The medical algorithmic audit

Xiaoxuan Liu, Ben Glocker, Melissa M McCradden, Marzyeh Ghassemi, Alastair K Denniston, Lauren Oakden-Rayner**

Artificial intelligence systems for health care, like any other medical device, have the potential to fail. However, specific qualities of artificial intelligence systems, such as the tendency to learn spurious correlates in training data, poor generalisability to new deployment settings, and a paucity of reliable explainability mechanisms, mean they can yield unpredictable errors that might be entirely missed without proactive investigation. We propose a medical algorithmic audit framework that guides the auditor through a process of considering potential algorithmic errors in the context of a clinical task, mapping the components that might contribute to the occurrence of errors, and anticipating their potential consequences. We suggest several approaches for testing algorithmic errors, including exploratory error analysis, subgroup testing, and adversarial testing, and provide examples from our own work and previous studies. The medical algorithmic audit is a tool that can be used to better understand the weaknesses of an artificial intelligence system and put in place mechanisms to mitigate their impact. We propose that safety monitoring and medical algorithmic auditing should be a joint responsibility between users and developers, and encourage the use of feedback mechanisms between these groups to promote learning and maintain safe deployment of artificial intelligence systems.

Lancet Digit Health 2022

Published **Online**
April 5, 2022
https://doi.org/10.1016/S2589-7500(22)00003-6

See **Online/Articles**
https://doi.org/10.1016/S2589-7500(22)00004-8

*Joint senior authors

Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, UK (X Liu PhD, Prof A K Denniston PhD); **Department of Ophthalmology, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK** (X Liu, Prof A K Denniston); **Moorfields Eye Hospital NHS Foundation Trust, London, UK** (X Liu); **Health Data Research UK, London, UK** (X Liu, Prof A K Denniston); **Birmingham Health Partners Centre for Regulatory Science and Innovation, University of Birmingham, Birmingham, UK** (X Liu, Prof A K Denniston); **Biomedical Image Analysis Group, Department of Computing, Imperial College London, London, UK** (B Glocker PhD); **The Hospital for Sick Children, Toronto, ON, Canada** (M M McCradden PhD); **Dalla Lana School of Public Health, Toronto, ON, Canada** (M M McCradden); **Institute for Medical Engineering and Science and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA** (M Ghassemi PhD); **National Institute of Health Research Biomedical Research Centre for Ophthalmology, Moorfields Hospital London NHS Foundation Trust, London, UK** (Prof A K Denniston); **University College London, Institute of Ophthalmology, London, UK** (Prof A K Denniston); **Australian Institute for Machine Learning, University of Adelaide, Adelaide, SA, Australia** (L Oakden-Rayner MBBS)

Introduction

Advances in artificial intelligence have attracted substantial interest for their potential applications in health care, particularly systems based on deep learning and neural networks. A large body of literature has been published proposing solutions based on artificial intelligence or machine learning for disease detection, classification, or prediction, or even as therapeutic interventions, including titration of drug dosages or offering mental health support through artificial intelligence chatbots.[1,2]

In the past 2 years, there has been a shift in emphasis from reporting impressive performance results to active investigation of algorithmic errors and failure modes.[3–6] Indeed, the analysis of error cases is a minimum reporting requirement of the SPIRIT-AI (Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence) and CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) guidelines for clinical trial protocols and reports of artificial intelligence interventions.[7,8] This change in focus from evaluating the best performance an artificial intelligence system can achieve, to identifying the worst mistake it could make, aligns with the foundational maxim embedded in medical safety: first, do no harm. The question of medical artificial intelligence safety is being asked at a crucial time, when this is no longer a theoretical concern but an immediate issue as, increasingly, artificial intelligence systems are receiving regulatory approval and being implemented in clinical care.

Why are artificial intelligence systems different from other medical interventions? Concerns have been raised that, unlike other interventions, artificial intelligence systems can yield errors that are difficult to foresee or prevent, due to the very nature of these systems. Modern artificial intelligence systems, particularly those based on deep learning, establish complex and opaque mathematical relationships between the input data and the output predictions, with little to no human control over how predictions are generated. Although this gives rise to a powerful machinery for learning patterns in the data, there is also a considerable risk of learning spurious correlations: relationships that appear useful in training but are unreliable when applied to real-world data. For example, an artificial intelligence system might learn to detect surgical skin markings to diagnose skin cancer, rather than looking for features related to the lesion itself.[9] Importantly, the errors of artificial intelligence systems appear to be quite distinct from the errors of human experts. In medical imaging, the majority of human errors (60–70%) are related to perceptual failure, caused by factors such as the subtlety of visual findings, incomplete search of the entire image, and so-called satisfaction syndrome (in which finding an abnormality makes the reader less likely to find a second one).[10] By contrast, artificial intelligence is not susceptible to incomplete searches or satisfaction syndrome. In this context, it is entirely reasonable to expect that artificial intelligence systems of equal performance to human readers will produce different errors, which can lead to different clinical outcomes.

The concept known as the artificial intelligence performance gap can be caused by a variety of factors, including those related to the algorithm's development, the input data, and interactions with users and the deployment environment. During development, the model design and training strategies, as well as the choice of training data (eg, poorly labelled or under-representative data) can directly influence the algorithm's performance. Mismatch or incompatibility of input data used during deployment can arise from various types of dataset shift (including population shift, annotation shift, prevalence shift, manifestation shift, and acquisition shift).[11] Interactions with users and the deployment environment are subject to automation bias, human error, and unintended or intended misuse.[12] Additionally, the reasons for unexpectedly poor performance can be

Correspondence to:
Dr Lauren Oakden-Rayner, Australian Institute for Machine Learning, University of Adelaide, Adelaide, SA 5000, Australia
lauren.oakden-rayner@adelaide.edu.au

non-obvious even after human inspection, and subtle or even unnoticeable differences in the input data might lead to catastrophic failure. This occurrence relates to the underlying mathematical approximations that artificial intelligence systems use to map input data (eg, a medical scan) to target outputs (eg, a diagnostic label). Generally, we can assume that artificial intelligence systems will operate well within the space mapped out by the training data (a process called interpolation), but perform poorly on out-of-distribution data that require extrapolation. Intuitively, the further an input sample is away from the statistical distribution of the training data, the more unpredictable becomes the behaviour and outputs of the artificial intelligence system. Unfortunately, given the complexity of most medical data, it can be difficult to define which cases are in-distribution and which are out-of-distribution. Furthermore, this drop in performance might not be obvious at the aggregate level of typical artificial intelligence testing, but rather in subsets of the target cohort or specific strata within the input data, a concept that has been described as hidden stratification.[13] These factors all contribute to the performance gap between preclinical testing and real-world deployment, and current evaluation strategies are ill-suited to identifying the problem.[14]

We define algorithmic errors as any outputs of the artificial intelligence system which are inaccurate, including those which are inconsistent with the expected performance and those which can result in harm if undetected or detected too late. Within these, there is a category in which the output might be correct but the algorithm is clearly informed by a flawed decision-making process. We suggest that these should also be considered algorithmic errors, because they indicate a high risk of future errors and should therefore be treated with similar levels of caution to standard output errors. If there is a pattern or systematic nature to the occurrence of errors, we refer to this as a failure mode: the tendency to malfunction in the presence of certain conditions. Whereas an error can be a single occurrence, failure modes represent errors which will repeatedly occur and often have similar consequences. Although individual errors might not always result in direct harm, their frequency or the summation of multiple errors could reach above an acceptable threshold and result in overall harm. By proactively investigating algorithmic errors and failure modes, the auditor becomes better placed to monitor artificial intelligence systems effectively and to understand the potential failure modes and their consequences.

Here, we propose an audit-based approach for investigating algorithmic errors. An algorithmic audit focuses on developing processes and embedding organisational principles and values in the algorithm design, and these values can vary widely depending on the organisation and context of the deployment. In medical artificial intelligence, the audit process focuses closely on the safety and quality of medical systems, the outcomes and perceptions of the patient and the general public, the responsible use of health-care resources, and the equitable distribution of health care and health-care outcomes.

