NICE National Institute for Health and Care Excellence

Evidence generation plan for artificial intelligence (AI) software to help clinical decision making in stroke

Implementation support

Published: 23 January 2024
Last updated: 2 May 2024

www.nice.org.uk

Contents

1 Purpose of this document

NICE's assessment of artificial intelligence (AI) software to help clinical decision making in stroke recommends that further evidence is generated for e-Stroke (Brainomix), RapidAI (Ischemaview) and Viz (Viz.ai), while they are being used in the NHS. Other AI software is recommended only for use in research and is not covered in this plan.

This plan summarises the evidence gaps and what real-world data needs to be collected for a NICE review of the technologies again in the future. It is not a protocol but suggests an approach to generating the information needed to address the evidence gaps.

The Department of Health and Social Care and NHS England have launched several initiatives that will support the generation of more evidence for the use of AI technologies. This includes:

- The AI Diagnostic Fund: In June 2023, DHSC announced funding for the creation of a ring-fenced £21 million AI diagnostics fund. DHSC, the National Institute for Health and Care Research (NIHR) and NHS England are collaborating to support 'in-service evaluations'. These evaluations, alongside plans for national collation of data and metrics from the deployments across multiple imaging networks, aim to answer the evidence gaps set out in the guidance and evidence generation plan.
- The AI Deployment Platform: DHSC is piloting a platform to help deploy multiple AI imaging technologies in radiology, in 2 NHS imaging networks. As part of this work, a mechanism will be set up to support the post-market surveillance of these AI models in clinical practice.

The companies are responsible for ensuring that data collection and analysis takes place. An approach to evidence generation is through the formation of a consortium bringing analytical partners and implementation sites together with technology developers.

Guidance on commissioning and procurement of the technologies will be provided by NHS England, who are developing a digital health technology policy framework to further outline commissioning pathways.

NICE will withdraw the guidance if the companies do not meet the conditions in section 4 on monitoring.

When suitable evidence has been generated, the developers should submit the evidence to NICE in a form that can be used for decision making. NICE will review all the evidence and assess whether the technologies can be routinely adopted in the NHS.

2 Evidence gaps

This section describes the evidence gaps, why they need to be addressed and their relative importance for future committee decision making.

The committee will not be able to make a positive recommendation without the essential evidence gaps (see section 2.1) being addressed. The company can strengthen the evidence base by also addressing as many other evidence gaps (see section 2.2) as possible. This will help the committee to make a recommendation by ensuring it has a better understanding of the patient or healthcare system benefits of the technology.

2.1 Essential evidence for future committee decision making

The impact of AI-derived software on a healthcare professional's ability to identify people for whom thrombolysis and thrombectomy is suitable

There is limited evidence on the impact of using AI software alongside healthcare professional interpretation to detect relevant features such as large vessel occlusions or intracerebral haemorrhage, and making decisions about use of thrombolysis and thrombectomy.

The impact of the software on how many people have thrombolysis or thrombectomy

Current evidence about the impact of AI software on how many people have thrombolysis or thrombectomy is limited. It is also confounded by issues such as lack of clarity about whether study populations were comparable before and after introduction of AI software, and the unknown influence of other changes to the care pathway around the time of implementation, for example, because of the COVID-19 pandemic. Further evidence, minimising the limitations and confounding issues affecting the current evidence, is needed to support future committee decision making.

The impact of using the software on time to thrombolysis or thrombectomy

The available evidence suggested that time to treatment with thrombolysis or thrombectomy reduced with the introduction of AI software. But, all studies were retrospective and limited. Further evidence is needed comparing time to treatment with and without AI software, accounting for confounding factors such as ring-fencing stroke beds and increasing staff numbers. This should also consider the impact on time to treatment for people transferred to other centres for thrombectomy, and whether image sharing functionality was used to facilitate this.

2.2 Evidence that further supports committee decision making

How often the software is unable to analyse CT brain scans and reasons for this

Software failure could delay diagnosis and access to time-sensitive treatments. However, only 1 study (Kauw et al. 2020) reported technical failure outcomes for any AI software and clinical experts advised that failure in clinical practice may be higher than the 11% reported. Evidence is needed to establish how often each AI software is unable to guide treatment decisions in stroke, and the reasons for this.

