AI deployment fundamentals

Introduction

Artificial intelligence (AI) has the potential to revolutionise radiology, for example by streamlining workflow, prioritising cases for reporting and enhancing diagnostic accuracy. With the increasing number of commercial imaging AI solutions, national guidance is needed on how to deploy AI safely and effectively in the National Health Service (NHS) and how to gather the best evidence to support its use in radiology.

The Royal College of Radiologists (RCR) has been tasked with developing guidance for independent benchmarking of AI algorithms following the Healthcare Services Safety Investigation Branch 2021 report into 'Missed detection of lung cancer on chest X-rays of patients being seen in primary care'.[1] In parallel, the National Institute for Health and Care Excellence (NICE) has conducted an early value assessment on chest X-ray AI software for suspected lung cancer in primary care.[2] Of note, NICE has highlighted the lack of supporting data and the need for more research. NICE recommends that current access to AI technology should be restricted to research or non-core NHS funding; centres already using AI should do so under an appropriate clinical evaluation framework.

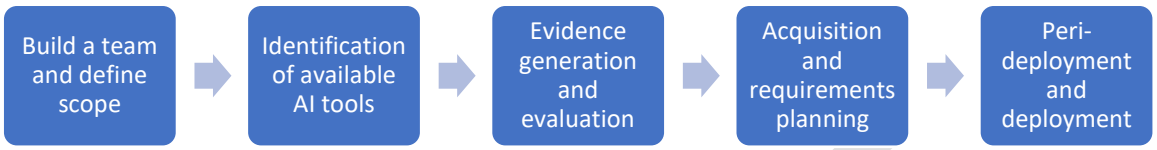
With the launch of the NHS England AI Diagnostics Fund (AIDF), this RCR guidance targets imaging networks and NHS trusts looking to evaluate and deploy AI solutions in radiology that have been certified and registered with the UK Medicines and Healthcare Regulatory Authority (MHRA). The guidance uses chest X-ray AI as an example but is broadly applicable to other certified AI solutions.

This guidance acknowledges the evidence gap and thus emphasises evidence generation and evaluation from the outset, given the need to ensure systems are safe and effective, from a risk–benefit and health economics perspective. However, the RCR recognises that appropriate infrastructure, expertise and funding are essential for formal evaluation and this will depend on the resources available to individual imaging networks. Post-market surveillance is outside the scope of this guidance.

This guidance is part of a wider RCR initiative to provide education in AI, share expertise and experience of using AI in radiology and shape the future of AI in healthcare. The

guidance was developed by an expert panel and incorporates feedback from a global expert reference group.

Key stages



1. Building a team to define the scope of the project

Selecting, deploying and evaluating AI solutions is a team effort, requiring input from and engagement by multiple stakeholders to ensure that the decisions made are clear and appropriate and have buy-in from everyone involved. This section sets out the stakeholders who should be engaged, the initial steps that should be taken to define the specific problem the AI is intended to address and its location in the pathway.

Engage stakeholders

1. Recruit a small working group of clinical pathway and imaging leads who are aware of current challenges in the pathway and service delivery and can see the potential opportunity for AI to deliver service improvement and enhance patient outcomes.

2. Engage in pre-market evaluation (see Glossary) by attending product demonstrations at regional and national events and arrange local demonstrations for the stakeholder group to consider the potential for conducting an AI evaluation project.

3. Engage with research and innovation leads to consider funding sources and opportunities to support AI projects, recognising that there may be insufficient evidence at first to commit core NHS funding through business cases. AI evaluation projects usually rely on industry, innovation or research funding applications.

4. Assemble a wider stakeholder group including clinicians, health professionals, clinical and operational managers (including clinical safety officers, chief clinical information officer, chief nursing information officers and/or chief information officer) and representatives from governance, patients, finance, IT and procurement to work up a project bid proposal, finance and implementation plan. Depending on the funding source this may include partnering with the supplier to submit a funding application, or it may require explaining the procurement approach.

Define the problem and pathway

5. Agree and clearly define the problem to be solved and its location in the pathway.

6. Agree the scope of the project and potential range of AI findings to be included, such as the detection or determination of:

    a. Normal versus abnormal

    b. Cancer only

c. Multiple pathology detection.

7. Consider where in the pathway AI is to be implemented (it may be at more than one site):

   a. Retrospective case finding for clinical audit and review

   b. Prioritisation of workflow and radiology reporting

   c. Clinical decision support (inside and outside of radiology).

8. Agree who will use the algorithm, considering they will need to be trained in its use.

9. Establish that the proposed algorithm is licensed to address the identified problem within the pathway and agree potential users with those involved in the pathway.[3] Review of the current National Pathway Guidance or NICE guidance should be included.

10. Include all those involved in the specific pathway in discussions on algorithm use to consider the potential impact of the algorithm on the pathway and patient care.

11. Agree on any changes to the clinical pathway enabled through the use of the AI, including fast-track referrals such as straight-to-test computed tomography (CT) and notification.

12. Recognise that introducing AI can be a catalyst for wider pathway improvement and potential benefits including report standardising, coding, report templating and clinical communication.

