Clinicians must participate in the development of multimodal AI

Christopher R. S. Banerji,[a,b,c,d,*] Aroon Bhardwaj Shah,[e] Ben Dabson,[f] Tapabrata Chakraborti,[a,c] Vicky Hellon,[a] Chris Harbron,[g] and Ben D. MacArthur[a,h,i,**]

[a]The Alan Turing Institute, London, UK
[b]University College London Hospital, University College London Hospitals NHS Trust, London, UK
[c]UCL Cancer Institute, Faculty of Medical Sciences, University College London, London, UK
[d]King's Comprehensive Cancer Centre, King's College London, London, UK
[e]Whittington Hospital, Whittington Health NHS Trust, London, UK
[f]Hammersmith Hospital, Imperial College London NHS Trust, London, UK
[g]Roche Pharmaceuticals, Welwyn Garden City, UK
[h]Faculty of Medicine, University of Southampton, Southampton, UK
[i]Mathematical Sciences, University of Southampton, Southampton, UK

eClinicalMedicine
2025;84: 103252
Published Online 23 May 2025
https://doi.org/10.1016/j.eclinm.2025.103252

Summary

Multimodal artificial intelligence (AI) is a powerful new technological advance, capable of simultaneously learning from diverse data types, such as text, images, video, and audio. Because clinical decisions are usually based on information from multiple sources, multimodal AI has the potential to significantly improve clinical practice. However, unlike most developed multimodal AI workflows, clinical medicine is both a dynamic and interventional process in which the clinician continually learns about the patient's health and acts accordingly as data is collected. In this article we argue that multimodal clinical AI must be fully attuned to the particular challenges and constraints of the clinic, and clinician involvement is needed throughout development—not just at clinical deployment. We propose ways that clinician involvement can add value at each stage of the multimodal AI development pipeline, and argue for the establishment of actively managed multidisciplinary communities to work collaboratively towards the shared goal of improving the health of all.

Keywords: Multimodal AI; Human-in-the-loop AI; Health policy; Community management; Clinical AI

Introduction

Artificial Intelligence (AI) is a powerful, rapidly developing technology that is attracting increasing public interest. The ability of AI tools to learn complex relationships from data has raised hopes that they could optimise clinical decision-making and dramatically improve human health.

Among many notable recent advances, developments in so-called multimodal AI (Glossary - Box 1) have been hailed as initiating a new era for the technology.[1] Unlike more established AI tools, which are typically limited to one type of input data (such as text), multimodal AI can learn from a range of qualitatively different data types at the same time (such as text, images, and audio), thereby assimilating different sources of information and enabling better informed predictions. The free availability of vast quantities of online multimodal data (such as subtitled films) has driven the development of multimodal AI tools for non-clinical purposes.[2–4] In recent years, large, clinical multimodal databases—that link multiple medical data sources, such as genomics, histopathology, and radiology[5–7]—have entered the public domain, raising hopes that advances in multimodal AI may soon translate to the clinic.[8,9]

However, most state-of-the-art multimodal AI models are not built with the clinical setting in mind[10–12] and so are not attuned to the particular challenges that clinical decision-making presents.

Clinician involvement throughout the multimodal clinical AI development pipeline is therefore essential—to guide model design, construction, development, deployment, and iterative refinement. However, clinician involvement in the AI development loop is not standard practice. Indeed, a recent systematic review of AI tools that directly sought clinician input found that just 22% of studies involved clinicians throughout development, while the majority (82%) involved clinicians specifically in the deployment phase.[13]

This lack of inclusion is important, because clinician trust in AI is vital to its successful deployment[14] and studies have shown that detachment of clinicians from

*Corresponding author. The Alan Turing Institute, London, UK.
**Corresponding author. The Alan Turing Institute, London, UK.
E-mail addresses: cbanerji@turing.ac.uk (C.R.S. Banerji), bmacarthur@turing.ac.uk (B.D. MacArthur).

Box 1.

Glossary of terms.

Artificial Intelligence (AI)—a general term used to describe computational models which can perform tasks normally associated with human intelligence, including the ability to learn from data, recognise patterns, and make decisions.