Principles underpinning the medical algorithmic audit

The importance of safety and quality for medical algorithms is embedded in the principles of medical ethics, which describe the obligations of clinicians to patients and the general public. Evidence-based practice reflects the ethical imperative to act to promote the patient's best interests (beneficence) while minimising harm (non-maleficence), with empirical data forming part of the foundation upon which these judgments are made in consort with patient values. Typically, the information gathered through the process of prospective evaluation is contextualised to a clinical setting based on factors relating to each individual patient.[15] For interventions like drugs, the intervention itself is identical for all iterations (eg, the chemical structure of a single pharmaceutical agent is the same for every patient who takes it) and it is within individuals that responses vary. With artificial intelligence systems, the intervention is acutely sensitive to between-individual and within-individual feature variation, because the very power of the computational technique is in its ability to use feature variations to make individual-level predictions. However, artificial intelligence systems cannot apply clinical knowledge and domain expertise (including previous experience, contextual understanding, and causal knowledge) or common sense to distinguish between relevant feature variation due to disease versus irrelevant feature variation due to other biological confounders or non-biological sources, potentially resulting in unreliable predictions. Therefore, to translate algorithms into clinical practice, more nuanced information is required on the algorithm's performance across a range of relevant features, which is the goal of medical algorithm auditing. This information then guides effective and beneficial translation of interventions.[15]

An often overlooked concern with artificial intelligence is that of fairness. As long as bias and social determinants of health exist, these patterns will entrench themselves within health care machine learning. In many cases, the performance of an artificial intelligence model differs across patient identities or social determinants of health (often proxies for identities), which can pose a threat to another core ethical principle: justice. In this case, we might consider distributive justice as a desirable property of artificial intelligence-enabled care delivery (ie, whether the benefits afforded by machine learning are conferred equally to all). Distributive justice also points us to the necessity of redressing disparities. If an audit reveals disparate performance among certain groups, compensatory mechanisms might help to ensure these patients are not disadvantaged by use of the algorithm. Medical algorithmic auditing can reveal areas in which these mechanisms are required and point to how potential disadvantages may be redressed (panel 1).

Elements of a medical algorithmic audit

Here, we build on the algorithmic audit approach proposed by Raji and colleagues,[19] who describe a qualitative structured audit process applying the SMACTR framework (scoping, mapping, artifact collection, testing, and reflection) to artificial intelligence, as a general purpose technology. Although this framework was originally proposed as a way of assessing whether artificial intelligence development was conducted in alignment with the principles of an organisation, its structure is highly applicable to local auditing of artificial intelligence performance due to its orientation towards internal auditing (and thus led by those closest to implementation). Each step of the SMACTR framework has its own set of documentation requirements, thus facilitating accountability and iterative, ongoing safety monitoring. There is also emphasis on other established auditing practices in medicine and other industries, including process mapping, failure modes and effects analysis (FMEA), risk prioritisation, and planning mitigating actions. We adapt this framework for use in medical artificial intelligence applications (figure 1) and approach the problem from two perspectives: that of the developer, who can modify the artificial intelligence system in response to audit results; and that of the user, who cannot modify the artificial intelligence system but has the means to implement risk mitigation plans specific to the deployment setting. We apply the principles of the FMEA tool, a known mechanism in engineering, to facilitate assessment, prioritisation, and mitigation of risk. For illustrative purposes, an example of an audit for a hip fracture detection algorithm is published as supplementary information in a study by Oakden-Rayner and colleagues,[20] alongside a detailed breakdown of the FMEA. The benefits of performing the FMEA is to initiate and guide a critical thought process, rather than to establish whether the artificial intelligence system is acceptable or unacceptable or to provide certainty that all risks can be anticipated and minimised. FMEA has previously been applied to clinical settings, although it must be interpreted with care due to issues around reproducibility and incompleteness.[21]

The medical algorithmic audit might be conducted by developers but it is also likely to be conducted by stakeholders with no involvement in algorithm design, such as health-care workers. During deployment, a myriad of human factors combined with a poor understanding of artificial intelligence systems could create a situation in which all errors are assumed to be a fault of the algorithm's design. Therefore, clinical auditors must have the necessary tools to identify error sources that are preventable (input data factors or user factors) and not preventable (factors that are intrinsic to the algorithm itself). Although those outside the development team might have no opportunity to change the algorithm, they might be able to control or influence the circumstances under which it is deployed, which is intrinsically tied to the likelihood of errors, as well as the ability to avoid them. To consider the medical algorithmic audit from both perspectives, table 1 describes tasks undertaken by users and developers for each audit step.

Panel 1: Why should we monitor for algorithmic errors?

Within the broader mandate to ensure artificial intelligence systems are safe, undertaking regular systematic analyses of the observed errors is helpful for a number of reasons:

- It is an essential component of safety monitoring and adverse event reporting.[16,17]
- It allows quantification of risk for the artificial intelligence system, which can be weighed against the potential benefits, to inform decision making around whether it is appropriate to apply the model clinically. Benchmarks might already exist within clinical practice (eg, estimated human radiologist error rates for a diagnostic task), which would inform the risk–benefit ratio for deploying the artificial intelligence system.
- It could reveal unknown failure modes of the artificial intelligence system, such as a tendency to produce higher error rates in certain populations, diseases, or settings, or in the presence of specific input data characteristics.[9,11,18]
- Before deployment, it can be used to derive a measurable adverse event rate, which can inform how closely safety monitoring and post-deployment auditing should be performed. It can also provide a baseline measurement against which ongoing performance can be benchmarked.
- It can inform risk mitigation strategies so that those overseeing deployment of the artificial intelligence system can anticipate errors if the conditions known to trigger failure do occur, implement measures to avoid failure modes, and pre-emptively set hard stop thresholds in high-risk situations.
- It can provide valuable feedback and information for future artificial intelligence development and model improvement, and also highlight the potential need for post-deployment calibration or localisation of artificial intelligence systems.
- It can reveal systematic differences in performance across features mapping onto a protected identity or social determinant. Insight into these performance differences can prevent a systematic disadvantage to those groups resulting from the implementation of the algorithm.

Scoping

Scoping is the process of defining the intended purpose of the artificial intelligence system and anticipating potential harms. In the work by Raji and colleagues,[19] the framework is intended for any domain in which artificial intelligence might be applied. In the setting of medical artificial intelligence, the scope of the audit is more clearly defined: the ethical and clinical motivation is common across medical artificial intelligence studies, with the intention to improve health-care outcomes (ie, quality and length of life, financial outcomes, and organisational outcomes) and to promote distributive justice. Therefore, scoping in medical algorithmic audit should focus on two key elements: the intended use and the intended impact.

The intended use is a regulatory requirement[22,23] that describes how the algorithm (as a medical device) is to be

