NICE National Institute for Health and Care Excellence

Evidence generation plan for artificial intelligence (AI) technologies for assessing and triaging skin lesions referred to the urgent suspected skin cancer pathway

Implementation support

Published: 1 May 2025

www.nice.org.uk

Contents

1 Purpose of this document

NICE's healthtech guidance on artificial intelligence (AI) technologies for assessing and triaging skin lesions recommends that Deep Ensemble for Recognition of Malignancy (DERM) can be used in the NHS during the evidence generation period. It can be used as an option in teledermatology services to assess and triage skin lesions referred to the urgent suspected skin cancer pathway.

This plan outlines 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 study protocol but suggests an approach to collecting 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 company is responsible for ensuring that data collection and analysis takes place.

Guidance on commissioning and procurement of the technology 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 company does not meet the conditions in section 4 on monitoring.

After the end of the evidence generation period (3 years), the company 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 technology can be routinely adopted in the NHS.

2 Evidence gaps

This section describes the evidence gaps that need to be addressed for future committee decision making. The committee will not be able to make a positive recommendation without these being addressed.

2.1 Essential evidence for future committee decision making

How accurate DERM used in teledermatology services is at detecting cancer and non-cancer skin lesions compared with teledermatology services alone

Collecting data on the accuracy of DERM used in teledermatology services to detect cancer and non-cancer skin lesions compared with teledermatology alone is essential for determining whether it can provide assessments suitable for routine clinical settings. The data will help to determine whether DERM can effectively discharge non-urgent cases from the suspected skin cancer pathway while maintaining diagnostic accuracy for detecting high-risk lesions. This data will help evaluate the potential of DERM to enhance the diagnosis of cancer skin lesions and optimise clinical resources by reducing the burden on dermatology services. The data will also help to assess whether DERM can increase staff capacity and benefit people with non-cancer dermatological conditions. Additionally, this can help inform whether AI technologies can be used autonomously and whether this is reliable and safe.

Accuracy of DERM in people with black or brown skin

Collecting information about the accuracy of DERM to detect cancer and non-cancer skin lesions across different skin colours is vital for assessing potential biases in performance and for ensuring equitable healthcare. Skin tone should ideally be measured using spectrophotometry.

The effect of using AI technology in teledermatology services on the number of referrals for face-to-face dermatology

appointments compared with established teledermatology services alone

Collecting data on the number of face-to-face dermatology referrals generated by AI technology compared with well-established teledermatology services alone is crucial for understanding the effect of AI technology on healthcare workflows. This information will help determine if AI can reduce unnecessary referrals and save dermatologist time, thereby optimising resources and improving timely access to care for people. Additional information about the number of referrals to dermatology before and after implementation of AI technology will provide further understanding of its impact.

3 Approach to research

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. Table 1 shows the evidence available to the committee when the guidance was published. No ongoing studies were identified in the external assessment group's report.

Table 1 Evidence gaps and ongoing studies

Evidence gap	Deep Ensemble for Recognition of Malignancy (DERM)
How accurate DERM used in teledermatology services is at detecting cancer and non-cancer skin lesions compared with teledermatology services alone	Limited evidence
Accuracy of DERM in people with black or brown skin	Limited evidence
The effect of using AI technologies in teledermatology services on the number of face-to-face dermatology appointments compared with a well-established teledermatology service alone	Limited evidence

3.2 Data sources

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.

Some data will be generated through the technology itself, such as the number of referrals that were assessed by the technology and the diagnostic outcomes predicted by it. This data can be integrated with other data collected.

The NHS England Secure Data Environment service could potentially support this research. This platform provides access to high-standard NHS health and social care data that can be used for research and analysis. Local or regional data collections such as NHS

England's sub-national secure data environments (see the blog on 'Investing in the future of health research') and databases like NHS England's National Cancer Registration and Analysis Service already measure outcomes specified in this plan. They could be used to collect data to address the evidence gaps. 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.

Datasets that are taken from general-practice electronic health records with broad coverage, such as the Clinical Practice Research Datalink and The Health Improvement Network could be used to provide individual patient-level data. These could provide some useful information on referrals, diagnostic outcomes and patient characteristics.

The quality and coverage of real-world data collections are of key importance when used in research. 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

Diagnostic accuracy study

A diagnostic accuracy study is used to assess the agreement between 2 or more methods. The study would assess the agreement between the diagnosis decision reached for each included case of suspected cancer by:

- AI technology alone (intervention)
- teledermatology unassisted by AI technology (comparator)
- a reference standard.

Potential approaches that can be taken to collect the reference standard include:

- A panel of experts or expert opinion: Ideally a consensus assessment by an expert panel or, if resources are limited, assessment by an experienced healthcare professional unassisted by AI technology, but ideally with access to clinical information that would be available at the time of intended AI use. This is the ideal approach for a

comprehensive assessment of both the AI technology and teledermatology.

- 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 can be affected by differential verification bias and may require a considerable follow-up period.

Representative image sets would be generated prospectively. These would then be processed by the AI technology within teledermatology services. Cases that the technology was unable to assess would be recorded. It is important to consider variation in skin colour as part of the study design, for example, ensuring a sufficient sample size to assess different skin colours (ideally measured using skin spectrophotometry).

A comparison between the AI technology alone (intervention), the teledermatology unassisted by AI (comparator) and the reference standard would allow an assessment of the diagnostic accuracy of the AI technology compared with teledermatology services. Cases with disagreements in the diagnosis between each method could be further explored to identify common characteristics, and reasons for disagreements could be considered.

