NICE National Institute for Health and Care Excellence

Evidence generation plan for artificial intelligence (AI) technologies to help detect fractures on X-rays in urgent care

Implementation support

Published: 14 January 2025

www.nice.org.uk

Contents

1 Purpose of this document

NICE's early value assessment of artificial intelligence (AI) technologies to help detect fractures on X-rays in urgent care recommends that BoneView, Rayvolve, RBfracture and TechCare Alert can be used in the NHS while more evidence is generated. The other AI technology considered in the guidance can only be used in research and is not covered in this plan.

This plan outlines the evidence gaps and what data needs to be collected for a NICE review of the technologies again in the future. It is not a study protocol but suggests an approach to generating the information needed to address the evidence gaps. For assessing comparative treatment effects, well-conducted randomised controlled trials are the preferred source of evidence if these are able to address the research gaps.

The companies are responsible for ensuring that data collection and analysis takes place.

Guidance on commissioning and procurement of the technologies will be provided by NHS England.

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

After the end of the evidence generation period (2 years), the companies 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

Diagnostic accuracy

To evaluate the efficacy of these technologies, it is essential to have further understanding about their diagnostic accuracy in urgent care settings that reflects the technology users in the NHS. Higher diagnostic accuracy for the presence or absence of fractures can minimise risks and costs associated with incorrect or delayed treatment.

Clinical and service outcomes

To assess the impact of these technologies on healthcare delivery and patient health, it is essential to measure outcomes related to both clinical effectiveness and efficiency of urgent care services. To evaluate clinical effectiveness, data collection should focus on changes in patient outcomes associated with reduced misdiagnosis rates and the impact of missed fractures. To provide further understanding on service efficiency, evidence should be collected on whether the technology can improve the service, influence decisions, and reduce the need for additional imaging and onward referral to fracture clinics.

2.2 Evidence that further supports committee decision making

Effectiveness in different subgroups

There was limited evidence for the use of the technologies in certain subgroups. Analysing the data collected to consider these groups will help the committee to understand the benefits of the technologies to broader populations. Subgroups to consider include:

- age (for example, children and young people)
- sex
- ethnicity
- socioeconomic status
- fracture types
- conditions that affect bone health (for example, myeloma, osteoarthritis, osteoporosis, osteogenesis imperfecta, Paget's disease, rickets, osteomalacia and metastatic bone disease).

People with conditions that affect bone health, and some people with joint replacements, may have different healthcare needs or present additional diagnostic challenges to healthcare professionals.

Costs associated with implementing the AI technologies

Collecting data on the costs associated with establishing the infrastructure for AI technologies is important for understanding the financial investment that is needed. It will also provide understanding of the feasibility and sustainability of integrating AI technologies into routine healthcare. This information can contribute to estimates of cost effectiveness.

3 Approach to evidence generation

3.1 Evidence gaps and ongoing studies

Table 1 summarises the evidence gaps and ongoing studies that might address them. Information about evidence status is derived from the external assessment group's report; evidence not meeting the scope and inclusion criteria is not included. The table shows the evidence available to the committee when the guidance was published.

Table 1 Evidence gaps and ongoing studies

Evidence gap	BoneView	Rayvolve	RBfracture	TechCare Alert
Diagnostic accuracy	Evidence is available Ongoing study	Limited evidence available	Limited evidence available Ongoing study	Limited evidence available Ongoing study
Clinical and service outcomes	Limited evidence available Ongoing study	Limited evidence available	Limited evidence available Ongoing study	No evidence Ongoing study
Effectiveness in different subgroups	No evidence	No evidence	No evidence	No evidence Ongoing study
Costs associated with establishing the infrastructure needed to implement the artificial intelligence (AI) technology	No evidence	No evidence	No evidence	No evidence

3.2 Data sources

Most of the data, particularly that relating to diagnostic accuracy, is likely best collected

through primary data collection. There are data sources that may collect some of the necessary outcome information, however, they will require linking to each other and the primary data collection.

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. Potential data sources include:

- NHS England's Diagnostic Imaging Dataset
- NHS England's Emergency Care Data Set
- NHS England's Hospital Episode Statistics.

