author marie@systematic-reviews.com

Disclosure of interests

Full disclosure of interests: Completed ICMJE forms for all authors, including all related interests, are available in the toolkit on the NIHR Journals Library report publication page at https://doi.org/10.3310/RDPA1487.

Primary conflicts of interest: None declared.

All authors have completed the unified competing interest form at www.icmje.org (available on request from the corresponding author) and declare: (1) no financial support for the submitted work from anyone other than their employer; (2) no financial relationships with commercial entities that might have an interest in the submitted work; (3) no spouses, partners or children with relationships with commercial entities that might have an interest in the submitted work and (4) no non-financial interests that may be relevant to the submitted work.

Published March 2024
DOI: 10.3310/RDPA1487

Scientific summary

Software with artificial intelligence-derived algorithms for analysing CT brain scans in people with a suspected acute stroke: a systematic review and cost-effectiveness analysis

Health Technology Assessment 2024; Vol. 28: No. 11
DOI: 10.3310/RDPA1487

NIHR Journals Library www.journalslibrary.nihr.ac.uk

Scientific summary

Background

The primary population for this assessment was people presenting or attending secondary care with a suspected acute stroke who were last known to be well within the previous 24 hours. Stroke is a serious life-threatening medical condition defined by the World Health Organization (WHO) as a clinical syndrome consisting of 'rapidly developing clinical signs of focal (at times global) disturbance of cerebral function, lasting more than 24 hours or leading to death with no apparent cause other than that of vascular origin'. Timely and effective management of the patients with suspected stroke substantially impacts patients' outcomes.

A number of software products with artificial intelligence (AI)-derived software technologies have been developed, which are intended to facilitate the review of computed tomography (CT) images of the brain in patients with suspected stroke. These products are not intended to provide a diagnosis but to support review and reporting healthcare professionals.

Objectives

This assessment aimed to evaluate the clinical and cost-effectiveness of using AI-derived software to support the review of CT brain scans in acute stroke, in the NHSs setting. Three research questions were considered.

(1) Does AI-derived software-assisted review of non-enhanced CT brain scans for guiding thrombolysis treatment decisions for people with suspected acute stroke represent a clinically and cost-effective use of NHS resources?
(2a) Does AI-derived software-assisted review of CT angiography (CTA) brain scans for guiding mechanical thrombectomy treatment decisions for people with an ischaemic stroke represent a clinically and cost-effective use of NHS resources?
(2b) Does AI-derived software-assisted review of CT perfusion brain scans for guiding mechanical thrombectomy treatment decisions for people with an ischaemic stroke after a CTA brain scan represent a clinically and cost-effective use of NHS resources?

Methods

Assessment of clinical effectiveness

Twenty-five databases, including MEDLINE and Embase, research registers, conference proceedings and a preprint resource, were searched for relevant studies from inception to July 2021; update searches were conducted in October 2021. Search results were screened for relevance independently by two reviewers. Full-text inclusion assessment, data extraction and quality assessment were conducted by one reviewer and checked by a second. The methodological quality of included diagnostic test accuracy studies was assessed using QUADAS-2 (Bristol Medical School, University of Bristol, Bristol, UK). The methodological quality of observational 'before and after' studies was assessed using a checklist, devised by the authors, for this review.

The hierarchical summary receiver operating characteristic (HSROC) model was used to estimate summary sensitivity and specificity with 95% confidence intervals (CIs) and prediction regions around the summary points, and to derive HSROC curves for meta-analyses of diagnostic test accuracy, where

four or more studies evaluated the same intervention for a given research question. All other results, including those of 'before and after' studies, were summarised in a narrative synthesis, grouped by research question addressed, AI-derived software evaluated and study type.

Assessment of cost-effectiveness

The health economic analysis focused on research question 2a:

(2a) Does AI-derived software-assisted review of CT angiography brain scans for guiding mechanical thrombectomy treatment decisions for people with an ischaemic stroke represent a clinically and cost-effective use of NHS resources?

All diagnostic accuracy studies identified by the systematic review conducted for this assessment assessed the accuracy of AI-derived software technologies as stand-alone interventions. As a result, information about how AI-derived software technologies would perform when used as an adjunct/aid to human readers (i.e. as recommended by the manufacturers, as specified for this assessment and as they would be used in clinical practice) is lacking. This is because the accuracy of the device by itself tells us nothing about how, or indeed whether, it might improve the accuracy of a human reader. It would not make sense to infer that any of the variation in sensitivity observed between stand-alone AIs can tell us something about precisely the variation in a hypothetical, small improvement in sensitivity of the human reader. To perform cost-effectiveness analysis (CEA), we elicited expert opinion to estimate the diagnostic accuracy of AI as adjunct to human reader. Experts were provided with the evidence on AI alone and human reader alone. Because it was considered too difficult for experts to differentiate between different AI-derived software-assisted review technologies, AI-derived software-assisted review in general (not specified by manufacturer or specific technology) is considered.

