Clinical Radiology xxx (xxxx) xxx

Contents lists available at ScienceDirect

Clinical Radiology

journal homepage: www.clinicalradiologyonline.net

Review

Beyond regulatory compliance: evaluating radiology artificial intelligence applications in deployment

J. Ross [a,*], S. Hammouche [a], Y. Chen [b], A.G. Rockall [a] and the Royal College of Radiologists AI Working Group[†]

[a] Department of Cancer and Surgery, Imperial College London, UK
[b] School of Medicine, University of Nottingham, UK

ARTICLE INFORMATION

Article history:
Received 25 August 2023
Received in revised form
24 January 2024
Accepted 29 January 2024

The implementation of artificial intelligence (AI) applications in routine practice, following regulatory approval, is currently limited by practical concerns around reliability, accountability, trust, safety, and governance, in addition to factors such as cost-effectiveness and institutional information technology support. When a technology is new and relatively untested in a field, professional confidence is lacking and there is a sense of the need to go above the baseline level of validation and compliance. In this article, we propose an approach that goes beyond standard regulatory compliance for AI apps that are approved for marketing, including independent benchmarking in the lab as well as clinical audit in practice, with the aims of increasing trust and preventing harm.

Introduction

There has been exponential growth in radiology artificial intelligence (AI)-related publications,[1] teamed with an increasing number of Food and Drug Administration (FDA) and conformité européenne (CE) marked radiology AI software as medical devices for clinical practice. Of the 692 AI-enabled FDA authorised medical devices in October 2023, 77% were primarily radiology focused[2]; however, the implementation of AI applications (apps) in routine practice is limited by practical concerns around reliability, accountability, trust, safety, and governance.[3] Until recently, the post-market surveillance of AI tools for clinical use in radiology has been relatively undefined,[4] and there are concerns about the differences between published and real-world performance of approved AI apps, as well as human factors, which consider the way humans will interact with AI.[5]

Radiologists are essential partners in ensuring the safe deployment and usage of AI apps, by collaborating with multidisciplinary teams to take part in clinical audit to evaluate the "real-world" performance of such tools. The

* Guarantor and correspondent: Jack Ross, Imperial College London, London W12 0NS, UK.
E-mail address: jackross@nhs.net (J. Ross).
[†] Members of the Royal College of Radiologists AI working group who co-authored this paper are: S. Alabed, M. Chen, K. Dwivedi, D. Fascia, R. Greenhalgh, M. Hall, K. Halliday, S. Harden, W. Ramsden, S. Shelmerdine.

https://doi.org/10.1016/j.crad.2024.01.026

grounds for a registry of AI deployed apps in radiology have been made,[6] and the Royal College of Radiologists has suggested proposals to overcome barriers to AI implementation in imaging.[7] Having an independent, centrally coordinated data trail of audits at different sites with deployed AI apps would be a major benefit. Standardising the way in which AI apps are evaluated will also allow for more direct comparisons between competing products. In addition, there may be opportunities to devise a benchmarking programme using retrospective expertly annotated datasets from multiple sites.

Reporting guidelines such as DECIDE-AI have been developed to help appraise early stage clinical studies of AI apps,[8] as well as audit techniques for potential algorithmic errors.[9] Audit of AI algorithms presents additional challenges and factors such as generalisation and adaptability, where an algorithm struggles to perform beyond its initial trained datasest. This inherent training bias should be considered when a tool is a applied to an unseen dataset which may include underrepresented cases. In addition, due to the risk of algorithmic errors, system failures and technical glitches affecting algorithm accuracy, safety and reliability should be included within the audit as these errors may lead to incorrect diagnoses or delays in patient care.

In this article we review the aims of clinical audit and the concept of beyond compliance, and then present proposed methods for “in the lab” benchmarking of approved apps, as well as clinical surveillance of AI applications once deployed.

Why should we undertake independent audit of AI applications?

To bring AI apps to market, industry developers need to demonstrate appropriate diagnostic performance to achieve regulatory approval, such as CE-marking in Europe and FDA approval in the USA. The industry developers are also responsible for providing post-marketing surveillance; however, this demand will inevitably be constrained by conflicts of interest, with one study suggesting fewer than half of studies for CE-marked products were independent from vendors.[10] Concerns about reliability and transparent audit of performance in clinical practice have led to calls for improved, independent, and comprehensive post-marketing surveillance of AI devices, including by the American College of Radiology.[11]