Machine learning—the development and/or study of algorithms to learn from data. Machine learning algorithms are typically optimised on seen 'training' data to identify salient patterns that generalise to unseen 'test' data. Machine learning methods are very varied and range from classical statistical techniques, such as linear or logistic regression, that make basic assumptions concerning any patterns in the data, to modern data-intensive approaches, such as artificial neural networks, that can flexibly learn patterns of arbitrary complexity without guidance.

Artificial neural network—a class of machine learning algorithm, loosely modelled on the human brain, in which signals are passed between nodes (representing neurons) via edges (representing synapses). Signals are typically passed from an input layer of nodes (associated with the variables in the data) to an output layer (associated with possible outcomes) via 'hidden' layers of nodes.

Deep learning—the area of machine learning that makes use of artificial neural networks with many hidden layers. By encoding different patterns in each layer, deep neural networks can learn complex, multi-scale, relationships between variables and can be used to address an extremely broad range of learning problems.

Data modality—a data type e.g., text or images or audio.

Multimodal AI—the area of AI concerned with developing tools that can combine and/or learn from data of multiple modalities (e.g., text and images and audio).

Feature—a distilled representation of the data that affects predicted outcomes. Features depend on both the data and the problem at hand. For example, raised white blood cell count is a useful feature for diagnosing presence of an infection, but is less useful for distinguishing between different kinds of infection.

Training—the process by which a machine learning model is optimised to solve a specified problem, typically by tuning model parameters to learn the relationships between data features and outcomes from a training dataset.

Generalisation—the ability of a trained machine learning model to make accurate predictions from data that it has not seen before.

Dataset drift—a gradual change over time in the distribution of a dataset and/or the relationships between data features and outcomes. Under dataset drift machine learning models trained on old data may gradually provide less accurate or relevant predictions.

Dataset shift—a sharp change in the distribution of a dataset and/or the relationships between features and outcomes. Under dataset shift, a machine learning model may suddenly provide less accurate or relevant predictions.

the development process is a driver of mistrust.[15] Conversely, studies reporting high levels of clinician involvement throughout development often benefit from high clinician confidence in the deployed tool.[16]

For reasons that we will articulate below, we contend that clinician involvement in multimodal AI development is particularly vital and will be essential for this technology to achieve its full potential and have a major impact on human health.

We first outline some of the main current approaches to multimodal AI that have been developed for non-clinical applications but are starting to be repurposed to the clinical setting. We interpret these models via parallels to clinical decision making and highlight that to various degrees all are mismatched to the clinical context. We argue that without careful clinician guidance, repurposing of such tools will likely have limited clinical application, and motivate the development of more clinically attuned multimodal AI models. To meet this challenge, we next provide an outline of the multimodal AI development pipeline, indicating the important role of the clinician at each step. We close by presenting some strategies to support clinician involvement in multimodal AI development, and for fostering and nurturing an integrated clinical AI community.

Multimodal clinical AI requires clinician input throughout development

Clinical decisions are rarely made from consideration of a single data modality. An initial clinical assessment typically comprises a verbal history followed by physical examination. Subsequent investigations vary and may include relevant blood tests, radiological imaging, pathological assessments and, increasingly, high throughput molecular tests such as genomics.

Clinical medicine is thus fundamentally multimodal, and though significant insights have been gained from clinical AI models trained on a single data modality—such as detecting abnormalities in chest X-rays[17] or malignancy in digital histopathology images[18]—to maximise impact, multimodal approaches to AI are needed.

Development of multimodal AI requires large computational resources and significant funds, and major breakthroughs are therefore often achieved by big technology companies. Though some of these companies have a dedicated clinical focus,[19] most leading multimodal AI tools are developed for non-clinical purposes, such as enhancing web searching[10] and augmented reality.[20] In recent years, largely due to computational advances and the free availability of text, image, and video data online, these approaches have been advancing extraordinarily rapidly, and the leading current AI models are now able to combine text, images, and audio interchangeably.[10]

Clinical data, however, are neither freely available nor interchangeable. Rather, clinical data are obtained via investigations performed in a specific order to provide distinct pieces of information, on a careful balance of many variables. These include necessity, risk of harm, suitability for the patient, patient preferences, system pressures and increasingly, financial cost. This naturally introduces a hierarchy and personalisation to clinical data acquisition. Moreover, the dynamic process of clinical investigation naturally leads to a shortening in the range (or, statistically, a bias) of pathologies

represented by later diagnostic investigations, as the clinician homes in on the correct diagnosis.