3 Approach to evidence generation

The Getting it Right First Time (GIRFT) review of stroke services noted that AI software is already widely used across stroke centres in the NHS, with estimates showing that 96% have access to it. The remaining centres are expected to be using it by the end of 2023, and it is possible that new centres implementing the software may choose to complete local studies or audits.

3.1 Evidence gaps and ongoing studies

The ongoing Health Innovation Oxford & Thames Valley study of e-Stroke, due to report in March 2024, uses data from the Sentinel Stroke National Audit Programme (SSNAP). This compares the time periods before and after AI software was introduced, and may address evidence gaps relating to the impact of AI on time to treatment, and numbers of people having thrombolysis or thrombectomy.

Several additional studies have been identified that may partially address the impact of the technology on treatment decision making and time to treatment. But these studies are limited by small sample sizes and only including known stroke cases.

Table 1 Evidence gaps and ongoing or newly identified studies

Evidence gap	e-Stroke	RapidAI	Viz
The impact of AI-derived software on a healthcare professional's ability to identify people for whom thrombolysis and thrombectomy is suitable	Limited evidence Ongoing studies	Limited evidence	No evidence
The impact of the software on how many people have thrombolysis or thrombectomy	Limited evidence Ongoing study	No evidence	No evidence
The impact of the software on time to thrombolysis or thrombectomy	Limited evidence Ongoing studies	Limited evidence	Suitable evidence

Evidence gap	e-Stroke	RapidAI	Viz
How often the software is unable to analyse CT brain scans, with reasons for this	No evidence Newly identified study	Limited evidence	Limited evidence

3.2 Data sources

There are several data collections that have different strengths and weaknesses that could support evidence generation. NICE's real-world evidence framework provides detailed guidance on assessing the suitability of a real-world data source to answer a specific research question.

SSNAP is a comprehensive dataset covering all NHS stroke centres in England. It records data on inpatient care, outcomes and interventions for up to 6 months after a person has had a stroke. The Core Dataset collects patient-level data including the date, time, and modality of first imaging, whether AI supported interpretation of these images, what treatment was given and when. It is feasible that this could be used to address some evidence gaps. But this dataset does not currently report which AI technology was used, or whether the software was unable to analyse an image. It may be possible to modify the Core Dataset to record this additional information on a per-image basis, but this could take up to 2 years. Alternatively, these modifications could more quickly be included in the Comprehensive Dataset for specific sites of interest. They would not be mandatory so engagement would be needed with sites to promote and encourage completion.

The Diagnostic Imaging Dataset, which collects monthly extracts from local Radiology Information Systems, may provide useful stroke imaging data such as type of CT requested and time between request and reporting.

Linking routine data sources has been considered in previous studies and is viable but may be challenging.

The quality and coverage of real-world data collections are of key importance when used in generating evidence. Active monitoring and follow up through a central coordinating point is an effective and viable approach of ensuring good-quality data with broad coverage.

3.3 Evidence collection plan

Before evidence generation, a user survey across all NHS stroke centres in England is proposed to elicit information about their use of AI software in the stroke pathway. This should include which software they use, when it was implemented, which staff groups or healthcare professionals use it, and on which CT scans (or where in the care pathway) it is used. Some of this information may be available directly from NHS England. From this, sites can be identified for the proposed studies to best represent acute stroke care in the NHS, and its variation across centres, to address confounders and produce the most generalisable evidence possible.

The suggested approaches to addressing the evidence gaps for AI software in stroke are an experimental concordance study with existing imaging data, plus evaluation of SSNAP data. How these approaches will address each evidence gap is considered, and any strengths and weaknesses highlighted.

Concordance study

A concordance study is used to assess the agreement between 2 or more methods. The evidence gaps that this will address are:

- The impact of the addition of AI-derived software on a healthcare professional's ability to identify people for whom thrombolysis and thrombectomy is suitable.
- How often the software is unable to analyse CT brain scans, with reasons for this.
- The impact of using the software on time to thrombolysis or thrombectomy (partially, since only speed of review can be assessed).