A systematic review of machine learning models for the diagnosis and prognosis of COVID-19 from chest radiography (CXR) or computed tomography (CT) images found that none of the models were of potential clinical use because of significant flaws in the methodology and underlying biases.[12] In Europe, a study reviewing CE-marked AI apps in imaging in 2021 found that only half of the available evidence was independent, only 18% showed potential clinical impact and 64% of the approved AI apps lacked peer-reviewed evidence of efficacy.[10] In the USA, a study looking at FDA-approved AI apps found a general lack of transparency and adequate evaluation datasets, with comprehensive prospective evaluation only performed for four of the identified 130 AI apps.[5,13] The recently published early value assessment of CXR for detection of early lung cancer was informed by evidence synthesis that found “no applicable evidence on which to evaluate the impact of adjunct AI software for analysing chest X-rays from people referred from primary care for suspected lung cancer” and as such made appropriate research recommendations.[14]

A significant challenge facing AI applications is how vulnerable their performance can be to systematic differences between their training data and the “real-world” data seen in clinical practice. Algorithm performance can vary in different environments (including different machine vendors, PACS systems, and AI deployment platforms)[15] and can drift as the environmental factors change, including patient population, resulting in significantly worse performance when evaluated at other sites.[13,16] This is of particular concern to AI apps that have been predominantly trained on cohorts from a small number of sites with limited geographic diversity.[17] AI is vulnerable to “hidden stratification”, where models can perform well overall but underperform on particular subgroups, and this has been shown to cause clinically meaningful failures in medical imaging AI.[18]

To ensure the safe and effective implementation of AI in clinical settings, the concerns around AI shortcomings must be addressed. This presents an opportunity for clinical radiologists to support independent evaluation and validation of AI. Organisations such as the Royal College of Radiologists, working together with other key partners, could provide a framework for clinical audit of radiology AI apps both “in the lab” (also known as “in silico”) and in real-world environments, providing curated validation datasets of “ground truth” for benchmarking new apps (see Fig 1). Those apps that are validated successfully could receive kite-marking or equivalent. In addition, this framework could be supplemented by evaluation and feedback of other important factors limiting clinical AI adoption, such as human factors and ease of use. It is likely that the Royal College of Radiologists will need to work in partnership with other organisations, and these proposals aim to go hand-in-hand with national programmes such as the NHS AI Lab[19] and the AI buyer’s guide,[20] as well as initiatives aiming to improve trust and generalisability in clinical AI such as the Health AI Partnership,[21] Coalition for Health AI,[22] STANDING together[23] and other commercial and academic coalitions.

When a technology is new and relatively untested in a field, professional confidence is lacking and there is a sense of a need to go above the baseline level of validation and compliance. In this article, we propose an approach that goes beyond standard regulatory compliance for AI apps that are approved for marketing, including both independent benchmarking in the lab as well as independent clinical audit in practice, with the aims of increasing trust and preventing harm.

Please cite this article as: Ross J et al., Beyond regulatory compliance: evaluating radiology artificial intelligence applications in deployment, Clinical Radiology, https://doi.org/10.1016/j.crad.2024.01.026