The NHS picture archiving and communication system (PACS) will also be a useful resource.

Local or regional data collections such as NHS England's sub-national secure data environments (see the NHS England blog on Investing in the future of health research) could potentially be used to collect information and link data sources together. Secure data environments are data storage and access platforms that bring together many sources of data, such as from primary and secondary care, to enable research and analysis. The sub-national secure data environments are designed to be agile and can be modified to suit the needs of new projects, as would be necessary in this instance.

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 to ensure good-quality data with broad coverage.

3.3 Evidence collection plan

The suggested approaches to address the evidence gaps are an experimental concordance study with existing imaging data and a real-world prospective study.

Centres that best represent urgent care centres in the NHS and variation across centres (for example, patient volume and number of readers) should be included to address confounders and allow subgroup analyses. Sample populations should be representative and consider subgroups (for example, age, sex, ethnicity and socioeconomic status).

Concordance study to assess diagnostic accuracy

A concordance study is used to assess the agreement between 2 or more methods.

Each case should include clinical data available at the time the X-ray is taken in line with standard care for each of the methods being compared. This study will assess the concordance between the diagnosis reached for each included case by the:

- healthcare professional assisted by artificial intelligence (AI) technology (intervention)
- healthcare professional unassisted by AI technology (comparator)
- reference standard.

There are several potential approaches that can be taken to collect the reference standard:

- Panel of experts: A consensus assessment by an expert panel, unassisted by AI technology, but ideally with access to clinical information that would be available at the time the AI is intended to be applied. This is the ideal approach for a comprehensive assessment of both the AI technology and the reporting of experienced radiologists or radiographers.
- Arbitration process: An arbitration process designed to resolve disagreements in diagnoses between the AI technology and the results from healthcare professionals unassisted by the AI. This method helps to determine the final diagnosis when there is discordance between the 2 approaches. A limitation of this approach is that it does not collect specific information about sensitivity and specificity of the technology. Therefore, this information may need to be sourced through other methods. This is particularly important to help populate future economic models.
- Follow up: Monitoring of clinical progression to identify and assess any false negatives or false positives, ensuring that the accuracy of the initial diagnosis can be confirmed or corrected over time. This approach will be affected by differential verification bias and may require a variable follow-up period depending on each presenting case.

Prospectively collected anonymised image sets would be provided by emergency centres and processed to determine the diagnosis by the intervention and comparator, and the reference standard. Ideally, cases should be randomly allocated to readers to minimise potential bias.

Any cases that the technologies were unable to analyse should be recorded for further investigation. Discordant cases could be further explored to identify common characteristics, and reasons for discordance.

Comparison between AI-assisted (intervention), and unassisted (comparator) readings, and the experienced radiologist or reference standard would allow assessment of diagnostic accuracy. It is possible that linked clinical outcomes could also provide evidence of whether a fracture was missed by the AI technology or human review when the patient returned at a later date.

As part of data collection process, performance of the AI technology alone should be collected. Although this data is not directly relevant to how the AI technology would be deployed in the NHS, it enables separation of software and human components of performance, and will allow monitoring, updating and direct comparison of future technologies. The combined performance may be sensitive to change in training level of users, or unassisted diagnostic practices. Measurement of AI performance alone is a useful marker as a lower bound to identify drift from intended use because of automation bias.

The diagnostic accuracy should be also assessed in applicable subgroups, such as children and young people, and people with conditions that affect bone health. It is important to also consider readers with varying levels of experience.

Prospective real-world study

To address the evidence gaps, a prospective real-world study is suggested. Ideally, this would compare outcomes in a period before implementation of the technology to a period after deployment.

This study could be done at a single centre or, ideally, replicated across multiple centres to show how the technology can be implemented across a range of services, representative of the variety in the NHS. Some outcomes may reflect other changes unrelated to the interventions that occur over time in the population. To control for these changes over time that might occur anyway, additional robustness can be achieved by collecting data in a centre that has not implemented the technology.

High-quality data on patient characteristics may be needed to identify and correct for any important differences between comparison groups and to assess who the technologies

would not be suitable for. Important confounding factors should be identified with input from clinical experts during the protocol development.