The de novo model (developed in R Shiny, R Foundation for Statistical Computing, Vienna, Austria) consisted of a decision tree (short-term) and a state transition model (long-term) to calculate the mean expected costs and quality-adjusted life-years (QALYs) for people with ischaemic stroke and suspected large-vessel occlusion (LVO).

The decision tree was used to estimate short-term costs and consequences (first 90 days). Subsequently, patients with LVO were classified as either eligible for thrombectomy or not eligible. Those with both LVO and eligibility for thrombectomy were further classified, based on the sensitivity of the diagnostic strategy, into whether a LVO was detected (and thus thrombectomy received) or not. Based on the classification in the decision tree, patients were subdivided into health states according to the modified Rankin Scale (mRS). Notably, patients without LVO were subdivided based on the specificity of the diagnostic strategy into whether a LVO was incorrectly detected or not. If a LVO was incorrectly detected (i.e. false positive), this had cost consequences only (e.g. due to potential unnecessary transfer to experienced stroke centre qualified to perform thrombectomy). The long-term consequences in terms of costs and QALYs were estimated using a state transition cohort model with a lifetime time horizon (annual cycle length) and health states defined as per mRS states.

Probabilistic sensitivity analyses, deterministic sensitivity analyses and scenario analyses were performed.

Results

Assessment of clinical effectiveness

A total of 22 studies (30 publications) were included in the review; for 9 of the 13 manufacturers of AI-derived software included in the scope, no studies were identified. All included studies concerned AI-derived software produced by Avicenna, Brainomix, iSchemaView or Viz. The majority (18/22 studies) reported data concerning research question 2a (i.e. evaluated AI-derived software for the interpretation of CTA). All included studies either assessed the diagnostic accuracy of

AI-derived software alone (i.e. *not* as it would be used in clinical practice, as recommended by the manufacturers and as specified in the inclusion criteria for this assessment) or were 'before and after' observational studies reporting information about the effects of implementing AI-derived software in treated patients.

Eleven studies provided information about the accuracy of various AI-derived software technologies for the detection of LVO on CTA scans in patients with acute ischaemic stroke. Where the target condition included occlusions of internal carotid artery, carotid terminus or the M1 or M2 segments of the middle cerebral artery (MCA), the sensitivity and specificity estimates were 95.4% (95% CI 92.7% to 97.1%) and 79.4% (95% CI 75.8% to 82.6%) for Rapid CTA (iSchemaView, Menlo Park, CA, USA), 91.2% (95% CI 77.0% to 97.0%) and 85.0 (95% CI 64.0% to 94.8%) for Viz LVO, 83.8% (95% CI 77.3% to 88.7%) and 95.7% (95% CI 91.0% to 98.0%) for Brainomix e-CTA, and 98.1% (95% CI 94.5% to 99.3%) and 98.2% (95% CI 95.5% to 99.3%) for Avicenna CINA LVO, based on one study each. There was some evidence to indicate that, where studies included more distal (e.g. M3 segment of the MCA) elements of the anterior circulation or included posterior circulation in their definition of the target condition, sensitivity was reduced. All four studies that provided information about the effects of implementing Viz LVO and one study that provided information about the effects of implementing Rapid CTA reported that implementation was associated with reductions in time to treatment for thrombectomy patients and, where reported, with no significant change in clinical outcomes (mRS). However, it should be noted that two of the studies of Viz LVO and the study of Rapid CTA evaluated implementation in the context of providing an automated alert system (i.e. *not* as specified in the scope for this assessment); it is plausible that reductions in time to intervention, observed in these studies, may be driven by this 'early alert' step. The information provided by studies of this type is also limited in that it concerns only treated (i.e. test positive) patients; no information is provided about test negative patients and hence there is no information about the extent to which AI-derived software, as implemented, may miss patients with LVO.

There is no evidence about the accuracy of AI-derived software when used as an aid to human interpretation; all evidence concerns only stand-alone AI. This might imply that a CEA is not feasible for any of the three research questions. However, we conducted a CEA in relation to the research question 2a, where there is most evidence about the performance of AI-derived software technologies alone and one study comparing an AI-derived software technology alone with human reader alone. These studies were not considered appropriate to inform cost-effectiveness modelling but formed the basis by which the accuracy of AI plus human reader could be elicited by expert opinion.