To illustrate this point, consider two patients, both presenting with weight loss but with significantly different patient journeys (Fig. 1A). The first patient's weight loss is due to an underlying malignancy, necessitating numerous often invasive investigations to diagnose and manage appropriately. The second patient's weight loss, however, is simply the result of increased exercise and requires only the clinical history to manage appropriately.

These contrasting examples emphasise the importance of personalisation in clinical data acquisition. For multimodal AI to be relevant in the clinic, it must first and foremost capture the inherently personal nature of clinical decision-making and not only be able to guide decisions about patient management, but also decisions about which investigation should be performed next for each particular patient. Multimodal AI must therefore be able to support decisions in the absence of some data modalities, and be attuned to the clinical, financial, and patient specific constraints of data collection. Incorporating such information into clinical AI models will require significant clinician input throughout development, not only at deployment.

These examples also highlight that the foundation of all clinical decisions is the patient history and examination, while the number and type of investigations performed subsequently will likely correlate with the presence of specific pathologies. This important fact means that large linked multimodal clinical databases—in which many investigations are performed for each patient—will likely overrepresent patients with specific, often serious, outcomes, and underrepresent patients with either clear cut or very positive prognoses (Fig. 1B). Training of multimodal models on these databases will therefore require close collaboration with clinicians, to identify and understand these biases and determine the scope of their clinical utility.

Leading multimodal AI models are generally not currently constructed with appreciation of such data collection hierarchies or outcome biases in mind. Indeed, different data modalities are typically combined as though they were acquired simultaneously, via an approach called data fusion. Broadly speaking, there are three main types of data fusion—known as early, intermediate, and late fusion—all of which have been applied to model development in clinical contexts.[8,21] Although none explicitly account for clinical data collection hierarchies, all these techniques can be compared to clinical decision-making, and show various degrees of relevance to the clinical context (Fig. 2 & Box 2).

More recently, new approaches to multimodal AI model design have been developed, which combine data modalities in a sequential manner.[8,22] By introducing a natural order (e.g., by invasiveness or risk of patient harm) these approaches—which go by many names including gradual intermediate data fusion[8]—may