Information to be collected in this study is detailed in section 3.4.

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.

3.4 Data to be collected

The following information has been identified for collection:

Diagnostic accuracy study

- Diagnoses made by the AI-assisted healthcare professional in urgent care, the unassisted healthcare professional in urgent care, and the experienced reviewer. Also, ideally, diagnoses by AI technology alone.
- Number and proportion of images not eligible for processing by the AI technology (for example, because of technical or software failure) and reasons.
- Performance of the different methods compared to the ground truth. Performance estimates should include overall accuracy, sensitivity, specificity, positive predictive value, negative predictive value and c-statistic. Number of true positives, false positives, true negatives and false negatives should also be reported.
- Performance of the different methods among different subgroups such as age, sex, ethnicity, socioeconomic status, fracture types and conditions that affect bone health.
- Cases of diagnostic disagreement and the likely reason for disagreement.
- Cases of missed fractures by the AI technology.
- Time spent on review, with or without the AI technology.

Prospective real-world study

- Clinical and service outcomes. These should be analysed considering factors such as type of fracture.
- Clinical outcomes associated with missed diagnosis or misdiagnosis, for example, unnecessary treatments, further diagnostic procedures, or complications from misdiagnosis, ideally with quality-of-life impact.
- The number and proportion of people being recalled to hospital after radiology review.
- Incidence of further injury or harm during the time between the initial interpretation and treatment decision in urgent care and the definitive radiology report.
- Total number of referrals to fracture clinics, and number and proportion of unnecessary referrals.
- Rate of detection of non-fracture-related conditions by the AI technologies or failure to detect non-fracture-related conditions highlighted by the reporting healthcare professional.
- Costs associated with establishing the infrastructure needed to implement the AI technologies.
- Costs associated with maintaining the infrastructure needed for the AI technologies, including software, hardware and staff training.
- Ongoing costs like system updates and technical support.

Information about the technologies

Information about how the technologies were developed, the update version tested, and how the effect of future updates will be monitored should also be reported. See the NICE evidence standards framework for digital health technologies.

3.5 Evidence generation period

This will be 2 years to allow for setting up and implementing the AI technologies, and for data collection, analysis and reporting.

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 Minimum evidence standards

The selection of the artificial intelligence (AI) technologies evaluated for this assessment was based on a limited evidence base gathered from company submissions, the external assessment group's review of the available literature and committee discussions.

The technologies were primarily assessed on diagnostic accuracy. Initial findings indicated a potential benefit in reducing missed fractures. However, this evidence had significant limitations such as risk of bias, small sample sizes and heterogeneity in study designs. Additionally, the evidence lacked consistency across different subgroups, such as children or people with conditions that affect bone health.

It is important to note that there was an absence of studies done in the NHS. This contributed to the decision for further data collection aimed at supporting the implementation of these technologies across urgent care settings in the NHS.

To have a deeper understanding of the full range of benefits these technologies can offer and to support their implementation, additional data is needed on diagnostic accuracy, efficacy in specific subgroups, clinical and service outcomes, and economic impact.

6 Implementation considerations

Companies should work with providers and central NHS England teams to begin the research. Planning for a prespecified period for the set-up of the technologies is advised. The following considerations around implementing the research process have been identified through working with system partners:

- the companies should provide training for staff on using the artificial intelligence (AI)-derived software
- services should be carefully selected to, when appropriate, maximise data collection for subgroups of interest
- ideally, a proportion of cases could be saved and sent for peer review as part of the prospective real-world study
- to assess the potential for automation bias after deployment, companies may want to track the rate of diagnostic disagreement over time
- all data transfer and processing should adhere to appropriate data protection legislation.

Potential barriers to implementation include:

- the availability of research funds for data collection, analysis and reporting
- the availability of NHS funding to cover the costs of implementing the technologies in clinical practice
- lack of expertise and staff to collect data
- burden on healthcare professionals, including the need to have training ahead of implementation, data collection and follow up
- differences in practice between NHS settings and the level of skills and experience that healthcare professionals have when reviewing X-rays.

ISBN: 978-1-4731-6736-0