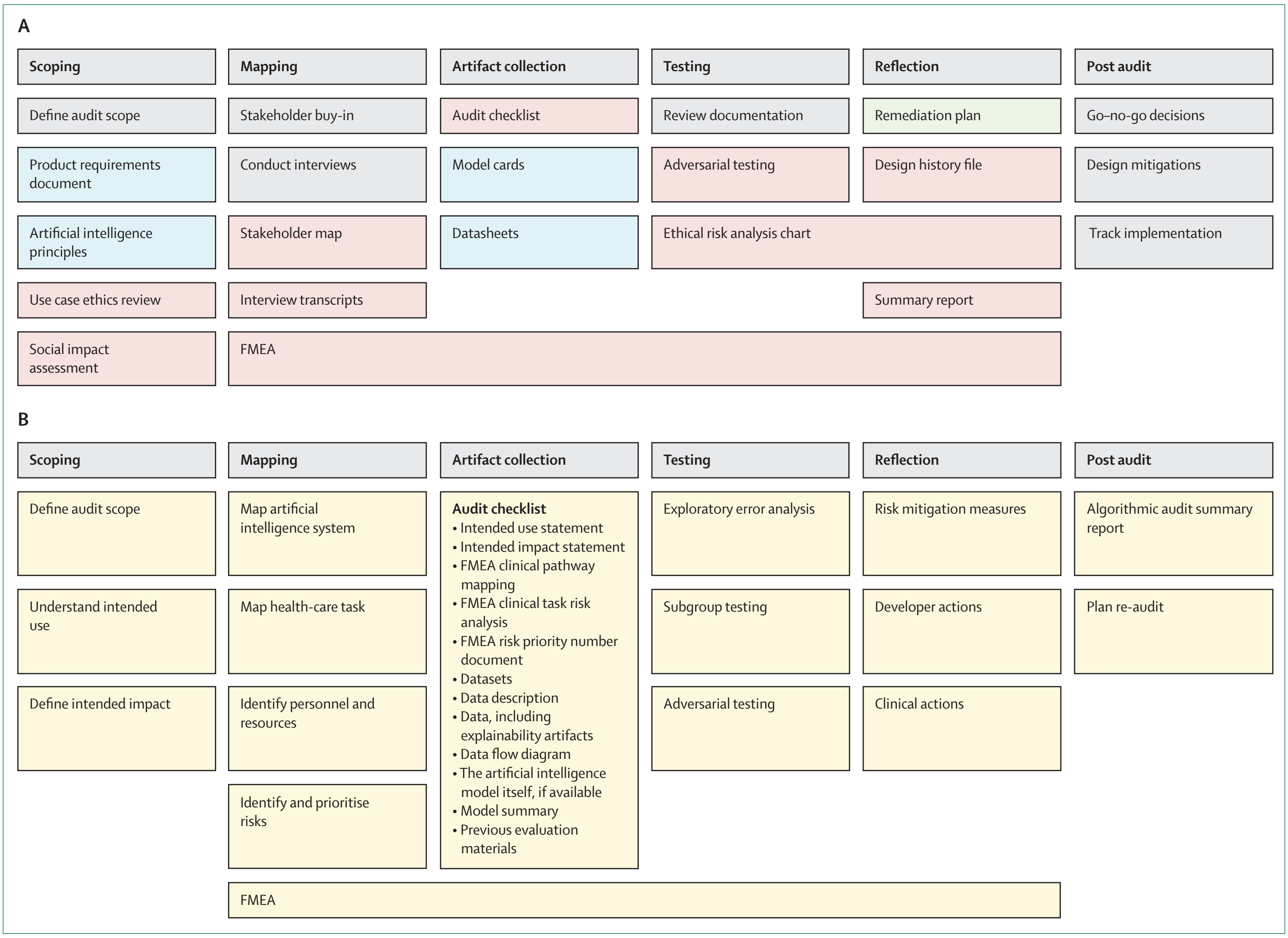

Figure 1: **Overview of the medical algorithmic audit**
(A) Overview of the internal audit framework, reproduced from Raji and colleagues.[19] Grey boxes represent processes; red boxes represent documents produced by the auditors; blue boxes represent documents produced by the engineering and product teams; and green boxes represent jointly developed documents. (B) Proposed modifications for the medical algorithmic audit. FMEA=failure modes and effects analysis.

applied. The US Food & Drug Administration premarket approval guidance states: "Indications for use for a device include a general description of the disease or condition the device will diagnose, treat, prevent, cure, or mitigate, including a description of the patient population for which the device is intended. Any differences related to gender, race/ethnicity, etc should be included in the labeling." The intended use specification is defined by the developer, who has knowledge of any previous evidence supporting indications for legal and safe use. It should also be known to the user, who decides whether the intended use statement matches the clinical task and clinical pathway in which the algorithm is intended to be deployed. For example, in the hip fracture audit,[20] scoping of the intended use refers to the function of the algorithm (detecting proximal femoral fractures) as well as its integration into a clinical pathway (in which detection leads to admission under an orthopaedic team and booking of further imaging if necessary). Other considerations include any limits on the health-care environment for use (eg, inpatient or outpatient) and the intended users or oversight (eg, health professionals, patients, or autonomous). A clear understanding must be established as to whether the current application falls within the artificial intelligence system's intended use, or if there are areas of ambiguity (eg, from missing or poorly defined intended use descriptions). Identified mismatches can motivate a targeted error analysis during the algorithmic audit.

The intended impact identifies the clinical or health-care target of the artificial intelligence system, accompanied by the ensemble of information that describes the boundaries within which the system is efficacious.[15] This statement describes how the artificial intelligence system will affect health-care outcomes if it is

	Developer and user actions	Developer actions	User actions
Scoping	..	Define intended use Anticipate intended impact(s)	Identify intended use Define intended impact(s)
Mapping	Mapping of the artificial intelligence system Define data flow Summarise risks in a risk priority number	Identify known risks of the artificial intelligence system in existing published and unpublished evidence and through knowledge of the training data	Identify known risks of the artificial intelligence system in existing published evidence Identify known risks of the health-care task Mapping of the health-care task, including elements before and after the artificial intelligence system in the clinical pathway, such as: relevant patient or data subgroups; potential sources of atypical input data; and relevant outcomes to be measured and how they will be captured in the audit
Artifact collection	Intended use statement Intended impact statement Failure modes and effects analysis: clinical pathway mapping, clinical task risk analysis, and risk priority number document Data flow diagram The artificial intelligence system Model summary Data for direct assessment, including explainability artifacts and adversarial testing artifacts Previous evaluation materials (including performance testing or user experience artifacts)	Datasheet for datasets (training and test data)	Datasheet for datasets (deployment data)
Testing	Exploratory error analysis: false positives and false negatives, and explainability methods (saliency maps and feature weights) Subgroup testing: patient-specific subgroup analysis and task-specific subgroup analysis	Adversarial testing	Adversarial testing if possible
Reflection	Compile algorithm audit summary report and share with relevant stakeholders	Risk mitigation actions: retrain the model, modify the model threshold, or modify the workflow or intended use	Risk mitigation actions: continue use with additional human oversight, modify or limit use, or withdraw use

Table 1: **Actions for developers and users at each stage of the medical algorithmic audit**

working as intended. The developer might be able to define, in theory, the intended impact, but the user is better placed to consider this. The hip fracture audit[20] has several intended impacts, including reduced time to admission and surgery, reduced resource use in the emergency department and unnecessary imaging, and downstream improvement in health outcomes. Different users of the same algorithm might have different target impacts specific to their health setting and needs. They might choose to implement the algorithm in different ways to produce different results and therefore their measures of success (and failures) will also be different. The auditor should define any unacceptably high-risk outcomes or adverse events; such events in medical safety are distinct, because they are considered so severe that they should never occur, such as surgical procedures performed on the wrong limb. It could be helpful to consider possible risks in the context of non-artificial intelligence systems with similar intended use and intended impacts.

Both the intended use statement and intended impact statement are used during the next phase (mapping), because they define the scope of algorithmic errors related to use of the artificial intelligence system.

Mapping

The mapping phase considers two main topics: the mapping of personnel and resources necessary for the audit, and the mapping of the risks and known vulnerabilities of the intended use as the first stage of the FMEA.

Personnel who might be helpful for a medical algorithmic audit include developers, users, and domain experts, particularly those who have experience with the artificial intelligence system. Developers have a substantial role to play in terms of providing periodic evaluations to guarantee expected performance, as is the case with other medical devices such as medical scanners, which often include 24-h service and support plans to ensure the device continues to meet operational, regulatory, quality, and safety requirements. If possible, developers should design mechanisms which allow users to carry out audits independently, at a local level. Resources required include, but are not limited to, access to suitable training or testing data and the associated labels (including non-target labels such as demographic information and hospital process factors); access to model predictions on the test data; access to any interpretability tools produced for use with the artificial intelligence model; and access to the model itself if a more in-depth introspection or further data challenges (such as adversarial testing) are required.

FMEA is a prospective risk analysis tool which first maps out a process or task, and then is used to identify foreseeable failures that could occur. In the mapping phase, there are two important elements of FMEA: mapping of the artificial intelligence system itself and mapping of the health-care task.