Document decisions

13. Document decisions made on the above points using the agreed template.

14. Record the current performance of the pathway prior to AI deployment, to enable comparison with post deployment. For example, for chest X-rays, depending on the intended purpose of the AI and problem to be addressed, this could include:

   a. Number of chest X-ray examinations performed on the specific pathway annually. If all chest X-rays, this should be categorised by referral source (eg general practitioner, emergency department) and into presentation and follow-up if possible.

   b. Number of those with prior chest X-rays – if this is difficult to determine then use a sample three-month period to estimate the annual number.

c. The time from chest X-ray referral (if not walk-in) to chest X-ray performed and reported.

d. Number already referred via the faster cancer diagnosis pathway.

e. Number of referrals onward for CT or clinic appointments and time from chest X-ray report to scan or appointment.

f. Number and proportion of normal and abnormal chest X-rays.

g. Number and proportion of cancers detected or missed on chest X-ray.

2. Identification of available AI tools

The RCR AI registry is currently under development and will aim to support the identification of possible AI applications and map to their intended purpose.

1. Review current sources of information about potential AI applications that are available, including:**Error! Bookmark not defined.**

   a. NHS guidance on the adoption of digital technology[4]

   b. AI for radiology database[5]

   c. NICE early value assessment for the use of AI to identify lung cancer on chest radiographs.[2]

2. Consider whether the potential AI applications will meet the needs of the outlined problem. A key part of implementation is that the tool's regulated intended purpose aligns with the problem that has been identified.

3. Request documentation describing intended use, functionality, limitations and possible risks from the manufacturer as required by regulations. A basic requirement for use in the NHS is a UK Conformity Assessed (UKCA) marking, which will have been provided through regulatory assessment by the MHRA. It is important to be aware that UKCA marking or regulatory approval does not necessarily equate to clinical effectiveness or cost-effectiveness in the proposed setting. A valid CE mark is still acceptable until 30 June 2028.

Understanding the evidence

4. Assess and evaluate the evidence base behind the tool.

5. Review the NICE medical technology evaluation programme or NICE early value assessment for medtech where these exist. This will outline the supporting evidence base, summarising the clinical effectiveness and cost-effectiveness for each technology.

6. Review published literature where NICE guidance is not available. This can be very challenging. It is important that the tool is shown to be effective in the population that has been proposed for the clinical question or role. It is also important that the application and any data have been assessed in a different cohort of patients to ensure that any results are reproducible and valid.

7. Explore national evidence on the available AI tools and consider whether they are from comparable populations and IT infrastructure.

8. Evaluate levels of evidence underpinning each of the available applications. It is likely the levels of evidence will increase as more are tested in the AIDF and other trusted research environments are developed, along with other forms of research in clinical pathways into the effectiveness and downstream effects of AI algorithms.

9. Assess the data for diversity to, as far as possible, minimise unknown biases in the tested applications. Diversity may be considered in terms of reflection of local population demographics or be related to protected characteristics, hard-to-reach groups and health inequalities.

Questions to assess

10. Discuss what the acceptable threshold level of performance will be in the setting or environment that is being proposed. It is very likely that performance will differ between training or testing and real-life clinical settings.[6,7]

11. Key questions to ask:

    a. Has the AI tool in question already been deployed in comparable healthcare organisations? A consultation with other trusts already using the tool could be considered.

    b. What is the status of integration with the local picture archiving and communication system (PACS) and radiology information systems (RIS)?

    c. How was the algorithm trained – has it been trained on representative patients and pathologies?

d. Is the performance acceptable – does it do the required task well?

e. Are the results generalisable – are the same results likely in your proposed population as those that have been tested?

f. Is local pre-deployment testing likely to be required?

g. What are the likely downstream effects of implementation? Consider the expected changes in existing pathways or services, whether services will be able to cope with those changes and the impact this may have on implementation.

Develop an evidence generation plan

12. Develop an initial local evidence generation plan and study protocol to assess whether the AI will deliver the anticipated benefits, including the impact on radiology services, staff and patient outcomes.

13. Consider the data collection requirements, key clinical performance indicators and local capability and resources to undertake the evaluation, recognising some data may need to be collected at baseline before AI implementation and to support the bid proposal.

14. Assess the AI tool's ability to support ongoing performance monitoring and data analytics, as well as its provision of audit data. Ongoing post-implementation evaluation is essential and requires a robust plan prior to deployment.

15. Evaluate this functionality based on its degree of automation to minimise the personnel effort required for data collection and auditing at a later stage.

16. Agree the process and frequency of product retraining and update.

17. Agree the procedures and process to follow in cases of immediate significant safety concerns. In this instance, product use should cease until remedial updates are available.

3. AI benefits evidence generation and evaluation

This section outlines the key areas to be addressed in the evaluation of AI tools in the clinical pathway.

1. AI should only be implemented in the NHS if the claimed accuracy has been confirmed and there is a clinical impact that is significant enough to justify using the product.

2. In addition to these, the impact of AI on the clinical pathway should be evaluated, including the impact on workflow, interaction with and acceptability for users, change of human decision-making and behaviour, what education is needed, monitoring of ongoing use, evaluation of updates, and acceptability for patients and the public.

Methods of evaluation

3. Methods used will depend on the focus of the evaluation of diagnostic accuracy and clinical impact. Two different approaches are recommended below.

Diagnostic accuracy study

4. A diagnostic accuracy study (see Glossary) is performed on a cohort of patients with the condition (eg lung cancer) to determine the sensitivity of the test and a separate cohort of patients without the condition to determine the specificity of the test.

