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 early value assessment 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 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.

The company is responsible for ensuring that data collection and analysis take place. NICE may 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, 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

Resource and care pathway impact

More data is needed about how using DERM affects the care pathway, compared with teledermatology alone, including:

- referral rates to dermatology from primary care
- impact on workload (for example, number of face-to-face appointments and biopsies, referrals to non-urgent pathways, patients discharged)
- time to diagnosis or discharge
- general costs related to the use of the technology
- proportion of lesions eligible for assessment by DERM.

Ideally, information about the impact on system indicators such as waiting times could also be collected.

This information will help provide a better understanding of whether the technology can help improve efficiency and provide benefits to the NHS.

Data should be collected when the technology is implemented as an autonomous tool and

when it is used with healthcare professional review. This data will support a better understanding of how effective the technology is when used in routine NHS practice.

Accuracy of DERM in people with black or brown skin

Collecting more information on the accuracy of DERM in people with black or brown skin (Fitzpatrick skin types 5 and 6) is necessary to assess the effectiveness of the technology across different skin tones.

2.2 Evidence that further supports committee decision making

Comparative analysis of the accuracy of DERM and teledermatology

It is important to understand how DERM performs compared with teledermatology. More information about the accuracy of teledermatology is needed to support this comparison and further data about the accuracy of DERM will enhance future analysis.

Understanding how well DERM or teledermatology discharges non-urgent cases from the suspected skin cancer pathway while maintaining diagnostic accuracy for detecting high-risk lesions will inform an understanding of the impact on the workload for dermatology services.

Subgroup analyses of the accuracy of DERM when it is implemented as an autonomous tool and with healthcare professional review will support a clearer understanding of the effectiveness of the technology.

3 Approach to evidence generation

3.1 Evidence gaps and ongoing studies

Table 1 summarises the evidence gaps and the status of evidence addressing them.

Table 1 Evidence gaps and status of evidence

Evidence gap	Deep Ensemble for Recognition of Malignancy (DERM)
Resource and care pathway impact	Limited evidence
Accuracy of DERM in people with black or brown skin	Limited evidence
Comparative analysis of the accuracy of DERM and teledermatology	Good indirect 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 may 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 can be integrated with other data collected.

Some data may be generated as part of post-market surveillance activities done by the technology manufacturer. This may include data relating to the number of referrals seen, or the proportion of outcomes or performance data for the technology in comparison with a pre-defined ground truth.

Local or regional data collections such as NHS England's secure data environments and databases like NHS England's National Cancer Registration and Analysis Service (NCRAS) 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.

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 to ensure good-quality data with broad coverage.

3.3 Evidence collection plan

Two potential methodological approaches are presented in this section. Both have their respective strengths and weaknesses and, depending upon the circumstances in which evidence is being generated, either may be the better approach.

Data should be collected that reflects the following different ways that the technology can be implemented:

- as an autonomous tool, and
- used with a healthcare professional review.

Real-world comparative cohort study

In this type of study, data should be collected from healthcare services where the artificial intelligence (AI) technology is offered and compared with services where it is not. People in both groups should be followed from the point at which they would typically be offered the AI technology.

The comparison group should include teledermatology services with comparable patient populations and standard care pathways but without access to the AI technology. Ideally, the study should be done across multiple centres to reflect the diversity of the NHS service provision.

Non-random assignment to interventions introduces a risk of confounding bias. So, appropriate methods such as matching or adjustment (for example, propensity score methods) should be used to minimise selection bias and balance confounding factors between groups. High-quality data on patient characteristics will be essential to support these methods. The identification of key confounders should be informed by expert input during protocol development.

Real-world before-and-after implementation study

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 about the service, for example:

- 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.

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 should be collected again. 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.

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.

3.4 Data to be collected

The following information has been identified for collection:

- patient demographics: age, sex, ethnicity, Fitzpatrick skin type (or other validated classification scale)
- lesion characteristics, for example, melanoma or squamous cell carcinoma (SCC)
- referral volumes to and from teledermatology services

- impact on workload, for example, numbers of biopsies, appointments and face-to-face appointments, and healthcare professional time
- number of lesions identified as benign and discharged
- time to diagnosis or discharge
- costs associated with implementing and using the technology, for example, set-up costs, staff needed and training
- diagnostic accuracy, for both DERM and teledermatology alone
- cases of diagnostic disagreement between DERM and current NHS practice and, ideally, reasons. For example, lesions that DERM was not able to assess or those identified as a cancerous lesion by DERM that were not identified by teledermatology
- site characteristics and data that can support matching or adjustment analysis
- ideally, system indicators such as waiting times.