Each patient case would include clinical data available at the time of scanning, and may include unenhanced CT, CT angiography (CTA) and CT perfusion (CTP) images, for review alongside each other, as appropriate according to standard care.

This study will assess the concordance between the treatment decision reached for each included case of suspected stroke by:

- Healthcare professionals assisted by AI software (intervention).
- Healthcare professionals unassisted by AI software (comparator).

- Consensus assessment by an expert panel, unassisted by AI software (reference standard).

Data should be selected carefully to be representative of existing sites, considering responses to the user survey. This is to make sure acute and comprehensive stroke centres, relevant staff roles and experience levels, and specific AI software are represented.

Representative image sets would be provided by stroke centres, anonymised, and processed by e-Stroke, RapidAI, Viz, and no AI software. They would then be presented to, and assessed retrospectively by, recruited healthcare professionals with appropriate qualifications and experience, recommending no treatment, thrombolysis, or thrombectomy. Cases that the software was unable to analyse would be recorded.

If every reader assessed every case, using every AI software and none, this would be a full factorial design. But a pragmatic approach to optimise healthcare resource usage is a 'split plot' design where each reader assesses a subset of cases, but each case is assessed using each AI software and none, enabling a comparison. The sample size (number of patient cases, and number of readers) should account for the following factors, and patient cases will be randomly allocated to readers to ensure each factor is fairly represented:

- AI used: e-Stroke, RapidAI, Viz, none
- staff role: radiologist and physician responsible for making treatment decisions
- years of experience in staff role.

The factors above would be controlled for in the design of the study, and the influence of other, explanatory, variables could be considered in cases of discordance. These variables may account for particular groups of interest, such as people aged over 80, people with cerebrovascular disease, specifications of different scanners used, type of stroke centre, and types of CT imaging included in the case.

Comparison between AI-assisted (intervention), and unassisted (comparator) readings, and the expert panel (reference standard) would allow assessment of concordance for each AI software, used as intended. By including 2 staff groups (radiologists and physicians), concordance could be measured within each group, and between groups. Discordant cases could be further explored to identify common characteristics, and reasons for discordance could be suggested by the expert panel.

A secondary outcome of this study would be the time spent assessing images to reach a treatment decision in each case, with and without AI assistance. Although this timing activity lacks some real-world validity by not reflecting all steps in the care pathway, this element of the proposed study may highlight differences in read times between different staff groups, and different experience levels.

Other concordance study designs are possible, which could better represent real-world use of the software, but may have less internal validity or need greater data collection.

Evaluation of SSNAP data

The evidence gaps that this will address are:

- The impact of using the software on time to thrombolysis or thrombectomy.
- The impact of using the software on how many people have thrombolysis or thrombectomy.
- How often the software is unable to analyse CT brain scans, with reasons for this.

The work supported by Health Innovation Oxford & Thames Valley is using SSNAP data to capture use of e-Stroke across 26 sites in England. Because of variation in how time to outcome data was collected before April 2021, use of a 'before-and-after' study design is limited. But because sites implemented AI software at different time points, time between CT scan and treatment decision could be monitored over time to determine whether this decreased as use of AI increased. Although the work of Health Innovation Oxford & Thames Valley is restricted to e-Stroke, its methodology could be applied to studies of other software, including RapidAI and Viz.

A random selection of cases could be extracted from SSNAP. Completion of the data field relating to the use of AI software could be used as a factor in subsequent analysis. Time since AI implementation and the AI software in use at each site (taken from the user survey) could also be treated as factors. Additional data fields may be needed to address how often the software is unable to analyse CT brain scans, and the reasons why, which could be added to the Comprehensive Dataset for specific sites of interest.

3.4 Data to be collected

The following data should be prioritised for collection within each of the studies described

above.

Information to be collected by surveying all stroke centres using AI software

The following data items should be collected for each site:

- AI software used, including version and date of implementation.
- Description of the pathway followed by people suspected of having an acute stroke, including which CT images are taken and when, and which staff role reviews these.
- Specification of CT scanner (or scanners) used.