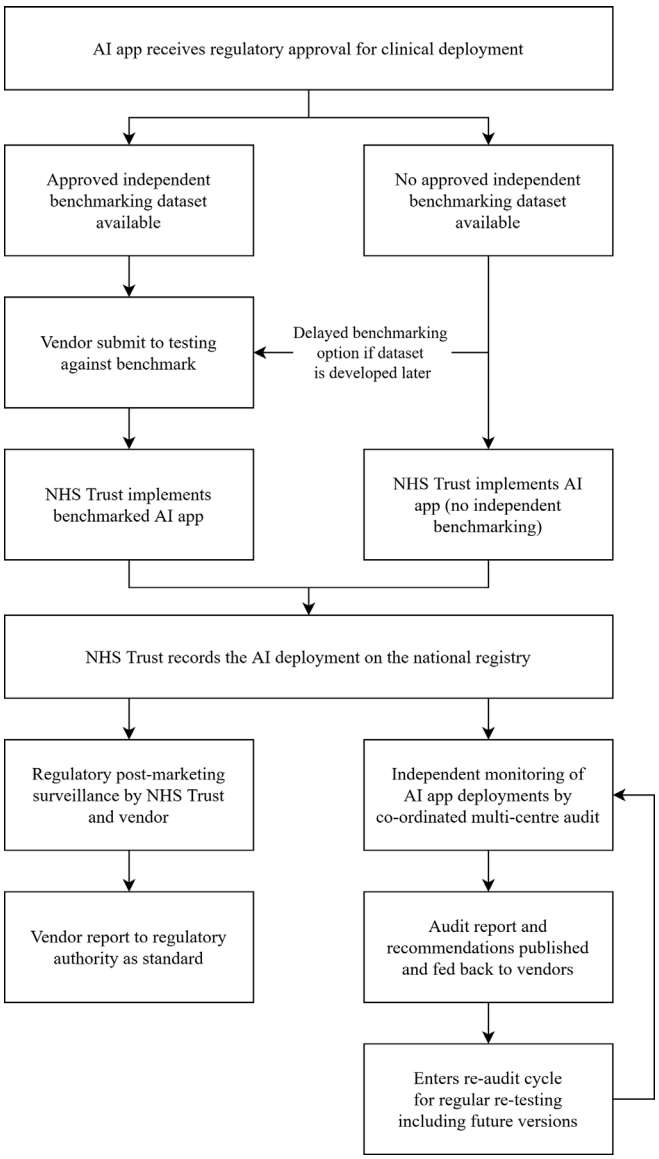

Figure 1 Flow diagram for proposed AI app deployment and evaluation.

Beyond regulatory compliance: an approach to ensuring the safety and effectiveness of deployed AI applications

Previous authors have compared evaluation of AI to pharmaceuticals and surgical innovation.[8] The concept of "beyond compliance" has been gaining traction in recent years, as organisations seek to ensure that their products and services are safe and effective for their intended use.[24] This approach transcends mere adherence to minimum regulatory requirements and instead concentrates on proactively identifying and mitigating potential risks and issues at an early stage.

The notion of "beyond compliance" has been particularly instrumental in the field of surgery, specifically in the context of orthopaedic implants.[24,25] This primarily involves the evaluation of new implants followed by their continuous monitoring. The genesis of this concept in surgical practice can be attributed to the significant clinical failures witnessed in certain metal-on-metal total hip replacements during the early part of the previous decade, despite initial approval by regulatory bodies such as the FDA and the Medicines and Healthcare products Regulatory Agency (MHRA).[26–29]

In the surgical domain, "beyond compliance" encompasses both pre-market assessment and post-market surveillance, incorporating various data sources including clinical research, user experiences, registry data, and adverse event analyses.[25,26,26,30,31] Regular feedback loops are established between manufacturers and the "beyond compliance" committee and advisors, fostering a proactive environment for identifying and resolving potential risks before they escalate. By employing this approach, early indicators of potential issues can be detected, prompting appropriate actions and recommendations. Consequently, this approach not only facilitates early risk identification but also bolsters confidence in the assessed products or services, thereby enhancing their overall impact.

A similar approach could be extended to the evaluation of clinical AI applications, particularly following deployment, to ensure their performance, effectiveness, and safety.[32,33] Potential risks could be detected at an early stage, facilitating local problem-solving and shared learning on a national scale, ensuring avoidance of clinical harm, enhancing the overall safety and effectiveness of AI applications. By demonstrating that an AI application has undergone thorough evaluation and satisfies high standards of safety and effectiveness, with planned on-going independent clinical audit, users are more inclined to trust the products or services being evaluated and feel assured in their usage. Consequently, this facilitates greater adoption of AI applications and amplifies their overall impact.

One ongoing challenge is the lack of health economic evaluation for many commercially available AI applications. A recent review from the National Institute for Health and Care Excellence in the UK (NICE) has published an early value assessment of AI-apps to analyse CXRs for suspected lung cancer, which identified the need for further evidence,[34] which may be collected through the NHS England AI Diagnostic Fund (AIDF).[35]

"In the lab" evaluation of AI applications

Retrospective studies have suggested that AI can achieve or even exceed human reader cancer detection performance[36]; however, studies that are based on an evaluation of an AI system using data collected retrospectively are subject to numerous biases and present an inferior level of evidence compared with prospective studies. The relevant dataset may not always perfectly reflect the final population in which the tool will be used. On the other hand, prospective evaluation of AI is time-consuming and requires large sample sizes so more difficult cases are included in sufficient numbers. There are several methods to undertake retrospective testing, none of which are perfect and each

Please cite this article as: Ross J et al., Beyond regulatory compliance: evaluating radiology artificial intelligence applications in deployment, Clinical Radiology, https://doi.org/10.1016/j.crad.2024.01.026

has pros and cons. A real-life dataset reflecting the population may be challenging to collect and annotate. An alternative approach is one that offers speedy evaluation against verified case collections that include greater proportions of the more challenging cases (i.e., an enriched dataset).

Poor reporting of in-lab evaluations of AI applications continues to pose a challenge in the clinical applicability of AI models.[37] To address this issue, recommendations have been devised based on systematic reviews of AI studies.[38] These recommendations emphasise the need for transparent reporting of data sources, including comprehensive descriptions of cases, eligibility criteria, and clinical characteristics to enable the assessment of the model's validity, relevance, and generalisability. Additionally, the description of model training should provide sufficient detail to ensure reproducibility. Consistency in reporting model performance enables meaningful comparisons between different models. Furthermore, reporting failures and limitations of the models plays a vital role in understanding potential biases inherent in AI models. Implementing these recommendations would greatly improve the overall quality and interpretability of in-lab evaluations of AI applications.