Data collection should follow a pre-defined 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.

3.5 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.

3.6 Following best practice in study methodology

Following best practice when doing studies is paramount to ensuring the reliability and validity of the research findings. Following rigorous guidelines and established standards is

crucial for generating credible evidence that can improve care. The NICE real-world evidence framework details some key considerations.

In the context of evidence generation, it is important to consider as part of the informed consent process that patients (and their carers, as appropriate) understand that data will be collected to address the evidence gaps in section 2. Where applicable this should take account of NICEs guidance about shared decision making.

4 Monitoring

NICE will contact the company:

- 6 months after 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

During the evidence generation period, new technologies may become available. This section summarises the minimum evidence requirements that a new technology would need to meet to be considered in the NICE evaluation after the evidence generation period.

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, but the current evidence base is not sufficient to fully recommend using it in routine NHS practice. More evidence is needed on:

- the impact of the technology on the care pathway, including referrals, workload and costs associated with its implementation
- accuracy of the technology in people with black or brown skin (Fitzpatrick skin types 5 and 6)
- diagnostic accuracy of the technology.

This evidence is essential to future NICE decision making. The evidence should be collected when the technology is implemented in the following different ways:

- as an autonomous tool, and
- used with a healthcare professional review.

Data about the impact of the technology when used in a pre-referral setting was outside of the scope of the assessment but is of interest to the clinical community so would be useful.

6 Implementation considerations

When the technology is implemented, steps should be taken to mitigate the potential risk of missed or delayed cancer diagnoses when using DERM during the evidence generation period by:

- doing a healthcare professional review for people with black or brown skin
- regular monitoring of DERM's performance to maintain accuracy.
- using additional protocols when necessary, such as:
  - a national governance framework to ensure local oversight of use of DERM
  - a healthcare professional review.

The company should work with NHS commissioners and providers to implement the technology in real-world settings. 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. NHS England has set up an independent data oversight group that will support evidence generation.

An inclusive user-led design approach should be followed, with individuals across key stakeholder groups at implementation sites brought together at the start to map the pathway and plan for implementation.

Care should be taken to ensure that:

- lesions in people with black or brown skin have a second read by a healthcare professional throughout the evidence generation period
- people with disabilities are supported to access the teledermatology service
- information is provided in:
  - clear, plain language to enable informed consent to be given
  - alternative languages, or local translation services are used as needed.

For the implementation of the real-world study, efforts should be made to select sites with

a range of different characteristics to ensure generalisability of the results. Site characteristics could include:

- type of centre, for example, community diagnostic centres versus hospital sites
- racial or ethnic diversity of the population
- image takers, for example, medical photographers versus healthcare assistants
- local IT systems
- local dermatologist workforce (for example, vacancy rate and seniority mix).

For safe implementation, when DERM is first introduced at a site, specialist dermatologists may wish to use the tool with a second read for an agreed number of interactions. This approach can be used to understand whether the technology can be deployed safely and what the influence on decision making would likely have been (for example, onward referrals). It may also collect some relevant data items (for example, test failure rate or number of indeterminate findings).

For people with black or brown skin (Fitzpatrick skin types 5 and 6), a second read by a healthcare professional should take place.

The company should provide training for staff in using the technology as well as ongoing support. 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:

- hesitation of clinical teams to adopt new technology
- the availability of NHS funding to cover the costs of implementing the technology in clinical practice
- burden on clinical staff, including the need to have training ahead of implementation, data collection and follow-up
- differences in dermatology pathways across the NHS
- differences in skill and experience among staff when using the AI technology.

7 Update information

March 2026:

- The 'comparative analysis of the accuracy of DERM and teledermatology' evidence gap was re-categorised from 'essential evidence for future committee decision making' to 'evidence that further supports committee decision making'.
- The 'accuracy of DERM in people with black or brown skin' evidence gap was updated to specify Fitzpatrick skin types 5 and 6.
- Several sections were updated to include data collection 'when the technology is implemented as an autonomous tool and when it is used with healthcare professional review'.
- The implementation section was updated to align with the NHS England Implementation Toolkit.

ISBN: 978-1-4731-9398-7