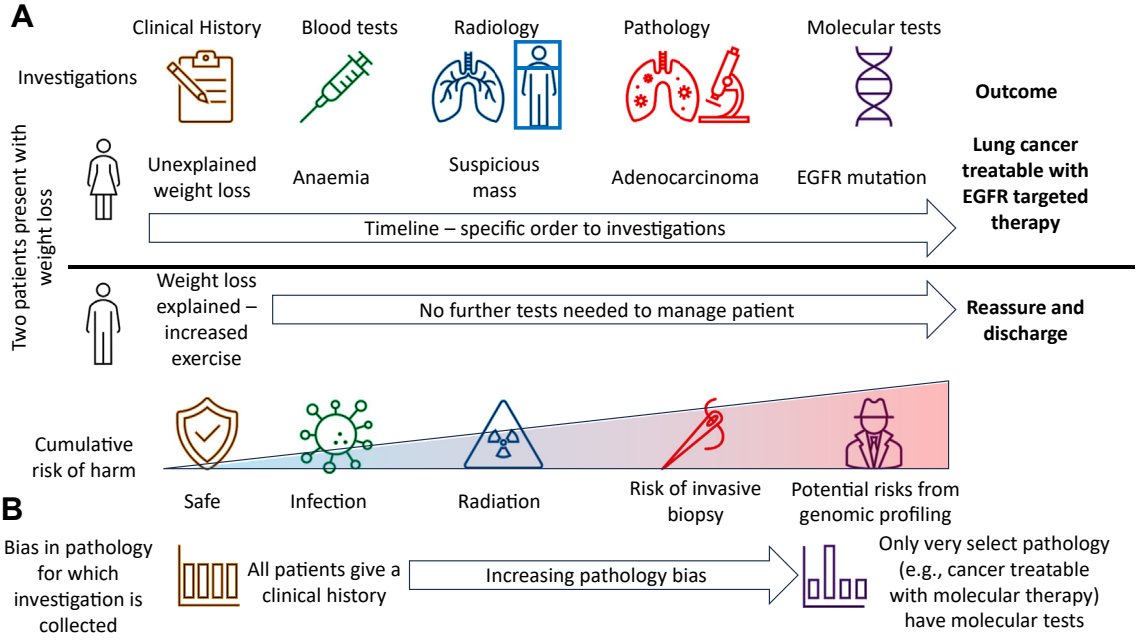

***Fig. 1:* Clinical data collection is a hierarchical, patient-personalised process.** A. Clinical data acquisition differs for two patients with the same presentation, highlighting the importance of clinical context to data acquired. B. Patients about whom significant multimodal clinical data is collected are likely to have significant underlying pathologies.

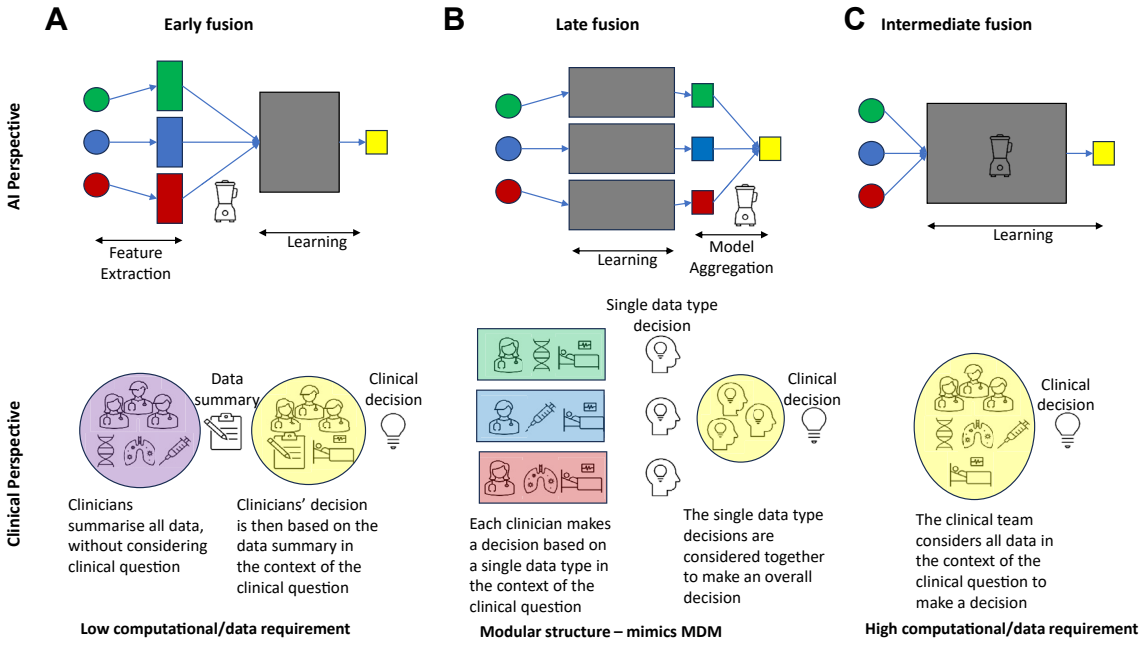

Fig. 2: **Multimodal data fusion: AI and clinical perspectives.** A comparison of AI perspectives and clinical interpretation is provided for the three most common multimodal data fusion approaches: A. Early fusion, B. Late fusion, C. Intermediate fusion.

provide a more clinically relevant way to combine data from different sources. However, typically, the choice of data ordering in these models must be defined a priori and is fixed for all individuals (i.e., decided at the population, rather than individual patient level). Thus, while they provide more flexibility than standard approaches, they do not fully capture the personalised nature of clinical decision-making. There is a need for clinician-informed ways of building multimodal AI that allow clinical data types to be fused in a personalised way that mirrors the uniqueness of each individual patient's journey.