Assessment of cost-effectiveness

Base-case analysis

The probabilistic results indicated that the addition of AI to detect LVO is potentially more effective (QALY gain of 0.003), more costly (increased costs of £8.61) and cost-effective for willingness-to-pay thresholds of £3380 per QALY and higher. The cost-effectiveness plane illustrated the negative correlation between incremental costs and incremental QALYs; that is if a technology is more effective it also tends to be less costly. The cost-effectiveness acceptability curve indicated that at willingness-to-pay values of £20,000 and £30,000 per QALY gained the probabilities of current practice with AI being cost-effective are 54% and 56%, respectively. The expected risks per patient associated with adding AI at willingness-to-pay values of £20,000 and £30,000 per QALY gained are £80 and £95, respectively (these were £122 and £163 respectively without adding AI; see expected loss curves). At a population level (assuming 87,635 patients imaged, per year, in the UK), the estimated annual risks associated with adding AI are £7.0 million and £8.4 million, at willingness-to-pay values of £20,000 and £30,000 per QALY gained, respectively.

Secondary analysis sensitivity and scenario analyses

Sensitivity analyses indicated that the sensitivity of both technologies (i.e. with and without the addition of AI-derived software-assisted review) was the most important input parameter. In addition, the proportion of patients with LVO that are eligible for mechanical thrombectomy is important to determine the most optimal strategy in terms of costs and QALYs. For the estimated costs (specificity), the additional costs of the AI technology, costs related to mRS 4 and mRS 5 were input parameters (in addition to those mentioned above), which can change the strategy that is most expensive. Consistently, the most influential scenario analyses were related to the sensitivity (for both strategies), the proportion of patients with LVO eligible for mechanical thrombectomy with AI, removing the general population mortality cap and the additional costs of the AI technology.

Conclusions

The available evidence is not suitable to determine the clinical effectiveness of using AI-derived software to support the review of CT brain scans in acute stroke.

The economic analyses did not provide evidence to prefer the AI-derived software strategy over current clinical practice. However, results indicated that if the addition of AI-derived software-assisted review for guiding mechanical thrombectomy treatment decisions increased the sensitivity of the diagnostic pathway (i.e. reduced the proportion of undetected LVOs) this may be considered cost-effective. Nevertheless, the sensitivity of AI-derived software-assisted review when added to current clinical practice is largely uncertain and probably depends on the implementation of AI-derived software-assisted review.

Study registration

This study is registered as PROSPERO CRD42021269609.

Funding

This award was funded by the National Institute for Health and Care Research (NIHR) Evidence Synthesis programme (NIHR award ref: NIHR133836) and is published in full in *Health Technology Assessment*; Vol. 28, No. 11. See the NIHR Funding and Awards website for further award information.

Health Technology Assessment

HTA TAR

ISSN 2046-4924 (Online)

Impact factor: 3.6

A list of Journals Library editors can be found on the NIHR Journals Library website

Launched in 1997, *Health Technology Assessment* (HTA) has an impact factor of 3.6 and is ranked 32nd (out of 105 titles) in the 'Health Care Sciences & Services' category of the Clarivate 2022 Journal Citation Reports (Science Edition). It is also indexed by MEDLINE, CINAHL (EBSCO Information Services, Ipswich, MA, USA), Embase (Elsevier, Amsterdam, the Netherlands), NCBI Bookshelf, DOAJ, Europe PMC, the Cochrane Library (John Wiley & Sons, Inc., Hoboken, NJ, USA), INAHTA, the British Nursing Index (ProQuest LLC, Ann Arbor, MI, USA), Ulrichsweb™ (ProQuest LLC, Ann Arbor, MI, USA) and the Science Citation Index Expanded™ (Clarivate™, Philadelphia, PA, USA).

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: journals.library@nihr.ac.uk

The full HTA archive is freely available to view online at www.journalslibrary.nihr.ac.uk/hta.

Criteria for inclusion in the *Health Technology Assessment* journal
Manuscripts are published in *Health Technology Assessment* (HTA) if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

HTA programme

Health Technology Assessment (HTA) research is undertaken where some evidence already exists to show that a technology can be effective and this needs to be compared to the current standard intervention to see which works best. Research can evaluate any intervention used in the treatment, prevention or diagnosis of disease, provided the study outcomes lead to findings that have the potential to be of direct benefit to NHS patients. Technologies in this context mean any method used to promote health; prevent and treat disease; and improve rehabilitation or long-term care. They are not confined to new drugs and include any intervention used in the treatment, prevention or diagnosis of disease.

The journal is indexed in NHS Evidence via its abstracts included in MEDLINE and its Technology Assessment Reports inform National Institute for Health and Care Excellence (NICE) guidance. HTA research is also an important source of evidence for National Screening Committee (NSC) policy decisions.

This manuscript

The research reported in this issue of the journal was commissioned and funded by the Evidence Synthesis Programme on behalf of NICE as award number NIHR133836. The protocol was agreed in July 2021. The assessment report began editorial review in January 2022 and was accepted for publication in December 2022. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' manuscript and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this manuscript.

This manuscript presents independent research funded by the National Institute for Health and Care Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, the HTA programme or the Department of Health and Social Care. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, the HTA programme or the Department of Health and Social Care.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Newgen Digitalworks Pvt Ltd, Chennai, India (www.newgen.co).