5. For clinical AI studies this should include the baseline accuracy of reporters without AI and the post-implementation accuracy of using AI in clinical practice. For AI decision support this is the accuracy of the reporters supported by the AI.

6. The RCR has produced an audit template with advice on how to identify a cohort of patients with lung cancer and determine the sensitivity of reporters with chest X-ray, with recommendations for reviewing the missed cases.[8] This can be adapted for other diseases and conditions.

7. When assessing patients without the condition to determine the specificity of the test, it is important that the sample is representative of the referral population as this will include other conditions that may mimic the disease. A test with a low specificity will overcall the number of patients with the condition and may result in additional tests.

8. The ability to run AI in ‘shadow mode’ (see Glossary) enables the algorithm to be run over these cohorts retrospectively to determine the relative sensitivity and specificity. It is also possible to predict the effect of AI in clinical practice by reviewing the cases that are known to be missed by the reporters to assess the likely impact in clinical use.

9. AI platforms can assist in performing diagnostic accuracy studies on contemporary validated imaging data sets that have ground truth for the measured outcome. These might be a large series of chest radiographs that each have some sort of robust confirmation of what they show, such as biopsy-proven cancer or long follow-up with no adverse consequences. Provided the images exactly reproduce what the AI will be

applied to in the NHS and cover the diversity of patients and conditions, these data sets can be used to check accuracy of any AI algorithm and the updates very efficiently.

10. Once such data sets are developed and there are systems for them to be regularly updated, the risk of deploying a system that does not perform as claimed will be significantly reduced.

11. The data derived from this approach will provide information about likely impact on the clinical pathway, for example the effect on workflow of the number of false positives or areas where AI can potentially be autonomous, allowing the workforce to be deployed elsewhere according to service and patient need.

Longitudinal clinical impact study

12. Some elements of clinical impact may be modelled using diagnostic accuracy studies and platform-based evaluations, but for many clinical outcomes a longitudinal study (see Glossary) is required to assess whether the addition of AI is better than the existing standard.

13. A longitudinal study is a research design that involves repeated observations of the same variables over periods of time. These can be based on real-world data (RWD) collected through routine clinical practice and can be supplemented by additional data captured as part of the study protocol.

14. A real-world historical control study is a type of longitudinal study where the baseline performance is assessed using a retrospective study, a change is implemented (eg AI is introduced into the pathway) and then the performance is reassessed after an appropriate interval. This is analogous to a clinical audit cycle. A repeat service evaluation is the same as a clinical audit except there is no predefined performance standard.

15. NICE has recommended the collection of data through real-world historical studies to generate evidence for AI to analyse chest X-ray for suspect lung cancer in primary care referrals.[9] The NICE real-world evidence framework provides further guidance on the planning, conducting and reporting of RWD studies.[10]

16. A prospective cohort study is a type of longitudinal study that follows patients over time to see who develops the health outcome under consideration. This is typically set up as a clinical trial and involves following up patients who have the intervention (supported by

AI) and those who do not and are managed by current best practice. A randomised control trial (RCT) is a prospective cohort study that randomises patients to help minimise the effect of co-variate factors that may influence the outcome.

17. Prospective cohort trials are time-consuming and expensive and run the risk of using AI that has been superseded by the time the study reports.

18. Ideally studies should be designed in such a way that the measured outcome is agnostic to the AI and only depends on the functionality of the product in influencing the outcome. For example, if immediate use of AI shows a marked reduction in time to diagnosis of cancer (an important clinical outcome) then any AI produced that has confirmed accuracy at least as good as the product tested in the trial could be deployed.

19. In this way the two methods of evaluation can be used to rapidly deploy products in a way that is proven to be clinically effective. Although these trials also give data about accuracy, they are inefficient, slow, expensive and may be 'single use'.

Evaluation priorities

20. Many other elements of evaluation of AI exist (see Table 1) and they are all important, but the imperative now is to at least confirm the two principal elements: assessing diagnostic accuracy and measuring clinical impact.

Table 1. AI evaluation in the clinical pathway

Topic category	Brief description	Time of evaluation	Method of evaluation
Accuracy (and safety)	External evaluation of the accuracy of the product	Before deployment and at regular intervals	Validated external test platform
Clinical outcome	A change to an important clinical outcome	Before deployment but time-consuming	Randomised trial or cohort study or similar
Bias	AI can be associated with a variety of biases, some based on non-representative data and others on human behaviour	Before and during deployment	Test platforms should eliminate data bias and clinical bias Bias due to change of behaviour (eg

			automation bias) requires training
Workflow	AI can impact positively and negatively	Before and during deployment, with ongoing evaluation	Some information from platforms; in-service evaluations and modification to practice
Human–AI interaction	Humans are influenced by AI and it is important to maximise the benefits and avoid harms	During deployment with ongoing monitoring	Nested psychological experiments, educational intervention testing, testing of accuracy, ethnography
Education and training	Use of AI will evolve and clinicians need to stay up to date	Before and during deployment	Education sessions, surveys of use, audits and qualitative document review including training plans and training records
Patient and public acceptance	Patients and the public have a right to know how their data are used	Before and during deployment	Surveys and information provision based on concerns of focus groups
Cost-effectiveness	AI should be cost-effective in the NHS	Before and during deployment	Health economics evaluation at baseline followed by in-service evaluation

4. Acquisition and further requirements planning

Once AI solutions that have the potential to solve the specific problem have been identified, the next steps are to agree how a preferred AI tool will be acquired and to articulate the requirements of both the tool and the vendor in greater detail.