Collectively these issues highlight the need for new approaches to clinically attuned multimodal AI. Importantly, such tools cannot be designed by AI developers alone. To be effective they will require deeply embedded clinician input throughout the full development pipeline—from the earliest stages of model planning and design, through training and testing phases, to their deployment in the clinic. Doing so, will ensure that models are selected based on their relevance to the clinical setting, rather than their general open-source availability.

Involving clinicians in multimodal AI development

AI development is a multi-step process that involves six main stages which iterate to form a cycle.[23] They are: (1) data collection, (2) data labelling, (3) data processing, (4) model selection, (5) model training and validation, and (6) testing and deployment (Fig. 3). As noted above, clinician involvement is often limited to the final testing and deployment stage. However, because clinical medicine is a dynamic, interventional process that involves collection of multiple data modalities, we believe that to be truly effective, clinicians must be involved throughout the multimodal AI development loop—not only as advisors but as codevelopers who ensure that tools are well matched to the clinical context and are addressing genuine clinical needs.

Data collection

In this first stage of AI development, the data needed to train the AI model are acquired. For clinical applications these data correspond to the outcomes of medical investigations and are acquired either directly from patients or indirectly via large databases. Regardless of how the data are acquired, they are almost always primarily obtained by a clinician. The data may be complex and/or require specific expertise to interpret, often related to knowledge of wider contextual issues. Clinicians are therefore vital to advise on the quality and factors that affect the reliability, meaning or biases inherent in the data.

This must include advice on technical issues, obtained from experience of handling or interpreting the data. For instance, knowledge that iatrogenic haemolysis during venepuncture occurs in association with specific pathologies[24] will condition the interpretation of blood

Box 2.

Clinical parallels to current multimodal AI approaches.

A key part of multimodal AI development is a step called data fusion, which defines how distinct data modalities are combined during the learning and decision-making processes. There are three major data fusion approaches: early, late, and intermediate (all with an ever-growing set of variations). These names relate to the point during the learning or decision-making process when the data modalities are first combined. Though decision making by multimodal AI tools differs from clinician decision-making, these three broad approaches can be viewed from both an AI and a clinical perspective and have different levels of clinical relevance.

Early data fusion

AI perspective

Features are extracted from each data modality and summarised in a way that makes all modalities compatible with one-another. A single AI model is then trained on the combined feature set. Because only compatible features are included in the combined feature set, important information may be discarded. This approach generally results in the simplest AI model of the three data fusion approaches, requiring the least data and computational resources (Fig. 2A).

Clinical perspective

Consider the following case study. A clinical team is provided with three types of data—blood test results, radiology images, and histopathology slides—once they have been collected, without knowledge of the order of acquisition or information about the clinical question they are expected to answer. Lacking this context, the clinical team 'extract features' from the clinical data which they hope will be relevant to the patient—e.g., the white cells count is elevated, there is a mass in the right upper lobe of the lung, there are numerous intra-epithelial neutrophils present in the biopsy. After extracting these features, the raw data is taken away from the clinical team and they are presented with the clinical question, which they must answer only with reference to their extracted clinical features.

Relevance to the clinical context

This approach may be sufficient for common, broad, clinical questions, such as 'Does the patient have an infection?', especially when the data shows gross abnormalities. However, for more specific questions, such as 'What chemotherapy should be prescribed?', the context is needed to guide extraction of relevant features from the data to answer the question, and early fusion will likely underperform. This form of data fusion is also not modular, and for patients where some data modalities are not collected (e.g., due to patient preferences or financial constraints) early fusion models cannot provide predictions, without significant adjustment (for instance, to impute the missing data). Importantly, extraction of information from clinical data without consideration of the wider clinical context, would be considered poor practice if done by a clinician.

Late data fusion

AI perspective

A separate AI model is trained for each data modality, and the outcomes of each model are then combined to give a consensus recommendation (Fig. 2B).

Clinical perspective

Consider the following case study. Three clinicians are asked a clinical question, without being allowed to confer. One is provided only blood test results, another only radiology images and the third only histopathology slides; none are informed about the order of tests. Each clinician provides a recommendation based on the limited evidence they have seen. The clinicians then reconvene to provide a collective answer to the question but are not allowed to bring the raw data to this meeting, only their overall recommendation and their certainty thereof. If one data type is not collected (e.g., the patient was not sent for biopsy) then the relevant clinician is excluded from the discussion.