Panel 2: The risk priority number

The risk priority number is an arbitrary value calculated through ranking and combining three elements: severity (severity of the failure effects), occurrence (likelihood of occurrence), and detection (effectiveness of mechanisms to detect the failure before it results in adverse consequences). The ranking of each item is subjective, but a scale should be defined so that risk priority numbers in future audits can be compared.

For example, in the hip fracture audit,[20] severity and occurrence were scored between 1 to 4. A severity score of 4 was catastrophic (failure could cause injury or death, and extreme loss of trust) and 1 was minor (no intervention needed and no injury to patient). An occurrence score of 4 was frequent (several times in 1 day) and 1 was remote (maximum frequency of once in 6 months). The auditor should decide whether all three elements (severity, occurrence, and detection) are applicable to the artificial intelligence system, and whether there are additional elements that should be added. In the hip fracture audit, only severity and occurrence elements were included because most risks could not be easily detected in the current clinical workflow (resulting in homogeneous scoring).

Mapping of the artificial intelligence system itself is a detailed expansion on the intended use statement and analysis of previous evidence documenting risks intrinsic to the artificial intelligence system by design, or which the artificial intelligence system has encountered previously. It could include an evaluation of the existing literature on known risks, or a scoping of other artificial intelligence systems with similar intended use for potential risks. Mapping the artificial intelligence system also involves mapping any prerequisite steps or minimal requirements that are essential to achieving expected performance. Crucial to this is the process for handling and selecting input data, which is sometimes underspecified and poorly reported.[1]

Mapping of the health-care task is a contextualised analysis of the artificial intelligence system as a component of clinical care. It requires clinical knowledge of the use-case, the clinical workflow (including existing safeguards for detecting errors), user behaviour (health-care provider, patient, and the general public), and knowledge about potential consequences of errors. This knowledge can complement mapping of the artificial intelligence system, to anticipate when and how failure modes can arise. Mapping the clinical pathway can identify upstream factors that might increase the chances of algorithmic error, and downstream consequences that could occur because of algorithmic error. It also involves identifying important patient or data subgroups and specific features of the input data that are atypical.

It can be helpful to map the artificial intelligence system in relation to the clinical task and intended impacts in the form of a causal diagram, to determine the direction of causality between variables measured in the audit.[24] This will inform the metadata required for the artifact collection phase and can help auditors in making sense of relationships between relevant components of the health-care task in the reflection phase.

The risks which are mapped out are summarised in a risk priority number, which ranks the identified risks (panel 2). It is crucial to understand that the actual risk priority number value is not a measure of safety, nor should there be an attempt to create arbitrary thresholds to determine the acceptability of risks. Rather, the risk priority number enables relative ranking of all risks to prioritise those which need urgent attention and to serve as a baseline for re-evaluation in future audits.

Artifact collection

The artifact collection phase involves gathering documents and materials identified in the mapping phase that might inform the audit (table 1). There are three components to consider for medical artificial intelligence systems (aside from those already identified in scoping and mapping): relevant datasets (training data, previous evaluation data, or prospectively collected audit data for the current audit); the model itself; and results of previous evaluations of the model or task.

The datasets are of primary importance in determining both the performance of the artificial intelligence system, and the potential limitations and failure modes. Various datasets are used throughout the development, evaluation, and monitoring of artificial intelligence systems, and all are relevant for the algorithmic audit, but they might reveal different information about errors and failure modes. The relevant datasets are the algorithm training data (for developing the algorithm, which might include data for internal validation), previous test data (for evaluation or validation of the algorithm), and deployment data (data generated as a by-product of the algorithm being used). Both the test data and deployment data can be used in an algorithmic audit, but the information provided within them might vary. Note that evaluation data, and in particular labelled evaluation data, can be difficult or impossible to obtain in live deployment situations (or in certain evaluation designs, such as randomised controlled trials of effectiveness), in which the ground truth for each case is not routinely collected. In these settings, the identification of sources of weak labels (such as adverse event registers and user feedback) will be important, and the limitations of these labels should be clearly indicated.

There is often little relationship between errors on the training set and errors that occur during deployment; therefore, access to the complete training data is a low priority in an algorithmic audit. Although training data might be used for conducting exploratory error analyses (discussed later), deep learning models can achieve negligible training errors but still perform poorly in a test or deployment environment. Access to training data can also be problematic for users and external auditors, given the commercial value of these data and the sheer size of datasets.

Although direct access to the relevant data is likely to be useful, understanding the data processes is equally important. This information can be formalised with a datasheet[25] and a data flow diagram. A datasheet provides an extensive description of the data generating process, dataset collection, dataset composition, and dataset processing and labelling. Datasheets can be extremely valuable during an audit, because the dataset composition (in particular, the training data composition) can suggest likely failure modes (for example, patient subgroups that are under-represented in the training data). Access to datasheets rather than the dataset itself should not be problematic and should be provided by the developers whenever possible. In addition to the datasheet, a data flow diagram should be available, which outlines the handling of data from point of acquisition to presentation to the algorithm. This flow diagram should include any preprocessing steps, such as data transformation and normalisation, as well as exclusions based on data quality and a traceability mechanism for unusable or discarded data.

The model itself is also important in the audit process. Basic information about the model design, version, and model developers should be collected as a minimum. Such information can be summarised in a model card.[26] If the artificial intelligence system consists of multiple components (eg, a segmentation step followed by a classification step[27]), artifacts should be collected for each individual component. If multiple audits have been conducted over time spanning updated versions of the artificial intelligence model, documentation regarding changes between updates and any published evaluations since the last audit should also be collected.

Although direct access to the model code and parameters (known as a white box audit) can be useful—for example, by performing stress-testing of the artificial intelligence system by intentionally modifying input data to induce errors—this is rarely possible due to intellectual property concerns. Most of the benefit that access would provide can be equally obtained with the ability to test the model on new cases and receive model outputs, usually via a web portal or application programming interface (known as a black box audit). Developers should provide such a mechanism for users to perform independent local testing using representative data samples, to ensure performance is as expected.

The evaluations performed previously on a given artificial intelligence system are extremely important during the preparation of an algorithmic audit. Typically, medical artificial intelligence development goes through several phases of evaluation, and artifacts of this process include internal and external evaluation summaries, published materials on preclinical and proof-of-concept testing, and summaries of any previous qualitative assessments or audits. Qualitative assessments might include developer and user experience materials, such as interviews, surveys, or other forms of feedback. Any results from previous explainability methods, such as saliency or attention maps, per-case feature importance measures, or feature visualisations, should also be collected at this stage.

In the context of the hip fracture audit,[20] the components of the scoping and mapping phases were all collated, but additionally, the auditor secured access to the validation and test datasets with explainability artifacts or saliency maps for these cases, the hip fracture model itself, and documents related to model development and previous testing,[28,29] including design documents for each component of the algorithm.

Testing

The most important part of the audit process, other than the implementation of recommendations, is the testing phase. It is also the hardest part of the process to standardise, because each artificial intelligence system faces different risks. Institutions are accountable for the choice to incorporate artificial intelligence systems into their clinical pathways, which necessitates the need to ensure their appropriateness and functionality for the patients they serve. Should an algorithm not perform as expected or if harm were to occur, an audit would provide a clear mechanism of showing institutional accountability.

We suggest several key components of testing of medical artificial intelligence systems during an algorithmic audit: exploratory error analysis, subgroup testing, and adversarial testing.

Exploratory error analysis

The auditors will review each example of algorithm error that has been provided (either from previous evaluations, or from detected errors or adverse events in deployment). Auditors will systematically examine false positive and false negative groups in the case of classification systems, or outliers with high numerical errors in the case of regression models. The aim of this process is to identify common elements among the errors (ie, specific types of case that might be more prone to error, as shown in the hip fracture detection example in figure 2), as well as examples of surprising errors (eg, a fracture detection model missing an extremely obvious fracture, as shown in figure 2E, F). Given the contrastive nature of this method, access to cases correctly analysed by the artificial intelligence system can also be helpful.