In addition, concern has been raised that changes in the operating environment over time can lead to less reliable algorithm performance.[39] This may be due to changes in the screening population, such as patient age or ethnic diversity, or even changes in other software, hardware, or applications deployed alongside the AI.[40] In the example of breast cancer screening, using real-life data to assess performance is problematic because obtaining an accurate measure of sensitivity and specificity in a timely fashion is difficult as true sensitivity may not be known for many years until interval cancer data are collected.[41] Similarly, for specificity measurement, the proportion of truly disease-free patients correctly identified as negative by AI will not become apparent until after the next screening round. Consequently, it may take several years for poor algorithm performance to be noticed by which time women may be harmed or the reputation of AI in breast cancer screening damaged.

The NHS Breast Screening Programme routinely uses a test set external quality assurance (EQA) scheme called Personal Performance in Mammographic Screening (PERFORMS) to assess reader performance.[41] PERFORMS is accredited by the Royal College of Radiologists and the European Accreditation Council for Continuing Medical Education as it helps participants to improve readers skills and knowledge through Continuing Professional Development (CPD), remain up to date in their specialities and comply with the relevant professional standards, providing them with appropriate CPD credits. As part of the scheme, a large dataset containing challenging two-dimensional (2D) full-field digital mammograms (FFDM) cases from multiple diverse sources, equipment vendors, and different geographic areas has been annotated by a panel of expert radiologists, each one of them with >20 years of experience.

Part of the PERFORMS dataset has already been used to compare the performance of human readers and a commercially available AI algorithm interpreting test sets. Each breast was considered separately, and the highest score was used to assess performance using a pre-defined recall threshold. Sensitivity, specificity, and ROC analysis was used to compare the performance of AI and human readers retrospectively (see Fig 2).

A similar platform of agreed use cases (e.g., stroke CT, lung cancer screening or CXR) of ground truth could be hosted by the Royal College of Radiologists for "in the lab" evaluation to ensure the safe deployment of AI apps, by this independent benchmarking approach. Diagnostic performance metrics such as sensitivity, specificity, negative predictive value, and positive predictive values will differ between a natural and "enriched" dataset. Regularly evaluating AI with an external quality assurance test set scheme has advantages as the outcome for each case is already known. An advantage of using cases from EQA test sets like PERFORMS is that AI performance can be immediately compared to a large cohort of human readers who have all read identical cases from the same patient population, thus providing a robust performance comparison between human readers and AI. Regular retesting on an up-to-date benchmark dataset will help detect drift in algorithm performance as operational factors change, although the frequency of retesting may vary between applications. It is important to note that, currently, re-training algorithms with local real-world data may invalidate the intended use authorisation aspects of the software and require further regulatory approval.

Developing validation datasets for benchmarking

Cases used in benchmarking datasets should reflect the population for the intended use. In the case of AI apps for screening tests, cases must be drawn from the screening programme where the AI product is to be deployed. For

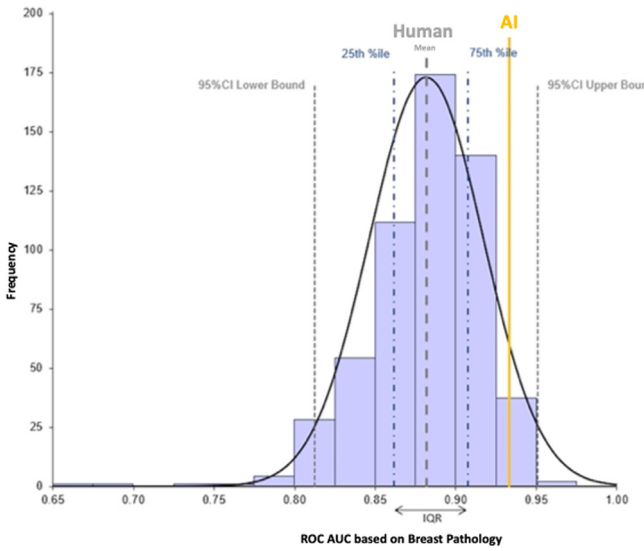

Figure 2 AI and human performance in mammographic screening.

Please cite this article as: Ross J et al., Beyond regulatory compliance: evaluating radiology artificial intelligence applications in deployment, Clinical Radiology, https://doi.org/10.1016/j.crad.2024.01.026

example, in lung cancer screening, cases should only be included from active or recent smokers deemed high risk of lung cancer using the same risk prediction models utilised in the screening programme. In breast cancer screening, a test set for evaluating AI performance in the UK breast screening would consist of mammograms from the women age between 50 and 70 years from the UK National Health Service Breast Screening Programme (NHSBSP).

