NICE National Institute for Health and Care Excellence

Evidence generation plan for artificial intelligence (AI)-derived software to analyse chest X-rays for suspected lung cancer in primary care referrals

Implementation support

Published: 28 September 2023

www.nice.org.uk

Contents

1 Purpose of this document

NICE recommends that artificial intelligence (AI)-derived software to analyse chest X-rays alongside clinician review for suspected lung cancer should only be used in research. Further evidence is needed to assess the risk and benefits of the technology in routine care.

This plan outlines the evidence gaps for the technology and what real-world data needs to be collected for a NICE review of the technology again in the future. It is not a study protocol.

The technology developers 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 developers.

The Department of Health and Social Care (DHSC) and NHS England have launched several initiatives that will support the generation of more evidence for 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. One area of focus of the fund is AI to support radiologists to read chest X-rays. DHSC, the National Institute for Health and Care Research (NIHR) and NHS England are collaborating to support the winning trusts to do '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. This may include those for chest X-rays. As part of this work, a mechanism will be set up to support the post-market surveillance of these AI models in clinical practice.

Guidance on commissioning and procurement of the technology will be provided by NHS England. NHS England is developing a digital health technology policy framework that will further outline commissioning pathways.

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 on evidence that is essential to allow the committee to make a recommendation in future) being addressed. The company can strengthen their evidence base by also addressing as many other evidence gaps (see section 2.2 on evidence that further supports committee decision making) as possible. Addressing these will help the committee to make a recommendation by better understanding the patient or healthcare system benefits of the technology.

2.1 Evidence that is essential to allow the committee to make a recommendation in future

Referrals to CT scan

The review of chest X-rays determines whether people will proceed to have a chest CT scan. This may be influenced by the assessed technologies. To understand their impact on resource use, it is necessary to understand how the software affects the proportion of people with chest X-rays who are referred on to chest CT scan and the overall number of referrals to CT scan.

Time to chest X-ray review, CT referral, and diagnosis

An advantage of the software may be in supporting quicker review and reporting of chest X-rays, leading to quicker referral to chest CT scan and diagnosis. An additional advantage of using artificial intelligence (AI)-derived software to interpret images is that the algorithm can prioritise the images for the reviewer if abnormal findings are detected.

This can be assessed through measuring:

- time from chest X-ray to report

- average number of chest X-rays assessed per reviewer per day.

For those who are referred on to CT scan, it is important to assess how the software affects the time from receipt of chest X-ray to CT scan.

The benefits of the technology when used by less experienced trainee radiologists and reporting radiographers should also be considered when collecting data for this outcome.

Diagnostic accuracy and technical failure rates

AI-derived software may improve a chest X-ray reviewer's ability to identify images with features suggesting lung cancer. Improving sensitivity to abnormalities could result in earlier diagnosis, but unnecessary referrals to CT incur costs to the NHS and may cause anxiety.

Information on diagnostic accuracy (positive predictive value) could be assessed by measuring the proportion of abnormal chest X-rays that are confirmed as abnormal by the chest CT scan. Information on the number of cancers detected and missed, and stage of cancer at diagnosis, is also important to inform the economic model.

Technical failure and rejection rates should also be collected.

2.2 Evidence that further supports committee decision making

Software impact on healthcare costs and resource use

Further information on resources needed to implement these technologies in clinical practice is important for use in economic modelling to inform decision making. For example, training and software implementation costs.

Evidence in populations with underlying conditions that could yield images that are challenging to interpret

There is currently a lack of evidence for these technologies in people with conditions that may result in images that are challenging to interpret. Lung nodules and other abnormalities may be difficult to recognise in people with conditions such as asthma,

scoliosis, obesity, or chronic obstructive pulmonary disease.

Clinician experience of using AI-derived software

Information on ease of use and acceptability of the software by clinicians is needed. This may include experiences around implementing the technology and any improvements in the delivery of diagnostic services, particularly around accuracy of the technology in identifying abnormalities, appropriateness of image triage, and the impact on speed of review and reporting.