Acquisition

1. Identify the means of acquisition of the AI tool(s), such as tender, national procurement framework, trial, national award process. For example, the NHS (via Shared Business Services) has a procurement framework for stroke AI tools and is interested in

developing that for other AI tools. This provides a compliant route to market for suppliers. See also, for example, the NHS England procurement framework strategy recommendations.[11]

2. Consider the potential benefits and drawbacks of available funding routes. For example, the ability to enable collective and collaborative procurement (eg across imaging networks or health boards) may deliver value and unlock potential savings, but it may necessitate the involvement of more decision-makers.

3. Be clear on who needs to recommend or take decisions once options that meet initial stakeholder requirements have been considered.

Requirements planning

4. **Stakeholder fundamentals:** Agree requirements of all stakeholders and use these to create AI tool and supplier assessment criteria. Two fundamental considerations are:

    a. Ensuring that the supplier's statement of intended use aligns with the clinical problem to be solved that the stakeholder group has agreed.

    b. Inclusion of evidence of independent validation of the efficacy of the AI tool as part of the bid (see below for further detail). Example AI tool supplier assessment criteria are available in the NHS AI buyer's guide[12] (see Appendix 1).

5. **Organisational requirements:** Define and document what stakeholders need from the organisation to deliver the project successfully. For example:

    a. Trusts must ensure adequate clinical, technical and project support resources with time allocated to staff leading the acquisition and requirements planning stage.

    b. The project team must design the planning stage to enable tool acquisition that is deliverable (to time targets), affordable, aligned to the original scope or enables achievement of outcomes, and achieves sufficient clinical evidence and technical reassurance.

    c. Set clear intended shadow mode (see 'Shadow mode' below) and go-live deployment dates, with realistic project timelines and targets agreed with all stakeholders at the planning stage.

d. Describe any post-implementation evaluation to which the vendor will be required to contribute, together with the nature of that contribution. Include any issues that may affect the vendor while the evaluation is being conducted.

e. Ensure the expected duties and responsibilities of the trust and the vendor are clearly described so that resources can be scheduled accordingly.

6. **Technical requirements:** Ensure assessment criteria for the tool and the supplier incorporate the following technical requirements as a minimum (see Appendix 2):

- Pillar 3 – Software/SaaS/Apps – Clinical.[12]
- The AI tool uses the NHS number as patient identifier.
- The AI tool is MHRA compliant.
- The AI tool deploys a web-based or mobile application user interface.
- The supplier's handling of data is GDPR compliant.
- The supplier is Cyber Essentials certified or ISO 27001 certified.
- The supplier and/or tool comply with relevant NHS policies, including:
  - Public cloud first
  - Internet first
  - HL7 FHIR conformant supporting FHIR UK Core
  - HL7 FHIR Code System, Value Set and Concept Map including all operations
  - DCB0129 conformant
  - SNOMED CT conformant
  - ICD10 conformant
  - ODS conformant
  - WCAG 2.1 at AA level for any web-based or mobile user interfaces
  - Must align to National Cyber Security Centre (NCSC) cloud security principles.

7. **Integration requirements:** Identify how the tool needs to integrate with existing local systems such as IT networks and firewalls and existing security, PACS and RIS, and specify any limitations of the current IT infrastructure and potential requirements for additional resource.

8. **Training requirements:** Identify how the vendor will support training of all staff who have access to the AI findings, and how the vendor will work with the PACS supplier to limit access to the AI report to trained members only.

Data protection impact assessment

A data protection impact assessment (DPIA) will need to be completed as a standard part of project documentation and will require approval through the organisation's governance channels. The DPIA is helpful both for those procuring an AI solution and for vendors, and completing it helps clarify data flow requirements.

9. Map out the data flow and planned integrations at the acquisition and requirements planning stage, and feed this into the DPIA. Knowing which population you will use the tool for and where data will flow internally (eg remapping digital imaging and communications in medicine [DICOM] headers) and capturing this in a detailed data map is essential if the DPIA is to be approved.

10. Share your detailed data map with potential vendors. Integration of AI tools with the same PACS and RIS providers is challenging due to variations in local implementation and configuration. Making a detailed data map available at this stage will enable vendors to indicate whether they have managed similar implementations in the past and demonstrate specifically how they will achieve the requirements for any procurement process.

Independent validation

11. Identify whether independent validation will be required as part of the bid. Currently there is little independent evaluation of AI performance in chest X-rays. Options would likely include a combination of real-world performance monitoring (as outlined in Section 3) from other sites, or evaluating tools against benchmark data sets, such as the use of the Personal Performance in Mammographic Screening (PERFORMS) database to benchmark AI in breast radiology.[13]

    a. When considering benchmarking, it is important to consider whether the AI tools have been benchmarked against a data set that reflects the real-world population, or an enriched data set that may more reliably identify limitations but may overstate algorithm performance.

    b. This may be limited by a current lack of availability of benchmarking data sets, though there are examples of AI software in radiology such as the Health AI Register.[5]

12. Identify whether suppliers will be willing to make data available for independent validation.

13. A post-market surveillance (see Glossary) plan should be developed as part of this stage.

Potential hazard and safety implications

14. Consider the potential hazards and safety implications of using AI in clinical practice prior to deploying an AI solution.