The dataset must also account for the heterogeneity observed within the real-life screening population. Taking the example of breast cancer screening, datasets should ideally include an accurate proportion of screen-detected cancers, interval cancers, radiological feature types, women with a personal history of breast cancer, women from different ethnic groups, mammograms acquired on equipment from different vendors, with the range of post-image acquisition processing software used throughout the NHSBSP. For any dataset used for benchmarking, it is critical to consider how representative the test population is compared to the patient population for this to be a fair test.

Accurate and reliable outcome information is required for each case in a test set. In order to provide ground truth information, cancer cases must be confirmed by biopsy. Normal or benign cases must either be confirmed by biopsy or adequate follow-up. For instance, for mammographic screening a mammogram should only be called truly normal when a normal outcome is also recorded for the mammogram at the next screening round, so for the NHSBSP this would be 3 years later as women are invited for screening every 3 years.

Real-world clinical evaluation post deployment

AI that has been deployed in the clinical setting can be evaluated prospectively to ensure safety and effectiveness. An example clinical audit tool is provided (see Table 1). Validating AI performance on local data will be important to help build clinicians' trust, allowing the comparison of both model performance and the impact on clinical outcomes. Clinical service evaluation will be able to assess the frequency of correct and misdiagnosis by the AI models, and compare cases where AI outperforms and underperforms compared to clinician readings. Examining these discordant cases can help identify areas where AI apps are vulnerable to underperform. Crucially, the subsequent clinical outcomes of these cases can be explored, allowing quantification of both adverse outcomes where AI could lead to a negative impact on patient care, as well as positive cases where the use of AI led to a beneficial clinical outcome or improved radiology productivity. Centrally coordinated audit and publication of these clinical evaluations is also important, as it will allow early identification of problems that may be occurring at multiple different departments and hospitals, with dissemination of the information to the wider community. This would also enable health economic assessments to be coordinated to calculate the cost-effectiveness of AI apps.

Table 1
Suggested components to include in AI clinical audit template.

Project title
Responsible clinician
Name of AI vendor, product name and version
Name of PACS vendor
Name of RIS vendor
Name of AI deployment platform (if applicable)
Integration method (e.g. PACS, RIS or AI deployment platform)
Identification of integration problems with hospital systems
Target clinical problem
Target population (e.g. age, gender, ethnicity, co-morbidities)
Inclusion and exclusion criteria for target population (if applicable)
Potential biases related to audit population
Measures of expected performance (e.g. diagnostic and/or triage performance, true positive, false positive, true negative, false negative)
Description of perceived risks
Basis and process for diagnostic reference standard
Evaluation date range
Processing time
Exclusion rate (proportion and number of eligible cases)
Failure rate (the proportion of eligible cases the tool failed to work for as expected)
Measured performance of AI tool independent of human read
Measured performance of AI tool with human in the loop
Clinical outcome measures
Discrepancy rate (rate at which clinician and AI tool agree)
Testing for automation bias
Perceived clinical impact of AI per case – beneficial, neutral, or negative (such as recall rate, number of cases which require non-invasive or invasive follow-up testing, outcome of follow-up testing)
Acceptance by users
Acceptance by patients
Carbon impact of processing
Comparison to historical data prior to AI use
Other issues encountered
Recommendation and feedback to sites and vendors

This table provides some suggested topic areas that could be included in AI audits, which may differ depending on the type of tool being evaluated and the stage of deployment.
AI, artificial intelligence; PACS, picture archive and communication system; RIS, radiology information system.

Clinical audit can be used to explore issues of bias and fairness, comparing AI performance and patient outcomes stratified by demographic data such as age, sex, and ethnicity. Real-world evaluation can also assess technical performance, in particular measuring how often AI application fails to work at all on scans in the workflow. The frequency of model use as a percentage of eligible patients can be calculated, and the evaluation can explore the reasons for AI failure, such as whether the AI tends to fail on cases that radiologists assess as being more challenging to report. Information captured during evaluation could be very helpful for sharing practical lessons about what worked and what did not during deployment. This could cover system usability, the approvals process, integration with hospital systems, and examples of deviation of human–computer interactions from expected use. Other measures might include the time taken for training radiologists, radiographers, and the PACS team, and radiologists' perception of the integration into clinical workflow. In

Please cite this article as: Ross J et al., Beyond regulatory compliance: evaluating radiology artificial intelligence applications in deployment, Clinical Radiology, https://doi.org/10.1016/j.crad.2024.01.026

addition, patient perspectives on the use of AI for their diagnosis and treatment planning can be evaluated in order to understand the impact on patient trust.