3 Ongoing studies

Table 1 summarises the available information on evidence gaps and ongoing studies that might address them. More information on the studies in the table can be found in the supporting documents.

Table 1 Summary of the evidence gaps and ongoing studies

Evidence gap	AI-Rad (Siemens Healthineers)	Red Dot (Behold.ai)	Lunit INSIGHT CXR (Lunit)	qXR (Qure.ai)
Impact of software on clinical decision making and number of people referred to have a CT scan	No relevant evidence identified	No relevant evidence identified	No relevant evidence identified	No relevant evidence identified Ongoing study
Time from chest X-ray review and report	No relevant evidence identified	Limited current evidence	No relevant evidence identified Ongoing study	No relevant evidence identified Ongoing study
Time to CT referral	No relevant evidence identified	No relevant evidence identified	No relevant evidence identified Ongoing study	No relevant evidence identified Ongoing study
Time to diagnosis	No relevant evidence identified	No relevant evidence identified	No relevant evidence identified	No relevant evidence identified Ongoing study

Evidence gap	AI-Rad (Siemens Healthineers)	Red Dot (Behold.ai)	Lunit INSIGHT CXR (Lunit)	qXR (Qure.ai)
Diagnostic accuracy, agreement, and technical failure rates	Limited current evidence	Limited current evidence	Limited current evidence Ongoing study	No relevant evidence identified Ongoing studies
Software impact on healthcare costs and resource use	No relevant evidence identified	No relevant evidence identified	No relevant evidence identified	No relevant evidence identified Ongoing study
Software performance for people with underlying conditions and high-risk groups	No relevant evidence identified	No relevant evidence identified	No relevant evidence identified	No relevant evidence identified
Clinician and patient perceptions on the use of artificial intelligence (AI)-derived software	No relevant evidence identified	No relevant evidence identified	No relevant evidence identified	No relevant evidence identified

Information about current evidence status is from the external assessment group (EAG) report. Evidence not meeting the scope and inclusion criteria is not included.

The following technologies did not have evidence that met the scope and inclusion criteria for these evidence gaps: Annalise CXR (annalise.ai), Auto Lung Nodule Detection (Samsung), Chestlink Radiology Automation (Oxipit), Chestview (GLEAMER), Chest X-ray (Rayscape), InferRead DR Chest (Infervision), Milvue Suite (Milvue), SenseCare-Chest DR Pro (Sensetime), and VUNO Med-Chest X-ray (VUNO).

4 Approach to evidence generation

An approach to addressing the evidence gaps through real-world data collection is considered, and any strengths and weaknesses highlighted.

Most technologies do not have ongoing studies that will address the evidence gaps, although Lunit INSIGHT CXR has ongoing research that may address some of the gaps. So, for these technologies, additional evidence generation is necessary.

qXR has ongoing research that may address all the essential and important evidence gaps and may not need additional evidence generation.

4.1 Evidence generation plan

For technologies lacking information about diagnostic accuracy and technical failure rates, diagnostic accuracy studies should be done to show this.

Other evidence gaps can be addressed through a real-world historical control study alongside a qualitative survey.

Diagnostic accuracy study

This could be done as a diagnostic cross-sectional study. The study would compare agreement between clinical reviewer alone and clinical reviewer aided by the software for identification of abnormal X-rays (needing CT follow-up). It would be possible to report accuracy (including sensitivity, specificity, negative predictive values and positive predictive values), variation across reviewers as well as technical failure rates.

Real-world historical control study

A historical control study could compare outcomes before and after the implementation of artificial intelligence (AI) software. This could assess the number and proportion of chest X-rays referred to CT scan, time from chest X-ray to completion of the report, number of chest X-rays assessed per reviewer per day, time from receipt of chest X-ray to CT scan report. The grade of NHS staff reviewing and reporting should also be collected.

This study could also collect additional diagnostic outcomes comparing AI-assisted review to reviewer alone. The study should assess whether abnormal findings on an X-ray correspond to disease-related abnormal findings on a follow-up CT scan (the reference standard). This would measure the positive predictive value aspect of diagnostic accuracy. Technical failure rates should also be reported. Information on number of cancers detected and stage of cancer at detection could be collected.