Relevance to the clinical context

This approach has parallels to clinical decision making by a multidisciplinary team, a common paradigm in complex clinical settings, such as oncology. Unlike early fusion, late fusion is modular, and decisions can still be made for patients when some data are missing (e.g., due to patient preferences or risk of harm). However, in contrast to a clinical multidisciplinary meeting, where the full data are presented and discussed, in late fusion only the model predictions are discussed, and therefore the underlying data cannot be considered in their wider context. This restriction clearly has important clinical ramifications. For example, in the context of diagnosing metastatic cancer, knowledge of primary tumour site (e.g., breast, lung etc.) is vital to determining the appropriate treatment course and prognosis. In this setting, radiology imaging may show masses at multiple sites, but to minimise patient risk only one biopsy is performed from an accessible site. Without knowledge of the presence of multiple masses and their locations provided by radiology, the histopathologist will be significantly hampered in their ability to provide an accurate primary site for the tumour and the team will be compromised in their ability to treat the patient effectively.

Intermediate data fusion

AI perspective

A single model is trained using all available data simultaneously, and features are determined by a specified learning objective. This approach typically requires the most complex AI model and therefore the most training data and the most computational resource (Fig. 2C).

Clinical perspective

Consider the following case study. A clinical team is provided with three types of data—blood test results, radiology images, and histopathology slides—once all have been collected and without knowledge about the order of acquisition. The team then considers all the data in the context of the clinical question. Discussion between clinicians is permitted, to facilitate an optimal decision.

Relevance to the clinical context

By allowing the data to be examined with reference to one another in the context of the clinical question, this approach most closely mirrors the clinical decision-making process. However, like early fusion, intermediate data fusion is not modular and so cannot provide predictions when one data modality is missing, without significant adjustment (for instance to impute the missing data).

These three broad approaches to data fusion clearly have varying relevance to the clinic. However—and importantly—clinical medicine is a both a dynamic and interventional process in which the clinician continually learns about the patient's health state, for instance by requesting relevant tests in a well-considered order and responds appropriately to improve that state (for instance by prescribing treatments). None of these methods accounts for such personal and interventional dynamics, and so do not fully reflect the richness of the clinical process.