15. Review the Health and Social Care Act 2012, Section 250, which sets out the statutory obligations to complete risk assessments for digital solutions deployed in the NHS.[14]

16. Review the DCB0129 manufacturer and DCB0160 deployment organisation information standards including the requirements to produce a clinical safety report and hazard log. These form part of the Digital Technical Assessment Criteria (DTAC) for deploying AI and are commonly included within the contractual requirements with the AI supplier.[15]

17. Ensure the clinical hazard log is tailored to the intended use of AI and consider the possibility of the AI inadvertently causing patient harm. This includes the likelihood and potential adverse consequences from AI 'overcalling' abnormalities (false positives) and AI missing significant abnormalities (false negatives). Overcalling findings can potentially lead to unnecessary further investigations and interventions, and missing abnormalities may delay the patient's diagnosis. The hazard log should record any mitigations to reduce the risk including any preclinical shadow mode assessments and staff training required prior to deployment.

Shadow mode

18. Evaluate AI in shadow mode as a standard deployment model for AI. This is where AI is enabled to run in the background on real patient data but the findings are not made available to be used in clinical practice. Enabling shadow mode provides data on how AI will perform in real-world conditions and enables comparison of the AI model with the current operational performance and clinical outcomes. This also serves as a baseline and a test of how data and outcomes are recorded, which may lead to recommendations for the subsequent clinical evaluation protocol including what metrics to record and code to support the analysis.

19. Test an enriched data set of positive and negative cases from the local institution to supplement shadow mode evaluation. The idea is not to redefine the performance metrics of the software, as this should have already been made clear by the manufacturer, but rather to ensure that the AI software is functioning as intended in the

context of the local population, staff, scanners, protocols and systems. For example, AI can be run in shadow mode on a retrospective sample of chest X-rays of patients with lung cancer, identified through performing the RCR audit of cancers at baseline.[8] This enables you to determine if AI can pick up any cases that were missed by the reporters, and conversely whether AI may miss any cancers that were detected by the reporters (false negative rate).

20. Use shadow mode prior to deployment to estimate the incidence of AI findings in the referral population. Review a sample of cases for each abnormality to predict how often AI may overcall abnormalities (false positive rate) to help set expectations in user training. Manufacturers do not normally provide these figures for deployment as the rates depend on the prevalence of the findings in the referral population.

Staff training

Staff will need to be sufficiently trained in issues specific to AI in healthcare as part of the acquisition and requirements planning stage. This includes understanding AI capabilities and an awareness of algorithm bias and human–AI interactions, clinical integration across the pathway and how the tool may have downstream effects.

21. Training needs to include general AI knowledge and domain (thoracic imaging) AI expertise plus training on the generic and specific risks and performance of the chosen AI algorithm. This will help ensure weaknesses in human and AI decision-making are minimised and the AI complements existing human expertise.

22. Consider how the introduction of AI will impact the training of radiologists and radiographers.

23. Identify appropriate training in how to interpret the findings where AI is to be used as clinical decision support. Radiologists and clinicians are used to assimilating evidence to help them come to a clinical diagnosis, some of which may be conflicting. The potential risk with AI is that if users are unaware of how AI works and its strengths and weaknesses, they may be unduly influenced by the technology, a phenomenon known as ‘automation bias’ (see Glossary). The purpose of AI training is to maximise the benefits of the AI while minimising the risk of automation bias.

24. Collate sample cases while in shadow mode to use for training, including ‘wow’ cases where AI can identify hard-to-spot abnormalities and make a difference to patient care. Balance these with examples where AI may overcall or miss findings. Educate users on

the anticipated false positive and false negative rates of AI to set appropriate expectations. Staff who have appropriate situational awareness of AI are potentially more likely to use it appropriately.

25. Train users to make their own interpretation first and then to review the AI findings. This can be supported by the technology through use of display protocols and requiring an extra step to click to view the AI. Some AI providers can also display a level of confidence in the AI findings, based on pre-market studies and the performance in shadow mode.

26. Train all staff who have access to the AI findings, and consider whether it is necessary to limit access to the AI report to trained members of staff.

27. Inform staff that if the AI report is accessible on the system, it must be clear that the interpreted findings are provisional and require validation by trained reporters, as appropriate to the terms of the AI product regulatory licence.

5. Peri-deployment and deployment

Peri-deployment

1. Identify whether the AI performs as expected and the mitigations are effective during peri-deployment activities including shadow mode.

2. Gather, analyse and act upon user feedback, which will provide valuable insights into user experiences, potential issues and areas for improvement.

3. Review existing incident reporting processes and radiology events and learning meetings (REALM) as these can help identify any consequences that may require adjustment to the AI algorithm or training and operational procedures.

4. Collate lessons learned for early live clinical evaluation, which can help inform other projects and should be shared with the supplier as part of the post-market surveillance activities.

Deployment

5. Continuous monitoring of ethical considerations such as bias and fairness is vital. Addressing any ethical concerns that arise during actual usage helps maintain trust in the AI system.