For pragmatism, this study could be done as part of the 'before' in the before-and-after study. Care would need to be taken to control for confounders and blinding. This could reduce the time to collect evidence and ensures the accuracy values are representative of the setting.

Real-world before-and-after implementation study

The results of the accuracy study should inform the population for a before-and-after implementation study, so that the technology is implemented in populations in which it has been shown to be effective.

A before-and-after study design allows for comparisons when there is considerable variation between services in the standards and mode of delivery of teledermatology. It also allows assessment of implementation costs, changes in referral rates, and the proportion of cases that are eligible for assessment by the AI technology.

Before the AI technology is implemented in a teledermatology service, data should be collected on the:

- total number of referrals to that service
- number of those referrals that resulted in a face-to-face appointment with a dermatologist
- number of biopsies
- number of referrals that resulted in a cancer lesion diagnosis.

If teledermatology is already established then the number of lesions that are not eligible for assessment by this service should also be recorded and the reasons why. The AI technology should then be implemented into the service and all implementation and training costs should be collected. After leaving a period of time to account for learning effects, the outcomes on referral rates, appointments and biopsies should be collected again in a period after implementation. The number of lesions that are not eligible for assessment by the AI technology should also be collected in the after-implementation study, and the reason why.

In a phased approach, a comparison between the AI technology's diagnosis and a dermatologist's opinion, and ideally, the final clinical outcome, can help to predict the likely impact of autonomous use of the AI technology before moving into, and testing fully autonomous use.

This study could be done at a single centre with an established teledermatology service or ideally, replicated across multiple centres. This could show how the AI technology can be implemented across a range of services, representative of the variety in the NHS. Outcomes may reflect other changes that occur over time in the population, unrelated to the interventions. Additional robustness can be achieved by collecting data in a centre that has not implemented an AI technology but is as similar as possible (in terms of clinical practice and patient characteristics) to a service where an AI technology is being used or ideally, a stepped-wedge design. This could help control for changes in referral rates over time that might have occurred anyway.

3.4 Data to be collected

The following information has been identified for collection:

Diagnostic accuracy study

- Classifications made using teledermatology unassisted by AI technology, and by AI technology alone, and by the reference standard.
- Proportion of lesions discharged from the urgent suspected skin cancer pathway onto the non-urgent pathway.
- Information on lesions that are not eligible for assessment by teledermatology and not eligible for assessment by AI technology, and the reasons.
- Whether or not the AI technology was able to process each image.
- Performance of the AI technology and teledermatology compared with the reference standard.
- Accuracy in people with black or brown skin (ideally measured using skin spectrophotometry).
- Cases of diagnostic disagreement and the likely reason for disagreement (given by reference standard).

Real-world before-and-after implementation study

- Patient information, for example, age, sex and ethnicity.
- Number and proportion of suspected skin cancer cases that are not eligible for teledermatology before implementation of the technology and the reasons why.
- Total number of referrals through the urgent suspected skin cancer pathway.
- Number and proportion of referrals that had appointments with a dermatologist.
- Number and proportion of appointments with a dermatologist that resulted in a biopsy or diagnosis of a cancer lesion.
- For the after-implementation period, the number and proportion of suspected cancer cases that are not eligible for assessment by the technology.
- The number and proportion of suspected cancer cases that are judged to be 'indeterminate' or cannot be processed by the AI technology (technical failure and rejection rate).

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.

Information about the technology

Information about how the technology was 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.

Evidence generation period

This will be 3 years to allow for setting up and implementing the AI technologies, and for data collection, analysis and reporting.

4 Monitoring

The company 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 company 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 company 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 evidence considered by NICE primarily focuses on diagnostic accuracy, clinical outcomes and the impact of the technology on the teledermatology pathway for urgent suspected skin cancer referral. The technology demonstrated potential to support triage for cancerous and non-cancerous lesions. Nevertheless, more information is still needed to fully understand the benefits it may provide. More evidence is needed on:

- diagnostic accuracy, and comparison with existing teledermatology services
- diagnostic accuracy in people with black or brown skin
- the impact on the care pathway.

This evidence is essential to future NICE decision making. It may also potentially inform whether AI technologies can be used autonomously to discharge non-cancer cases while sustaining safe diagnostic accuracy rates.

6 Implementation considerations

The company should work with providers and central NHS England teams to begin the research. Planning for a prespecified period for the set-up of the technology is advised. During this period, training and implementation should be done before data collection is started, to account for learning effects. The following considerations around implementing the research process have been identified through working with system partners:

- For safe implementation, developers could initially do 'silent evaluation' (see Kwong et al. 2022) before full deployment into services. This approach deploys the technology without any influence on clinical decision making until the technology is fully deployed. This approach can be used to understand whether the technology can be deployed safely (including in subpopulations), what the influence on decision making would likely have been (for example, onward referrals), and may collect some relevant data items (for example, test failure rate or number of indeterminate findings).
- The company should provide training for staff in using the technology.
- Services should be carefully selected to, when appropriate, maximise data collection for subgroups of interest.
- To assess the potential for automation bias after deployment, the company may want to track the rate of diagnostic disagreement over time.

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 technology in clinical practice
- lack of expertise and staff to collect data
- burden on clinical staff, including the need to have training ahead of implementation, data collection and follow up
- differences in practice between primary care settings across the NHS
- differences in skill and experience among staff when using the AI technology.

ISBN: 978-1-4731-7744-4