There may be disease- and context-specific considerations for different AI tools and environments. For example, an AI tool that suggests diagnoses for CXRs performed in the emergency department may be read by emergency department staff prior to being reviewed by a radiologist. The evaluation will need to determine the appropriate benchmark against which the AI model should be measured, be that a newly qualified doctor reading of a scan or a consultant radiologist's report. Although some models may be able to be evaluated as simple yes/no diagnoses, others may be subtle, requiring assessments of likelihoods or heat maps. Several AI applications have in built tools for simplifying error reporting, such as a button to highlight cases where the AI response was incorrect (or particularly good), which should feed into evaluation.

This evaluation can be used to design longer-term post-marketing surveillance, to assess whether the AI continues to perform over time. Regular reporting and threshold alerts could be used to identify when there is a risk of data drift, such as a significant change in either the input data (such as from a new scanner) or the AI outputs. Special consideration should be given to assessing the impact of "automation bias", the propensity for humans to favour suggestions from automated decision-making systems even when they are incorrect.[42] Federated learning is increasingly considered in medical imaging AI,[43] and federated networks could be used to support evaluating algorithms across several institutions.[44]

These evaluations may not address some of the wider challenges that hospitals and institutions may have in adopting clinical AI, such as identifying whether current digital and informatics resources are able to deal with the additional demands of AI systems, the cost of the system in relation to improvements in care, or improving the involvement of patient and public involvement in the successful acceptance of AI in patient pathways and clinical workflows.

Conclusion

Streamlined and expedited implementation of AI apps may be accelerated through providing robust, multicentric, real-world clinical AI audit data of the applications in radiology practice. Independent audit going beyond regulatory compliance will help address safety and effectiveness concerns, allow for earlier identification of errors, and provide benchmarking. This could consist of evaluation of AI applications "in the lab" on specific validation datasets, and by prospective centrally coordinated audit in the clinical setting. Validation datasets based on the natural patient population should be used where available, but the use of enriched datasets may be practical to allow safe deployment and appraisal in a timely manner when facing a rapid deployment of AI tools. These evaluations could also provide health economic evidence for the cost effectiveness of these applications to support business planning. Given the exponential growth of possible AI applications, now is the time to agree a framework for evaluating these products with the key stakeholders.

Glossary

This paper includes a glossary of terms used in AI research (see Table 2), and an overview of risks and mitigations in clinical AI development and deployment (see Electronic Supplementary Material Fig. S1).

Table 2
Glossary.

Term	Definition
Artificial Intelligence (AI)	The capability of a machine to imitate intelligent human behaviour. In the context of radiology, AI can be used to analyse images and detect abnormalities, among other tasks
Model overfitting	A modelling error that occurs when a function is too closely fit to a limited set of data points. An overfitted model may perform well on training data but poorly on new, unseen data
Data drift	The change in input data distribution over time. This can lead to a decrease in model performance if the model is not updated or retrained to reflect the new data distribution
Hidden stratification	A situation where an AI model performs well overall but underperforms on certain subgroups. This can lead to significant failures in clinical settings
Benchmarking	The process of comparing business processes and performance metrics to industry bests or best practices from other industries. In the context of AI in radiology, this could involve comparing the performance of different AI apps against each other or the reference standard of practice
CE marking	A certification mark that indicates conformity with health, safety, and environmental protection standards for products sold within the European Economic Area
FDA approval	The Food and Drug Administration (FDA) approval signifies that the agency has determined that the benefits of the product outweigh the known risks for the intended use
Post-marketing surveillance	The practice of monitoring the safety of a pharmaceutical drug or medical device after it has been released on the market
Automation bias	The propensity for humans to favour suggestions from automated decision-making systems even when they are incorrect
Ground truth	The term refers to the accuracy of a dataset, or the certainty that the dataset is a true representation of the world's features. Ground truth is used as a standard to train models and evaluate their performance. It is crucial for supervised learning where the model learns from labelled data and for validating the results of unsupervised learning
Validation datasets	A sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters

Please cite this article as: Ross J et al., Beyond regulatory compliance: evaluating radiology artificial intelligence applications in deployment, Clinical Radiology, https://doi.org/10.1016/j.crad.2024.01.026

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

The authors thank Keith Tucker, chair of Beyond Compliance, as well as Oliver Reichardt and Tosin Olufeko from the Royal College of Radiologists for their comments and support with this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crad.2024.01.026.