References

1. Healthcare Safety Investigation Branch. Missed detection of lung cancer on chest X-rays of patients being seen in primary care. HSIB, 2021. www.hssib.org.uk/patient-safety-investigations/missed-detection-of-lung-cancer-on-chest-x-rays-of-patients-being-seen-in-primary-care

2. National Institute for Health and Care Excellence. Artificial intelligence-derived software to analyse chest X-rays for suspected lung cancer in primary care referrals: early value assessment. NICE, 2023. www.nice.org.uk/guidance/hte12

3. https://pard.mhra.gov.uk

4. NHS AI and Digital Regulations Service for health and social care. Guidance for adopters. www.digitalregulations.innovation.nhs.uk/regulations-and-guidance-for-adopters

5. Health AI Register. https://radiology.healthairegister.com.

6. Kim C, Yang Z, Park SH *et al*. Multicentre external validation of a commercial artificial intelligence software to analyse chest radiographs in health screening environments with low disease prevalence. *Eur Radiol* 2023; **33**: 3501–3509. doi: 10.1007/s00330-022-09315-z.

7. Maiter A, Hocking K, Matthews S *et al*. Evaluating the performance of artificial intelligence software for lung nodule detection on chest radiographs in a retrospective real- world UK population. *BMJ Open* 2023; **13**: e077348. doi:10.1136/bmjopen-2023-077348.

8. www.rcr.ac.uk/career-development/audit-quality-improvement/auditlive-radiology/audit-template-to-assess-compliance-of-low-radiation-dose-ct-used-in-the-targeted-lung-health-check-scans

9. National Institute for Health and Care Excellence. Evidence generation plan for artificial intelligence-derived software to analyse chest X-rays for suspected lung cancer in primary care referrals. NICE, 2023.

10. National Institute for Health and Care Excellence. NICE real-world evidence framework. NICE, 2022.

11. www.england.nhs.uk/nhs-commercial/central-commercial-function-ccf/procurement-framework-strategy-recommendations

12. NHSX. *A buyer's guide to AI in health and care: 10 questions for making well-informed procurement decisions about products that use AI*. NHSX, 2020. https://transform.england.nhs.uk/ai-lab/explore-all-resources/adopt-ai/a-buyers-guide-to-ai-in-health-and-care

13. Chen Y, James JJ, Cornford EJ *et al*. The relationship between mammography readers' real-life performance and performance in a test set–based assessment scheme in a national breast screening program. *Radiol Imaging Cancer* 2020; 2e200016.

14. Health and Social Care Act 2012, section 250. www.legislation.gov.uk/ukpga/2012/7/section/250

15. https://digital.nhs.uk/services/clinical-safety/clinical-risk-management-standards

Abbreviations

AIDF	Artificial Intelligence Diagnostics Fund
CT	computed tomography
DPIA	data protection impact assessment
DTAC	Digital Technical Assessment Criteria
ED	emergency department
GP	general practitioner
NICE	National Institute for Health and Care Excellence
MHRA	Medicines and Healthcare Regulatory Authority
PACS	picture archiving and communication system
RIS	radiology information systems
RWD	real-world data
UKCA	UK Conformity Assessed

Glossary

Automation bias – occurs when users are unaware of how AI works and its strengths and weaknesses and they may be unduly influenced by the technology.

Diagnostic accuracy study – measures the reliability of diagnostic tests outside of the highly controlled research environment.

Longitudinal clinical impact study – a research design that involves repeated observations of the same variables over periods of time.

Post-market surveillance – monitoring device safety and performance.

Pre-market evaluation – completed to ensure that the product's design, functionality, performance and safety are sufficiently predictable and that the predicted standard of each of these aspects is acceptable.

Shadow mode – provides data on how AI will perform in real-world conditions and enables comparison of the AI model with the current operational performance and clinical outcomes.

Appendix 1 AI buyer's guide assessment template

from NHS AI Lab AI buyer's guide (adapted from the template by Haris Shuaib, Clinical Scientific Computing, Guy's & St Thomas' NHS Foundation Trust)

0.0	**Background information on product**
0.1	Vendor or manufacturer's name:
0.2	Name of product:
0.3	Short description of product:
0.4	Intended users of product:
0.5	Anticipated timescale for potential implementation in your organisation:
0.6	Main point/s of contact within your organisation for liaising with vendor:

1.0	Problem to be solved
1.1	**Challenge-driven**
1.1.1	What is the problem you are trying to solve?
1.1.2	What is the rationale for choosing AI to solve your problem? What is it about AI – over and above other solutions – that makes it a powerful choice?
1.1.3	What is the appropriate scale for addressing your challenge (eg organisational, system, regional or even national)?
1.2	**Credible business case**
1.2.1	What is the baseline you are looking to improve, and what metrics matter in measuring this improvement?
1.2.2	What do you expect the quality improvements and/or savings and efficiencies to be for your organisation?

2.0	Regulatory standards
2.1	What is the intended use of the product? What can it be used for and under what conditions can it be used? What can it not be used for?
2.2	If the product is defined as a medical device, does it have CE marking? What is the product's risk classification, and do you agree with this designation?
2.3	If the product carries out regulated clinical activity independently of clinicians, has it been registered as a service through the Care Quality Commission (CQC)?
2.4	If the product is categorised as operational healthcare software, has the manufacturer developed it in line with ISO 82304?
2.5	If the product is categorised as healthcare software in general, have you asked to see documentation enabling you to monitor the product manufacturer's compliance with DCB0129?