Because this analysis is exploratory in nature, access to additional tools can be useful, which might require access to the algorithm itself or support from the algorithm developers. Examples of useful tools include artificial intelligence explainability methods, such as saliency

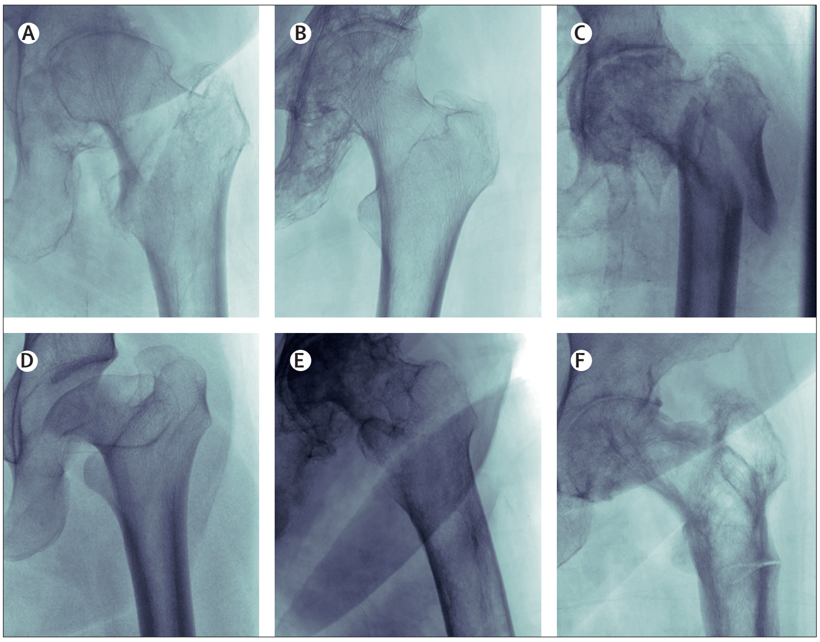

Figure 2: **Audit of a hip fracture detection system**
An example audit of a hip fracture detection system[20] showed that cases with abnormal bones or joints (Paget's disease and femoral head deformity) were overrepresented among the errors. The overall error rate was 2·5%, but the error rate for this subset was 50%. (A, B) True negatives. (C) True positive. (D) False positive. (E, F) False negatives.

maps and feature visualisations for image data, attention maps, feature weights, or importance measures for text and tabular data. Similarly, data clustering methods have been shown to be useful for some audit tasks, such as cryptic subset detection.[30] An example of the use of data visualisation is shown in figure 3A, where, in the absence of data normalisation, the largest modes of variation in brain MRI data after principal component analysis on the input images is between hospital sites. There is a high risk that a disease classification model trained on such data might pick up features associated with the site rather than the pathology (for example, if one site contributes more cases than controls). A careful data normalisation pipeline could mitigate such site differences, as shown in figure 3B. A model trained on the normalised data might be more robust when employed on new data. Although these exploratory tools are not powerful for risk assessment in isolation, they can be very useful during exploratory error analysis.

Subgroup testing

Subgroup testing, or secondary performance analysis, is widely used in medical and epidemiological research to investigate the possibility of confounding or stratification: patient or data variables that indicate a subset of cases in which performance will significantly differ from the overall cohort.

Importantly, subgroup analysis is not performed to test hypotheses. Given the reduction in power due to the lower sample sizes in subsets, as well as the inflated type 1 error rate (false positives) caused by multiple testing, these results should not be considered reliable or definitive in the same way that a primary analysis might be. Instead, the goal is to identify possible high-risk subpopulations within the target group. Although the subgroup analyses can be useful for identifying possible error patterns, these findings should be confirmed through investigation in a sufficiently powered sample.

There are three main forms of subgroup testing: patient-specific subgroup analysis, task-specific subgroup analysis, and exploratory error analysis-discovered subgroup analysis.

In a patient-specific subgroup analysis, the baseline characteristics table is used to describe important forms of variation in the dataset that might cause confounding relationships within the data or have wider implications on the results. Almost all studies report demographic and non-demographic characteristics of the study population, such as age, sex, ethnicity, socioeconomic status, disease severity, and comorbidities, and some studies also include information on geographical location and study site. Because these variables are often reported in the first table of a study manuscript, this is sometimes referred to as a table 1 subgroup analysis. These variables are highlighted during audit because they have well-known stratifying relationships with disease and treatment outcomes, and, if they are included in a table, the data are readily available and the subgroup analysis is straightforward to perform. Many studies already report this information[31–33] and an example is shown from the evaluation by Ting and colleagues[31] of a retinal imaging artificial intelligence system (table 2).

The number of possible confounding and stratifying factors in medical artificial intelligence evaluation is near to infinite. A task-specific subgroup analysis seeks to analyse the most concerning of these factors and is informed by the FMEA risk analysis and risk priority number. Like the patient-specific subgroup analysis, the subgroups in the task-specific subgroup analysis are defined prospectively, based on an understanding of the clinical task, often informed by domain experts. The main difference between a patient-specific subgroup analysis and a task-specific subgroup analysis is that the subgroups are often cryptic (unlabelled) in task-specific subgroup analysis. It is often necessary to undertake additional labelling to identify data that are part of the subgroup of interest. Because it might require considerable time and resources to undertake the labelling of relevant data, the risk prioritisation performed in the FMEA might inform which additional labelling should be prioritised. Factors that might be considered in task-specific subgroup analysis include collision groups (such as a combination of features from the patient-specific subgroup analysis) and process variables, such as the scanner used to obtain medical images or the presence of artifacts of medical care within the data (eg, a chest drain on a chest x-ray for a patient being treated for pneumothorax [figure 4],[13] or a surgical mark on the skin of a patient suspected of melanoma[9]). Special consideration should be given to data subgroups which would not be captured in a typical patient-specific subgroup analysis, such as visually distinct subsets in a medical image analysis task (eg, subsolid *vs* solid lung nodules).[34] The task-specific subgroup analysis might also be informed by important clinical implications associated with certain subgroups, such as the need for diagnostic certainty when differentiating infectious from non-infectious skin lesions (because the treatment for non-infectious lesions, topical steroids, will often worsen infectious lesions and might make subsequent diagnosis more difficult).[35]

During the exploratory error analysis process, distinct subgroups of error cases or error features might be identified, which are not considered during patient-specific subgroup analysis or task-specific subgroup analysis. Of note, whereas the subgroups in a task-specific subgroup analysis are defined prospectively based on expert knowledge, exploratory error analysis might identify new subgroups, which should then be evaluated as if for a task-specific subgroup analysis or a patient-specific subgroup analysis. However, unlike prospectively defined subgroups, subgroup cases identified during the exploratory error analysis are even less likely to be labelled and the auditor might need to invest time and resources to carry out further targeted labelling of the audit dataset. The risk priority number is helpful in this context, to help the auditor rationalise whether this investment is necessary. In the hip fracture detection audit,[20] the discovery of algorithmic errors in cases with abnormal bones prompted an additional labelling exercise of all hips with abnormal bones and joints to find that the error rate in those cases was 50%, compared with 2·5% in the overall dataset (figure 2).

Adversarial testing

Major unpredictable changes to input data are generally less of a concern in medical settings (in which data generation and processing are heavily standardised and monitored), but considering worst-case scenarios for targeted testing can be useful. The term adversarial here means actions that a hostile actor might take to break the artificial intelligence system, but in the medical context we could consider adversarial testing analogous with counterfactual reasoning, in which users can explore or simulate changes in data inputs to observe how the model behaves. This can be done in a safe environment to simulate high-risk situations and their potential consequences. For example, DeGrave and colleagues[36] used multiple adversarial testing approaches for a COVID-19 detection model for chest radiographs to show that the model can be misled by laterality markers and shoulder positioning.