References

1. Rockall AG, Shelmerdine SC, Chen M. AI and ML in radiology: making progress. *Clin Radiol* 2023;**78**(2):81–2. https://doi.org/10.1016/j.crad.2022.10.010.
2. FDA. Artificial intelligence and machine learning (AI/ML)-Enabled medical devices. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices (accessed 1 January 2024).
3. Saw SN, Ng KH. Current challenges of implementing artificial intelligence in medical imaging. *Phys Med* 2022;**100**:12–7. https://doi.org/10.1016/j.ejmp.2022.06.003.
4. Rockall A. From hype to hope to hard work: developing responsible AI for radiology. *Clin Radiol* 2020;**75**(1):1–2. https://doi.org/10.1016/j.crad.2019.09.123.
5. Ebrahimian S, Kalra MK, Agarwal S, *et al*. FDA-regulated AI algorithms: trends, strengths, and gaps of validation studies. *Acad Radiol* 2022;**29**(4):559–66. https://doi.org/10.1016/j.acra.2021.09.002.
6. Silkens MEWM, Ross J, Hall M, *et al*. The time is now: making the case for a UK registry of deployment of radiology artificial intelligence applications. *Clin Radiol* 2023;**78**(2):107–14. https://doi.org/10.1016/j.crad.2022.09.132.
7. Royal College of Radiologists. Overcoming barriers to AI implementation in imaging. https://www.rcr.ac.uk/our-services/artificial-intelligence-ai/overcoming-barriers-to-ai-implementation-in-imaging/(accessed 2 January 2024).
8. Vasey B, Nagendran M, Campbell B, *et al*. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *BMJ* 2022:e070904, https://doi.org/10.1136/bmj-2022-070904.
9. Liu X, Glocker B, McCradden MM, *et al*. The medical algorithmic audit. *Lancet Digit Health* 2022;**4**(5):e384–97. https://doi.org/10.1016/S2589-7500(22)00003-6.
10. van Leeuwen KG, Schalekamp S, Rutten MJCM, *et al*. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol* 2021;**31**(6):3797–804. https://doi.org/10.1007/s00330-021-07892-z.
11. Dreyer KJ, Allen B, Wald C. Real-world surveillance of FDA-cleared artificial intelligence models: rationale and logistics. *J Am Coll Radiol* 2022;**19**(2):274–7. https://doi.org/10.1016/j.jacr.2021.06.025.
12. Roberts M, Driggs D, Thorpe M, *et al*. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell* 2021;**3**(3):199–217. https://doi.org/10.1038/s42256-021-00307-0.
13. Wu E, Wu K, Daneshjou R, *et al*. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med* 2021;**27**(4):582–4. https://doi.org/10.1038/s41591-021-01312-x.
14. Colquitt J, Jordan M, *et al*. Artificial intelligence software for analysing chest X-ray images to identify suspected lung cancer. *Warwick Evidence. Early Value Assessment report commissioned by the NIHR Evidence Synthesis Programme on behalf of the National Institute for Health and Care Excellence.* 2023. https://www.nice.org.uk/guidance/hte12/documents/diagnostics-assessment-report.
15. de Vries CF, Colosimo SJ, Staff RT, *et al*. Impact of different mammography systems on artificial intelligence performance in breast cancer screening. *Radiol Artif Intell* 2023;**5**(3):e220146, https://doi.org/10.1148/ryai.220146.
16. 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**(11):e077348, https://doi.org/10.1136/bmjopen-2023-077348.
17. Kaushal A, Altman R, Langlotz C. Geographic distribution of US cohorts used to train deep learning algorithms. *JAMA* 2020;**324**(12):1212. https://doi.org/10.1001/jama.2020.12067.
18. Oakden-Rayner L, Dunnmon J, Carneiro G, *et al*. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proc ACM Conf Health Inference Learn Toronto Ontario Canada: ACM* 2020:151–9. https://doi.org/10.1145/3368555.3384468.
19. NHS England. Transformation directorate. The NHS AI lab. https://transform.england.nhs.uk/ai-lab/(accessed 1 January 2024).
20. Joshi I, Cushnan D. A buyer's guide to AI in health and care. https://transform.england.nhs.uk/ai-lab/explore-all-resources/adopt-ai/a-buyers-guide-to-ai-in-health-and-care/a-buyers-guide-to-ai-in-health-and-care/(accessed 2 January 2024).
21. Health AI Partnership. https://healthaipartnership.org/(accessed 1 January 2024).
22. Coalition for Health AI. https://www.coalitionforhealthai.org/(accessed 1 January 2024).
23. STANDING together. https://www.datadiversity.org/(accessed 1 January 2024).
24. Beyond compliance. https://www.beyondcompliance.org.uk/(accessed 1 January 2024).
25. Patel NG, Napier RJ, Phillips JRA, *et al*. The first knee prosthesis to go through beyond compliance: a new standard for the safe introduction of orthopaedic implants. *Surgeon* 2020;**18**(6):e27–32. https://doi.org/10.1016/j.surge.2020.06.005.
26. Lidgren L, Alriksson-Schmidt A, Ranstam J. Arthroplasty watch—beyond borders, beyond compliance. *BMJ* 2013;**346**(feb19 2):f1013. https://doi.org/10.1136/bmj.f1013.
27. Smith AJ, Dieppe P, Vernon K, *et al*. Failure rates of stemmed metal-on-metal hip replacements: analysis of data from the National Joint Registry of England and Wales. *Lancet* 2012;**379**(9822):1199–204. https://doi.org/10.1016/S0140-6736(12)60353-5.
28. Godlee F. The trouble with medical devices. *BMJ* 2011;**342**(may18 3):d3123. https://doi.org/10.1136/bmj.d3123.
29. Hwang TJ, Sokolov E, Franklin JM, *et al*. Comparison of rates of safety issues and reporting of trial outcomes for medical devices approved in the European Union and United States: cohort study. *BMJ* 2016:i3323. https://doi.org/10.1136/bmj.i3323.
30. Tucker K. How registry data can improve outcomes from joint replacement - a seminal paper. *Acta Orthop* 2020;**91**(3):230–1. https://doi.org/10.1080/17453674.2020.1763567.
31. Herberts P, Malchau H. Long-term registration has improved the quality of hip replacement: a review of the Swedish THR Register comparing 160,000 cases. *Acta Orthop Scand* 2000;**71**(2):111–21. https://doi.org/10.1080/000164700317413067.
32. UK Government. *Establishing a pro-innovation approach to regulating AI*. 2022 https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement (accessed 2 January 2024).
33. Benrimoh D, Israel S, Fratila R, *et al*. Editorial: ML and AI safety, effectiveness and explainability in healthcare. *Front Big Data* 2021;**4**:727856, https://doi.org/10.3389/fdata.2021.727856.
34. NICE. Artificial intelligence-derived software to analyse chest X-rays for suspected lung cancer in primary care referrals: early value assessment. Health technology evaluation. Ref.: HTE12. 28 September 2023. https://www.nice.org.uk/guidance/hte12 (accessed 1 January 2024).
35. NHS England. Transformation directorate. AI diagnostic Fund. https://transform.england.nhs.uk/ai-lab/ai-lab-programmes/ai-in-imaging/ai-diagnostic-fund/(accessed 1 January 2024).