521

522

523

3.0	**Valid performance claims**
3.1	Does the prediction generated by the AI model result in an output that supports practical action?
3.2	**Model performance metrics**
3.2.1	**If classification model:**
3.2.1.1	What are the sensitivity and specificity metrics of the model? Does the trade-off between these metrics give you confidence, given the context of your use case?
3.2.1.2	What are the positive predictive value and negative predictive value metrics of the model? Does the trade-off between these metrics give you confidence, given the context of your use case?
3.2.1.3	Is there an issue of class imbalance to take into account?
3.2.1.4	What is the model threshold? Does the choice of threshold correspond to the use case?
3.2.1.5	What is the area under the curve (AUC) metric of the model?

3.2.2	**If regression model:**
3.2.2.1	What is the root mean square error (RMSE) of the model?
3.2.2.2	What is the mean absolute error (MAE) of the model?
3.2.2.3	What is the R-squared (R^2) value of the model?
3.2.2.4	How much of an issue are outliers for your use case data set, and how does this influence which of the metrics above should be prioritised?
3.3	**Model validation**
3.3.1	What are the results from validation tests, to understand the model's predictive performance on data it hasn't seen before? Was the validation internal or external?
3.3.2	Has the separation of training and validation data been clearly documented?
3.3.3	Do you understand the characteristics of the validation data set and what it was used to test for?

	Was the validation data set: • Similar to the original data set in terms of its population and setting? • Different to the original data set in terms of its population and/or setting? • Representative of the same or new populations over time? • Different to the original data set on account of technical reasons (eg images taken on different scanners)?
3.3.4	Was the validation data set sampled fairly and representatively, and did it incorporate edge cases?
3.4	**AI safety**
3.4.1	How does the vendor evidence model robustness? Can the model make reliable predictions, given that data are subject to uncertainty and errors? Does the model remain effective even in extreme or unexpected situations?
3.4.2	How does the vendor evidence model fairness? What measures are in place to prevent the model from discovering hidden patterns of discrimination in its training data, reproducing these patterns and making biased predictions as a result?
3.4.3	How does the vendor evidence model explainability? Can predictions made by the model be explained in terms that both a trained user of the product and a patient or service user would understand?

3.4.4	How does the vendor evidence model privacy? Is the model resilient against attempts to reidentify individuals whose data was contained in the model's training set?
3.5	**Comparative performance**
3.5.1	How does reported model performance compare with the current state (how things are currently done without use of the AI product)?
3.5.2	Is it possible that the seemingly obvious comparator current state may not be the best place to look for potential value offered by the AI product? Are there any less obvious comparators?

524

4.0	**Will the product work in practice**
4.1	**Evidence base for effectiveness**
4.1.1	What is the evidence base for demonstrating the product's effectiveness? Is the standard of this evidence sufficiently robust, taking into account the function and associated risk of the product?
4.2	**Insight from other organisations**

4.2.1	What insight is available on the product's effectiveness in other health and care settings?
4.3	**Deliverability**
4.3.1	If significant changes to your organisation's ways of working are needed to realise the benefits promised by the product, is this possible?
4.3.2	If implementation of the product will cause short-term disruption, how will you manage this?
4.3.3	If you are replacing an older system with the new technology, have you factored in time, costs and potential complications of dealing with a legacy system?
4.3.4	Have you considered starting off with a pilot project with a tightly defined scope and set of success metrics before scaling up?
4.3.5	What artefacts does the product produce? For example: Does it produce additional data or files? Does it trigger an alert? If so, what kind of alert?

4.3.6	Does the product record and make available operational data (eg processing time, product usage)?
4.4	**Usability and integration**
4.4.1	How will the product interface with different technology systems that are implicated in your deployment, and how will you ensure clear and reliable workflows?
4.4.2	Have you asked the vendor for the software architecture diagram?
4.4.3	Does the product make use of open standards to promote interoperability?
4.4.4	If you want to automatically access the product's internal data, have you considered whether the product has an application programming interface (API)?
4.5	**Data compatibility**
4.5.1	What are the product's data requirements, and how will it ingest this data for processing?

4.5.2	Does your organisation have the data needed, in the right format? What are the sources and types of data needed?
4.5.3	Can your organisation's data be labelled and stored in the right way?
4.5.4	How reliable is the quality of this data?
4.6	**Data storage and computing power**
4.6.1	What are the data storage and computing power requirements of the product? How much data will the product need and generate, and how long will the data be stored for?
4.6.2	If your project will use cloud-based servers, are you clear about where these are based?
4.6.3	If data storage and computing infrastructure is not provided by the vendor, can your organisation cover the associated costs?

4.6.4	As your use of the product scales and data-processing requirements increase, will the infrastructure costs increase in a linear or exponential way?
4.7	**Auditing and evaluation**
4.7.1	Have you considered how you will audit and evaluate the product and its implementation? Have you factored this into your costs?

525

5.0	**Support from staff and service users**
5.1	**Staff**
5.1.1	Have you directly involved staff who will be end-users of this prospective product in the procurement exercise?
5.1.2	Which staff groups have you engaged and gathered input from regarding this procurement?
5.1.3	How confident are you of widespread clinical, practitioner and operational support for the product? What will you do to cultivate this?

5.1.4	Will your vendor supply any induction or training that is needed in your organisation?
5.2	**Service users**
5.2.1	How compelling a story can you tell about the expected improvement in health and care outcomes?
5.2.2	How will you communicate with patients and service users about how the AI product is being used, how their data are being processed and, where relevant, how an AI model is supporting decisions that affect them?