Unlike subgroup analyses, adversarial testing might require access to the model itself. It could also be tested empirically through gathering real-world examples of specific subgroups for which performance is known to be poor, or by the use of simulated data that are expected to challenge the model. Although this is more common in tabular data, recent advances in generative models can allow for the simulation of more complex data, such as images and text. With either real-world or simulated data, the purpose of adversarial testing is to better understand the prevalence and source of errors in worst-case subgroups.

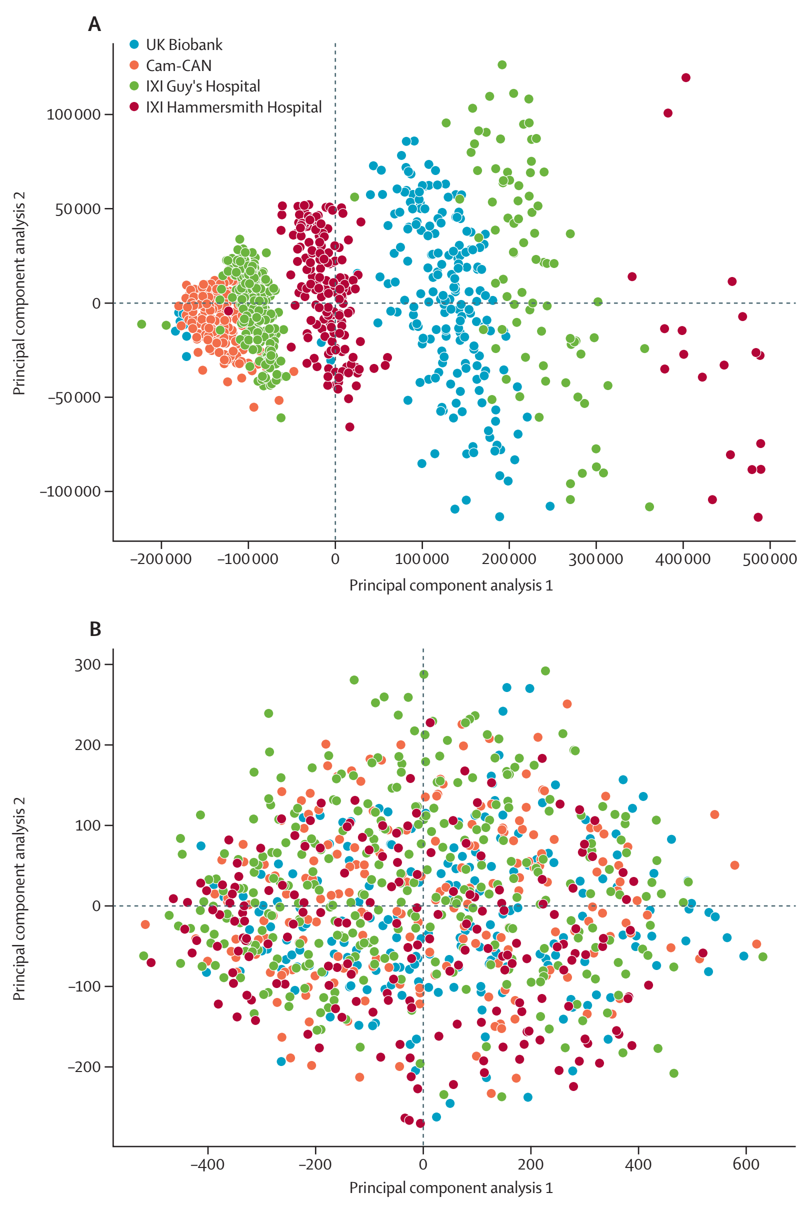

Figure 3: **Principal component analysis of brain MRI data, with (A) and without (B) data normalisation, across four hospital sites**

Cam-CAN=Cambridge Centre for Ageing and Neuroscience dataset. IXI=information extraction from images dataset. UK Biobank data were accessed under application number 12579; Cam-CAN data were made available upon application; the IXI datasets are publicly available.

	External validation dataset: non-referable eyes		External validation dataset: referable eyes					Diagnostic performance of artificial intelligence system in external validation		
	No diabetic retinopathy	Mild non-proliferative diabetic retinopathy	Moderate non-proliferative diabetic retinopathy	Severe non-proliferative diabetic retinopathy	Proliferative diabetic retinopathy	Diabetic macular oedema	Ungradable	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Community-based										
Guangdong	5665	1235	737	0	154	0	108	0·949 (0·943–0·955)	98·7 (97·7–99·3)	81·6 (80·7–82·5)
Population-based										
Singapore Malay Eye Study	1143	215	113	18	9	53	28	0·889 (0·863–0·908)	97·1 (95·1–99·9)	73·3 (70·9–75·5)
Singapore Indian Eye Study	1639	422	125	5	17	71	48	0·917 (0·899–0·933)	99·3 (95·1–99·9)	73·3 (70·9–75·5)
Singapore Chinese Eye Study	759	131	60	1	7	17	10	0·919 (0·900–0·942)	100 (92·5–100·0)	76·3 (72·7–79·6)
Beijing Eye Study	493	4	11	4	0	12	2	0·929 (0·903–0·955)	94·4 (72·7–99·9)	88·5 (85·4–91·2)
African American Eye Disease Study	807	50	37	5	16	28	41	0·980 (0·971–0·989)	98·8 (93·5–100·0)	86·5 (84·1–88·7)
Clinic-based										
Royal Victoria Eye and Ear Hospital	432	121	159	123	191	249	125	0·984 (0·972–0·991)	98·9 (97·5–99·6)	92·2 (89·5–94·3)
Mexican	38	284	192	51	18	223	3	0·950 (0·934–0·966)	91·8 (88·4–94·4)	84·8 (80·4–88·5)
Chinese University of Hong Kong	224	114	235	43	11	96	0	0·948 (0·921–0·972)	99·3 (97·3–99·8)	83·1 (77·9–87·3)
University of Hong Kong	1984	1485	155	14	0	214	1	0·964 (0·958–0·970)	100 (99·0–100)	81·3 (80·0–82·6)

These data are taken from a study by Ting and colleagues[31] and show the performance of their artificial intelligence model stratified by the clinical origin of the data. Reproduced with permission from *JAMA* 2016. **316:** 2402–10. Copyright © 2016 American Medical Association. All rights reserved. AUC=area under the receiver operating characteristic curve.

Table 2: **Example of a patient-specific subgroup analysis by dataset source or setting**

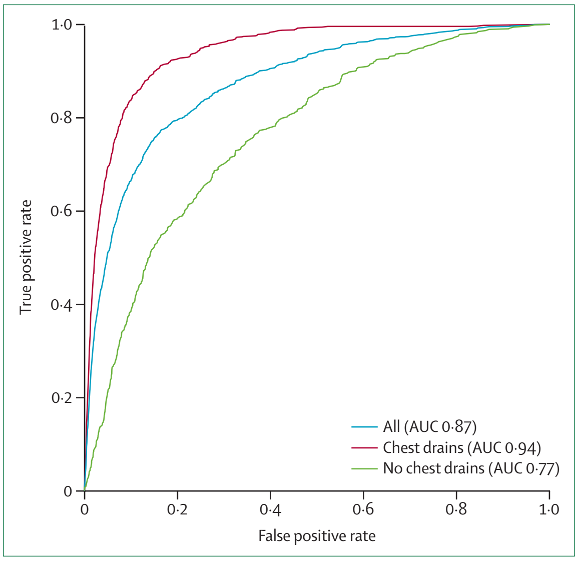

Figure 4: **Example of a task-specific subgroup analysis for a model detecting pneumothorax on chest radiographs**

In this example, reproduced from Oakden-Rayner and colleagues,[13] the artificial intelligence model learns to detect the artifacts of clinical care (chest drains) and fails to adequately learn the features of the pathology itself. AUC=area under the receiver operating characteristic curve.