Concordance study

- Treatment decisions made using relevant clinical information and CT scans with and without AI support, and by the expert panel.
- Whether or not the AI software was able to process each image.
- Time spent assessing CT scans for each case, plus relevant clinical information, with and without AI support.

Evaluation of SSNAP data

The following data items should be extracted, or calculated from the SSNAP dataset, for periods before and after AI implementation:

- time between CT scan and thrombolysis (if indicated)
- time between CT scan and arterial puncture for thrombectomy (if indicated)
- number and proportion of people having treatment with thrombolysis
- number and proportion of people having treatment with thrombectomy
- number and proportion of people referred to another site for thrombectomy
- how often the software is unable to analyse CT brain scans, with reasons for this.

A subgroup analysis could consider the difference in time to treatment for those referred to another site for thrombectomy, or shorter or longer time after stroke symptom onset. For outcomes lacking comparable data from a 'before' phase, analysis could be repeated using quarterly extracts from SSNAP to determine whether increased use led to changes over time.

Data collection should follow a predefined protocol and quality assurance processes should be put in place to ensure the integrity and consistency of data collection. See NICE's real-world evidence framework, which provides guidance on the planning, conduct, and reporting of real-world evidence studies.

3.5 Evidence generation period

Because 99 of 107 stroke units in England are already using AI technologies, with the rest expected to implement AI software by the end of 2023, it is feasible that sufficient robust evidence could be generated within the next 3 years.

4 Monitoring

The companies must contact NICE:

- within 6 months of publication of this plan to confirm agreements are in place to generate the evidence
- annually to confirm that the data is being collected and analysed as planned.

The companies should tell NICE as soon as possible of anything that may affect ongoing evidence generation, including:

- any substantial risk that the evidence will not be collected as planned
- new safety concerns
- the technology significantly changing in a way that affects the evidence generation process.

If data collection is expected to end later than planned, the companies should contact NICE to arrange an extension to the evidence generation period. NICE reserves the right to withdraw the guidance if data collection is delayed, or if it is unlikely to resolve the evidence gaps.

5 Implementation considerations

The following considerations around implementing the evidence generation process have been identified through working with system partners:

- There may be variation in the care pathway at different centres. For example, different types of CT scan may be used depending on local preference. The most effective method of recruiting readers for the experimental concordance study should be considered. This should ensure representation across acute stroke centres and comprehensive stroke centres, all staff groups assessing CT images in stroke, and ensuring that those recruited are truly representative of the wider groups fulfilling the same role.
- Evidence should be generated in such a way that it facilitates ongoing or future assessment of AI software in stroke (for example, supporting repeatability for future updates or other technologies) in line with standard 16 of NICE's Evidence standards framework for digital health technologies.
- The SSNAP dataset only collects data for confirmed cases of stroke. So, even if it is modified to collect data relating to the AI software being unable to analyse CT images, this may not be generalisable to negative cases, which may have features on CT imaging that make it more or less likely that the software will fail.
- It is important to note that certain outcomes from the implementation of AI software that would influence its accuracy or time taken to reach a decision, cannot be easily measured. For example, readers may have different reliance on AI depending on their experience.
- Even if AI software speeds up the isolated task of reading images, this may not reduce the time to thrombolysis or thrombectomy, because of unavoidable delays elsewhere in the stroke pathway.
- Baseline use of thrombolysis and thrombectomy varies across the NHS, potentially because of process factors and attitudes to judging suitability for thrombolysis. Qualitative work by Allen et al. (2022) discusses further contributing factors, and a clinical expert noted that access to treatment differs by site, and by time of day. So, AI software is only one factor to influence changes in treatment rates, and other sources of variance should be considered during evidence generation.

- CT perfusion imaging may guide referrals, especially between 6 and 24 hours after symptom onset, when salvaging brain tissue is less likely. This is increasingly available at acute stroke centres, but analysis support may be needed from comprehensive centres, which could influence uptake of thrombectomy.

ISBN: 978-1-4731-7437-5