36. Hickman SE, Woitek R, Le EPV, *et al*. Machine learning for workflow applications in screening mammography: systematic review and meta-analysis. *Radiology* 2022;**302**(1):88–104. https://doi.org/10.1148/radiol.2021210391.
37. Alabed S, Maiter A, Salehi M, *et al*. Quality of reporting in AI cardiac MRI segmentation studies—a systematic review and recommendations for future studies. *Front Cardiovasc Med* 2022;**9**:956811, https://doi.org/10.3389/fcvm.2022.956811.
38. Maiter A, Salehi M, Swift AJ, *et al*. How should studies using AI be reported? lessons from a systematic review in cardiac MRI. *Front Radiol* 2023;**3**:1112841, https://doi.org/10.3389/fradi.2023.1112841.
39. Daye D, Wiggins WF, Lungren MP, *et al*. Implementation of clinical artificial intelligence in radiology: who decides and how? *Radiology* 2022;**305**(3):555–63. https://doi.org/10.1148/radiol.212151.
40. Kim H-E, Kim HH, Han B-K, *et al*. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *Lancet Digit Health* 2020;**2**(3):e138–48. https://doi.org/10.1016/S2589-7500(20)30003-0.
41. 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;**2**(5):e200016, https://doi.org/10.1148/rycan.2020200016.
42. Dratsch T, Chen X, Rezazade Mehrizi M, *et al*. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology* 2023;**307**(4):e222176, https://doi.org/10.1148/radiol.222176.
43. Pati S, Baid U, Edwards B, *et al*. Federated learning enables big data for rare cancer boundary detection. *Nat Commun* 2022;**13**(1):7346. https://doi.org/10.1038/s41467-022-33407-5.
44. Rehman MH ur, Hugo Lopez Pinaya W, Nachev P, *et al*. Federated learning for medical imaging radiology. *Br J Radiol* 2023;**96**(1150):20220890, https://doi.org/10.1259/bjr.20220890.

Please cite this article as: Ross J et al., Beyond regulatory compliance: evaluating radiology artificial intelligence applications in deployment, Clinical Radiology, https://doi.org/10.1016/j.crad.2024.01.026