Reflection

The final stage of the audit is a reflection on test results in light of the intended use and the intended impact as outlined in the scoping step. A final assessment of risk is formalised at this stage, risk mitigation strategies are proposed, and recommendations are made on whether the errors fall above the threshold for continued use of the artificial intelligence system. This decision will be specific to the clinical setting and the ability of the responsible team to mitigate risks. Those overseeing the algorithmic audit should be vigilant to deviations from the artificial intelligence system's intended use. It might become apparent during testing when a mismatch between intended use and actual use has occurred, but additional auditing measures such as root cause analysis might be required to retrospectively determine whether errors were due to gaps between the intended and actual use. In any case, errors should be reported to the relevant regulatory bodies, especially if the errors found invalidate the artificial intelligence system's intended use claims. It is also important to report errors even if adverse outcomes were mitigated through other measures in the health system (ie, near misses), because other deployment sites might not have the same mitigation measures in place.

The feasibility of risk mitigation strategies will be specific to the deployment setting and they require regular review as clinical systems change over time. The measures that can be put in place also depend on who the auditor is and which aspects of the artificial intelligence system and health-care system they are able to modify. Developers might be limited by legal requirements, such that substantial changes to the artificial intelligence model, deployment infrastructure, or the intended use could require reconsideration by regulatory agencies.

Developer actions

Potential risk mitigation strategies to be taken by developers include modifying the artificial intelligence model, modifying the model threshold, or modifying the instructions for use or intended use.

Modifications of the artificial intelligence system could target any part of the system, but they might involve targeted retraining of the model itself. In general, the intention would be to train an improved version of the artificial intelligence model using more diverse and representative data, targeting areas of weakness by enriching the training dataset with examples of cases associated with errors. If further data are not available, a similar effect might be achieved by reweighting training examples (amending the attributed value of one case versus another) or rebalancing the training data to increase the relative value of these cases (amending the number of positive and negative cases), or by producing simulated examples of these error cases.

The model threshold in classification systems determines the cutoff to discriminate between positive and negative cases, and it is also known as the operating point. This can be altered without retraining the model—for example, if user feedback suggests that the model produces too many false positives, shifting the threshold can reduce these (at the expense of increasing the rate of false negatives). The operating point of a model might be prespecified or suggested by the developers, but could also need tuning after deployment based on the specific clinical needs or performance at a particular site.

Modifications can also affect the non-model components of the deployed infrastructure and artificial intelligence workflow. These could include changes to data acquisition and preprocessing steps, or, in more extreme cases, modifying the intended use of the system. Such modifications could involve excluding some types of input data from the artificial intelligence system, changing how the model outputs are presented to the users, or even redefining the intended user group (eg, by increasing training requirements for users).

Clinical actions

Clinical actions can also be taken to mitigate risk, including continuing use with additional human oversight, modifying or limiting use of the model, or withdrawing its use. Some errors might be acceptable for continued use if the likelihood of harm is very low, or if the consequences can easily be mitigated given adequate human oversight. Depending on the use-case, reducing the level of autonomy of the artificial intelligence system and necessitating human verification might be sufficient to mitigate risks. In the FMEA, the risk prioritisation number could be informative because such errors would score low for severity or high for detection (or both). Human oversight might be implemented for all use or reserved for certain subgroups in which performance is known to be lower.

If modifiable risks are identified (eg, confounding visual features, such as laterality markers on chest radiographs), processes can be implemented to prevent recurrence (in this example, by standardising placement or the removal or digitisation of laterality markers in chest x-ray images). Thus, modifying the input data acquisition protocol or integrating additional pre-processing steps into the workflow to minimise the effects of spurious input data elements might be required.

If modifications are unfeasible or insufficient, limiting use of the artificial intelligence system on certain input data or subgroups that are prone to errors is another option. This strategy can be implemented if the subgroup can be identified upstream in the clinical pathway (eg, subgroups of certain demographic, input data type, or known task-specific feature variants that could be identified by the imaging technicians performing a scan) and those patients can be routed to an alternative care pathway. In the hip fracture detection audit,[20] the risk of false positives in cases of femoral deformities was found to require further monitoring, with a potential modification of the intended use (to exclude cases with deformities) as an appropriate mitigation action, if they were confirmed to cause a failure mode. However, this option is not always feasible if the subgroups are not readily identifiable prior to analysis by the artificial intelligence system. There should also be consideration as to whether such modifications inadvertently legitimise a two-tier health system, with particular groups receiving worse care, compromising the principle of distributive justice.

The last option is withdrawal of the artificial intelligence system altogether and reverting to previous care models. If the likelihood of errors is so severe that continued use of the algorithm is no longer safe or ethical, the only option is to stop use of the artificial intelligence system until modifications can be made. The potential harms of sudden and complete withdrawal of the artificial intelligence system should be weighed against the harms caused by continued use with or without modifications and limitations.

An alternative method to consider is the use of so-called hard stop thresholds, which are common components of medical device deployments.[37] These involve prespecified minimum performance levels, and if performance falls below the threshold during ongoing and active monitoring, then the device is immediately removed from use. These thresholds can be jointly agreed with all relevant stakeholders and informed by relevant organisational values (including equity and justice). Prespecification of actions can simplify the often complex decisions on whether to stop use if the hard stop threshold is met.

Algorithmic audit summary report

Findings of the medical algorithmic audit are summarised in a report, which includes all collected artifacts, the FMEA, datasets, test results, risk mitigation plans, and final decisions made. Any learning derived from the audit process that extends beyond the current application should be recorded to assist future artificial intelligence evaluations or deployments. Key audit findings that carry direct implications on clinical care should also be disseminated to users. Updates or changes made to the artificial intelligence system should be made apparent to the user, ideally with reasons reported. A frequent and open dialogue of findings from the algorithmic audit summary report should be shared between developers and users.

Conclusions

Artificial intelligence systems for health care could bring considerable benefits to patient care, but, like any other medical intervention, they also have the potential to cause harm. For artificial intelligence systems, the nature of errors might make them particularly difficult to identify, explain, and mitigate. At a time when artificial intelligence systems are being rapidly adopted into clinical care, providing a framework for ongoing performance monitoring and scrutiny of error and harm is essential. Implementation of artificial intelligence systems can be especially high risk, given that it often coincides with the establishment of new clinical pathways with no clear comparators for expected outcomes or standards for quality (such as the creation of new telemedical and virtual care pathways).

Of note, although many artificial intelligence systems are supported by evidence showing superior or equivalent performance compared with human experts, such monitoring of human performance is not routinely carried out in a task-specific fashion in clinical practice. In fact, whereas clinical artificial intelligence evaluations have provided valuable insights into human performance by measuring and benchmarking human performance at specific diagnostic tasks, routine monitoring of human grader accuracy, such as those introduced by UK national screening programmes for diabetic retinopathy[38] and breast cancer screening,[39] is not performed for most other clinical tasks. Gaining an understanding of human performance will not only reveal for which tasks artificial intelligence systems truly provide value, but should also drive improvements to the standards of care among clinicians.

The medical algorithmic audit proposed here is a process to investigate and pre-empt errors and harms that could be caused by artificial intelligence systems. It is a general framework which promotes thoughtful interrogation of errors and unexpected results in evaluations both before and during real-world deployment. Performing the audit requires clinical and technical expertise and contextual knowledge for anticipating the potential effects of the deployment environment, which might expose vulnerabilities of the algorithm and increase the likelihood of errors.

One question yet to be answered is who should conduct the medical algorithmic audit. The skills and knowledge required to undertake such an audit cross computational, bioinformatics, and clinical skill sets, and are not taught in combination in standard medical curricula. In order to fulfil this responsibility, health providers need to invest in upskilling clinical personnel to oversee the piloting, deployment, and ongoing monitoring of artificial intelligence systems, broadly described as the science of algorithmovigilance.[40] In the UK, the need to invest in digital leaders with the necessary capabilities (such as clinical information officers) has been recognised by the National Health System and Health Education England.[41,42] In Australia, the Royal Australian and New Zealand College of Radiologists have recommended that medical imaging departments and practices appoint a responsible radiologist with the necessary skills and knowledge to perform regular algorithmic audits of artificial intelligence systems in deployment.[43] In both nations, concerns have been raised that appropriately trained clinicians are rare, and that there remains considerable work to build an artificial intelligence-ready workforce. Structured processes and guidelines, such as the ones described here, are necessary to accelerate the development of quality and safety assurance capabilities for artificial intelligence in clinical settings.