The study could also collect information on missed cancers among those who were not referred for chest CT during the study period, although this would give a biased estimate of false-negatives because not all missed cancers may be picked up over the observation period.

Data collection for each technology could be at a single centre or ideally across multiple centres. The study should also collect data on implementation costs for these technologies in routine clinical practice.

Qualitative survey

A qualitative survey is suggested to collect information on ease of use and acceptability of the software by clinicians. The format of the survey should include open-ended questions to give people the freedom to provide detailed insight. A range of views and perspectives should be collected that is representative of participating clinical reviewers at the sites where the technology is implemented.

4.2 Real-world data collections

The NHS England Secure Data Environment (SDE) service could potentially support evidence generation. This platform provides access to high standard NHS health and social care data that can be used for research and analysis. The Diagnostic Imaging Data Set within this service may be useful because it collects information about diagnostic imaging that people have and can be linked to other datasets.

There may be local or regional data collections that collect outcome measures specified in the research recommendation. The sub-national secure data environments could be a regional data collection alternative.

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 high coverage. NICE's real-world evidence framework also provides detailed guidance on assessing the suitability of a real-world data source to answer a specific research question.

4.3 Data to be collected

The following outcomes have been identified for collection through the suggested studies:

Quantitative

- time from chest X-ray to report
- time from chest X-ray to CT scan report
- time from chest X-ray to diagnosis
- number of chest X-rays reviewed per reviewer and centre per day
- of those who had a chest X-ray, the number and proportion of people referred to have a chest CT scan
- grade of NHS staff reviewing and reporting chest X-ray
- agreement between AI-derived software and clinician review for normal and abnormal interpretation of chest X-ray
- number and proportion of chest X-rays defined as abnormal confirmed as abnormal by CT
- number of cancers detected
- stage of cancer at detection
- number of cancers missed, that is, those initially not picked up as abnormal, later referred to chest CT in the study period, and any subsequent cancer diagnosis
- technical failure and rejection rates
- all training and software implementation costs
- characteristics of patients, including age, sex, weight and height or body mass index

(BMI), and comorbidities such as asthma, scoliosis, interstitial lung disease, chronic obstructive pulmonary disease (COPD), family background of lung cancer or young people who do not smoke.

Qualitative

- X-ray ease of use and acceptability
- perceived accuracy of the technology in identifying abnormalities
- perceived appropriateness of image triage
- perceived impact on speed of review and reporting
- perceived software's performance for people with underlying conditions and high-risk groups
- clinician perspective on the use of AI-derived software.

Other information

The company should describe the process for monitoring the performance of the technologies while they are used in clinical practice. See NICE's evidence standards framework for digital health technologies for guidance on post-deployment reporting of changes in performance. This should include:

- future plans for updating the technology, including how regularly the algorithms are expected to retrain, re-version or change functionality
- the sources of retraining data, and how the quality of this data will be assessed
- processes in place for measuring performance over time, to detect any impacts of planned changes or environmental factors that may impact performance
- processes in place to detect decreasing performance in certain groups of people overtime
- whether there is an independent overview process for reviewing changes in performance
- an agreement on how and when changes in performance should be reported and to

whom (evaluators, patients, carers and healthcare professionals).

The company should describe any actions taken in the design of the technology to mitigate against algorithmic bias that could lead to unequal impacts between different groups of people.

5 Implementation considerations

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

- Developers should provide training for staff in using the artificial intelligence (AI)-derived software.
- 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; the need to have a training ahead of its implementation, data collection and follow-up
  - differences in practice between large tertiary referral centres and smaller hospitals
  - differences in skill and experience among staff members when interpreting a report and defining severity of risk
  - possible governance issues because the software needs to send images and personal data to a cloud-based server before emitting a report
  - the software may not be compatible with other computer packages and different scanners used in the NHS (for example, NHS PACS [Picture Archiving and Communication System] systems)
  - the availability and ability of NHS information technology departments to install the software.

ISBN: 978-1-4731-7714-7