526

6.0	**Culture of ethics**
6.1	Are you confident that your AI project, and the product in question, is: • Ethically permissible? • Fair and non-discriminatory? • Worthy of public trust? • Justifiable?
6.2	Have you assessed your project against the principles of the Data Ethics Framework? Are there any areas of the project that need revisiting as a result?
6.3	Have you carried out a stakeholder impact assessment? What are the key insights from it?

527

7.0	Data protection and privacy
7.1	Will you be able to create a data flow map that identifies the data assets and data flows pertaining to your AI project?
7.2	Will you be able to develop a data-processing contract (otherwise known as an information-sharing agreement) with the vendor?
7.3	Is your organisation's use of data for this project covered under its data privacy notice?
7.4	What will be in place in terms of data protection to mitigate the risk of a patient or service user being reidentified – in an unauthorised way – from the data held about them?
7.5	In cases where you will be processing personally identifiable data, will you be able to complete a data protection impact assessment (DPIA)?

528

8.0	**Ongoing maintenance**
8.1	**Vendor's responsibilities**
8.1.1	Is the vendor providing a managed service for the product?
8.1.2	What is the vendor's approach to product and data pipeline updates? Who pays for these?
8.1.3	What is the vendor's plan for mitigating adverse events (if the AI product fails or is compromised)?
8.1.4	What is the vendor's plan for addressing performance drift? Have you agreed a suitable margin of acceptable drift? Does performance need continuous monitoring or is an interval audit sufficient?
8.2	**Your organisation's responsibilities**
8.2.1	If you are not buying into a managed service, do you have the IT capability in-house?
8.2.2	Can your organisation develop a sufficiently robust understanding of relevant data feeds, flows and structures, such that if any changes occur to model

	data inputs you can assess any potential impacts on model performance or signpost questions to the vendor?
8.2.3	Are you clear about your organisation's reporting requirements for adverse events?
8.2.4	What are the vendor's expectations of your organisation sending back data to support its iteration of the model or development of other products? Have you clarified what the vendor means by model iteration and development, and have you ensured that your information governance arrangements address this?
8.3	**Decommissioning**
8.3.1	On decommissioning the product, what will happen to any data that are stored outside of your organisation's systems? Will it be deleted, or archived?
8.3.2	How will you ensure that you have access to any data or analysis you require that is due to be deleted or archived?
8.3.3	On decommissioning the product, how will you ensure that the vendor's access to any part of your organisation's infrastructure is revoked in full?

529

9.0	Compliant procurement
9.1	Have you clearly documented and justified instances of your organisation talking to or inviting specific vendors to bid for the project?
9.2	If you are being offered a product for free, what steps have you put in place to ensure that you remain compliant with public procurement guidelines?

530

10.0	Robust contractual outcome
10.1	**Commercial**
10.1.1	Are you clear about exactly what you are buying? For example: Is it a lifetime product? Is it a licence? What is the accompanying support package?
10.1.2	Have you set out a clear specification and service-level agreement? Do these secure the quality, availability, flexibility and performance of service that you need?
10.1.3	What provisions are in place for contract termination and handover to another supplier?

10.1.4	To what extent will you be able to publish details of your contract?
10.2	**Intellectual property**
10.2.1	How will you ensure that any agreement with your prospective vendor is fair, in the sense that it recognises and safeguards the value of the data you are sharing?
10.3	**Liability**
10.3.1	With regard to product liability, is the vendor providing any indemnities, and are they clearly set out in the contract?
10.3.2	Is it clear what is considered as product failure versus human error in using the product?
10.3.3	What is the extent of cover your own indemnifier or insurer can provide in the event of product failure or human error? Do you need to purchase additional cover or extend existing cover?
10.3.4	Does your contract and information governance documentation clearly set out what measures you expect the vendor to have in place for compliance with data protection regulation?

531

Appendix 2 Technical requirements for the tool and the supplier that should be incorporated

NHS policy or requirement	Purpose or expectation
Public cloud first	Digital services should move to the public cloud unless there is a clear reason not to do so.
Internet first	All new health and social care digital services should be internet facing.
HL7 FHIR conformant supporting FHIR UK Core	UK Core is an implementation guide that provides a four-nation approach to Fast Healthcare Interoperability Resources (FHIR), which applies across jurisdictions and care settings.
HL7 FHIR Code System, Value Set and Concept Map including all operations	Adherence to HL7 FHIR requirements supports integration with other digital products in use within the service or network.
DCB0129 conformant	DCB0129 is a clinical safety standard that requires suppliers of digital health solutions to verify the safety of their products.
SNOMED CT conformant	SNOMED CT is a structured clinical vocabulary for use in an electronic health record.
ICD10 conformant	The World Health Organization (WHO) International Classification of Diseases (ICD) is the global standard that categorises and reports diseases in order to compile health information related to deaths, illness or injury.
ODS conformant	The Organisation Data Service (ODS) issues and manages unique identification codes (ODS codes) and accompanying reference data for organisations that interact with any area of the NHS.
WCAG 2.1 at AA level for any web-based or mobile user interfaces	Web Content Accessibility Guidelines (WCAG) 2.1 defines how to make web content more accessible to people with disabilities.
Aligns to NCSC cloud security principles	Application of the cloud security principles assists in choosing a cloud provider that meets minimum cybersecurity needs.