Ultimately, the responsibility and benefits of investigating and improving the safety of the artificial intelligence system is shared between developers, health-care decision makers, and users, and should be part of a larger oversight framework of algorithmovigilance to ensure the continued efficacy and safety of artificial intelligence systems. For the medical algorithmic audit to have the highest chance of success, we advocate for the process being carried out jointly between these stakeholders, so that each party enables the other in developing a deeper and more contextualised insight into the findings and possible mitigation strategies.

Contributors

LO-R, XL, BG, AKD, MG, and MMM were responsible for project conception, study design, and drafting and review of the manuscript. LO-R did the example algorithmic audit.

Declaration of interests

BG is a part-time employee of HeartFlow and Kheiron Medical Technologies and holds stock options with both as part of the standard compensation package. He was a part-time employee at Microsoft Research until May, 2021 and a scientific advisor for Kheiron Medical Technologies until September, 2021. All other authors declare no competing interests.

Acknowledgments

We acknowledge Inioluwa Deborah Raji and her co-authors for their seminal work on internal algorithmic auditing,[19] upon which we based our medical algorithmic audit. We are grateful for her comments and feedback on our adaptation of this framework. XL and AKD receive a proportion of their funding from the Wellcome Trust, through a Health Improvement Challenge grant (200141/Z/15/Z). BG receives funding from the European Research Council under the EU Horizon 2020 research and innovation programme (grant agreement number 757173, project MIRA, ERC-2017-STG).

References

1 Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health* 2019; **1:** e271–97.

2 Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020; **368:** m689.

3 Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019; **25:** 1337–40.

4 Schulam P, Saria S. Can you trust this prediction? Auditing pointwise reliability after learning. *Proc Mach Learn Res* 2019; **89:** 1022–31.

5 Pooch EHP, Ballester P, Barros RC. Can we trust deep learning based diagnosis? The impact of domain shift in chest radiograph classification. In: Petersen J, Estépar RSJ, Schmidt-Richberg, et al, eds. Thoracic image analysis. Cham: Springer, 2020: 74–83.

6 Mahajan V, Venugopal VK, Murugavel M, Mahajan H. The algorithmic audit: working with vendors to validate radiology-AI algorithms—how we do it. *Acad Radiol* 2020; **27:** 132–35.

7 Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020; **26:** 1364–74.

8 Rivera SC, Liu X, Chan A-W, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Med* 2020; **26:** 1351–63.

9 Winkler JK, Fink C, Toberer F, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol* 2019; **155:** 1135–41.

10 Degnan AJ, Ghobadi EH, Hardy P, et al. Perceptual and interpretive error in diagnostic radiology—causes and potential solutions. *Acad Radiol* 2019; **26:** 833–45.

11 Du-Harpur X, Arthurs C, Ganier C, et al. Clinically relevant vulnerabilities of deep machine learning systems for skin cancer diagnosis. *J Invest Dermatol* 2021; **141:** 916–20.

12 Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. *J Am Med Inform Assoc* 2017; **24:** 423–31.

13 Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proc ACM Conf Health Inference Learn* 2020; **2020:** 151–59.

14 McCradden MD, Stephenson EA, Anderson JA. Clinical research underlies ethical integration of healthcare artificial intelligence. *Nat Med* 2020; **26:** 1325–26.

15 Kimmelman J, London AJ. The structure of clinical translation: efficiency, information, and ethics. *Hastings Cent Rep* 2015; **45:** 27–39.

16 US Food & Drug Administration. Guidance documents (medical devices and radiation-emitting products). https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/guidance-documents-medical-devices-and-radiation-emitting-products (accessed Dec 4, 2020).

17 European Union. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) no 178/2002 and Regulation (EC) no 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. 2017. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745 (accessed March 16, 2022).

18 Wang P, Berzin TM, Glissen Brown JR, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* 2019; **68:** 1813–19.

19 Raji ID, Smart A, White RN, et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. New York, NY: Association for Computing Machinery, 2020: 33–44.

20 Oakden-Rayner L, Gale W, Bonham TA, et al. Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. *Lancet Digit Health* 2022; published online April 5. https://doi.org/10.1016/S2589-7500(22)00004-8.

21 Shebl NA, Franklin BD, Barber N. Failure mode and effects analysis outputs: are they valid? *BMC Health Serv Res* 2012; **12:** 150.

22 US Food & Drug Administration. PMA labelling. 2018. https://www.fda.gov/medical-devices/premarket-approval-pma/pma-labeling (accessed Sept 29, 2020).

23 International Organization for Standardization. IEC 62366-1:2015. Medical devices—part 1: application of usability engineering to medical devices. 2020. https://www.iso.org/standard/63179.html (accessed Sept 29, 2020).

24 Castro DC, Walker I, Glocker B. Causality matters in medical imaging. *Nat Commun* 2020; **11:** 3673.

25 Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. *arXiv* 2018; published online March 23. http://arxiv.org/abs/1803.09010 (preprint).

26 Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. New York, NY: Association for Computing Machinery, 2019: 220–29.

27 De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018; **24:** 1342–50.

28 Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. *arXiv* 2017; published online Nov 17. http://arxiv.org/abs/1711.06504 (preprint).

29 Gale W, Oakden-Rayner L, Carneiro G, Palmer LJ, Bradley AP. Producing radiologist-quality reports for interpretable deep learning. In: 2019 IEEE 16th International Symposium on Biomedical Imaging. Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2019: 1275–79.

30 Sohoni N, Dunnmon JA, Angus G, Gu A, Ré C. No subclass left behind: fine-grained robustness in coarse-grained classification problems. *arXiv* 2020; published online Nov 25. https://arxiv.org/abs/2011.12945 (preprint).

31 Ting DSW, Cheung CY-L, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* 2017; **318:** 2211–23.

32 Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; **316:** 2402–10.

33 McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. *Nature* 2020; **577:** 89–94.

34 Ciompi F, Chung K, van Riel SJ, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Sci Rep* 2017; **7:** 46479.

35 Liu Y, Jain A, Eng C, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med* 2020; **26:** 900–08.

36 DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *medRxiv* 2020; published online Oct 8. https://doi.org/10.1101/2020.09.13.20193565 (preprint).

37 US Food & Drug Administration. Recalls, corrections and removals (devices). 2020. https://www.fda.gov/medical-devices/postmarket-requirements-devices/recalls-corrections-and-removals-devices#4 (accessed Nov 26, 2020).

38 Public Health England. Diabetic eye screening: participation in the grading test and training system. Nov 25, 2020. https://www.gov.uk/government/publications/diabetic-eye-screening-test-and-training-participation/diabetic-eye-screening-participation-in-the-grading-test-and-training-system (accessed Dec 9, 2020).

39 NHS Cancer Screening Programmes. Quality assurance guidelines for breast cancer screening radiology. Sheffield: National Health Service Cancer Screening Programmes, 2011.

40 Embi PJ. Algorithmovigilance—advancing methods to analyze and monitor artificial intelligence-driven health care for effectiveness and equity. *JAMA Netw Open* 2021; **4:** e214622.

41 Topol E. The Topol review: preparing the healthcare workforce to deliver the digital future. National Health Service Health Education England, 2019.

42 National Health Service. The NHS long term plan. 2019. https://www.longtermplan.nhs.uk/wp-content/uploads/2019/08/nhs-long-term-plan-version-1.2.pdf (accessed March 16, 2022).

43 The Royal Australian and New Zealand College of Radiologists. Standards of practice for artificial intelligence. 2020. https://www.ranzcr.com/documents-download/quality-and-standards-except-professional-documents/5187-standards-of-practice-for-ai (accessed Nov 25, 2020).