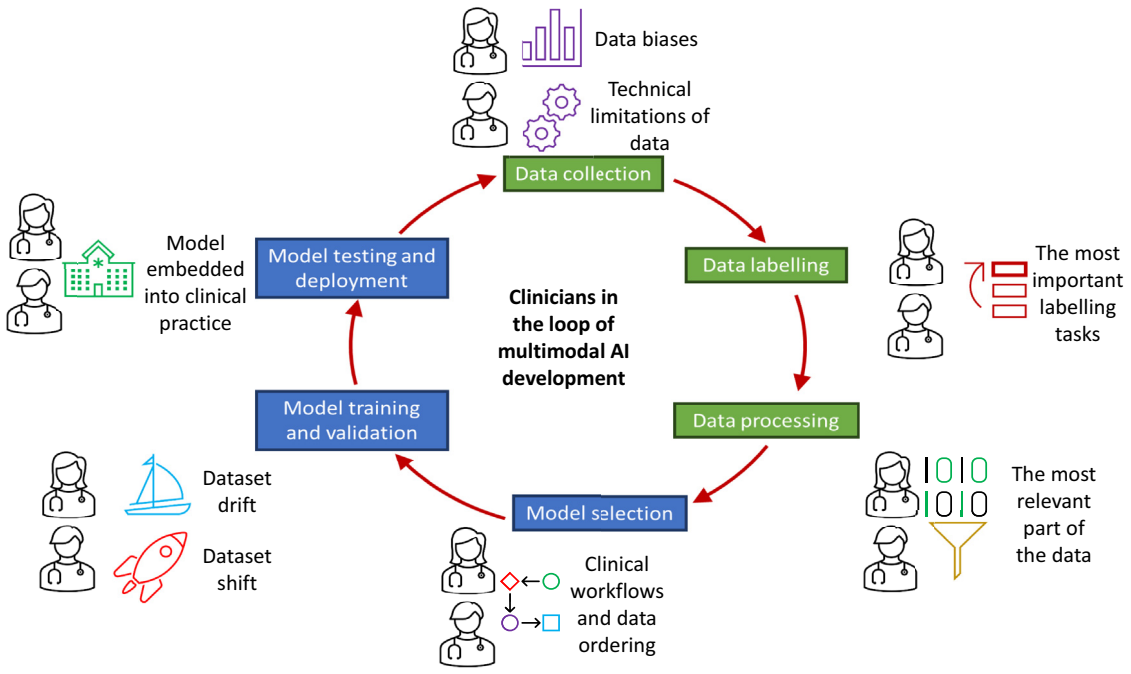

***Fig. 3:* Clinicians in the loop of multimodal AI development.** Barriers to translation of multimodal AI can arise at any point in model development and clinician involvement is required throughout this loop to ensure that the AI tool is matched to its clinical context.

test results; knowledge that biopsies processed in different pathology laboratories often have different staining intensities,[25] will similarly condition the proper interpretation of histopathology images. Clinicians are also more aware of a wide variety of demographic factors that impact data collection—for instance, that anteroposterior chest X-rays are more commonly performed on frail patients than posterioanterior chest X-rays, or that more ultrasounds but fewer fluoroscopies are performed on women than men by the UK National Health Service.[26] This understanding enables better interpretation of the distribution of collected data types.

Such insights into the data collection process can be highly nuanced, yet are essential to properly understand clinical data, assess issues of bias, set the scope of clinical questions that can reasonably be asked of training data, and determine the patient populations that predicted outcomes might reasonably relate to. In the case of multimodal data, the order in which investigations are performed and the time between them is also vital and is itself a form of clinical information that must be interpreted by clinicians—for instance, to provide insight into the rarity and severity of any pathologies observed.

Data labelling

Once it has been collected, each raw datapoint (i.e., the data associated with a unique patient) is typically assigned an appropriate clinical label, usually by an expert clinician. These labels allow clinical AI models to learn associations between the data features that are predictive of the label of interest (such learning is said to be supervised; there are other important forms of AI that do not rely on such labelling which we will not discuss here). Depending on the problem at hand these labels can be very diverse and are often both data and problem dependent. Examples include diagnoses associated with electronic health records,[27] fracture sites associated with radiological images,[28] and tumour regions associated with histopathology slides.[18] In the setting of multimodal AI, many labels can be associated with each datapoint to reflect the fact that several clinically important outcomes—such as treatment response, survival time, side effects, etc.—may be deduced from the combined dataset. Clearly these labels will often be related and will require clinical expertise to interpret. Clinician input is therefore essential to guide the most appropriate ways to label multimodal clinical data and specify the tasks for which AI tools can provide the most useful support. Depending on the clinical need, these tasks may accelerate time-consuming clinical processes—for instance, by triaging large volumes of data—and/or supporting complex clinical decisions.

In many AI applications, such as the classic computer vision task of discerning pictures of cats from dogs, it is reasonable assume that the training labels (here 'cat' or 'dog') assigned to a data point (here an

image) represent ground truths that will not change over time. In the complex world of the clinic, decisions are often not so clear cut. Two issues are particularly relevant. First, some clinical decisions are made on judgement, and can therefore vary by practitioner. A significant example is histopathological assessment of tumour grade, which has notoriously large inter-clinician variability, especially for intermediate grade categories.[29] This fact, and the clinical settings under which inter-clinician variability is most prevalent or problematic, must be accounted for in the design of clinical AI tools (see Model selection). Second, clinical guidelines that define standards of best practice are regularly updated in light of new knowledge—for instance as better medications, surgical procedures, and biomarkers are developed. In consequence, a clinical decision that accorded with best practice a year ago may not do so today (indeed, it has been estimated that 50% of clinical guidelines are out of date within 5.8 years[30]). It is important that AI tools keep abreast of evolving guidelines and doing so requires ongoing clinical engagement.

When providing an appropriate data label, say of a diagnosis, a clinician must often consider a large amount of temporal data and extract relevant context to guide the labelling process, applying causal reasoning often over long periods of time. For example, a diagnosis of haemophilus influenza B in an adult may be deduced through knowledge of the patient's childhood vaccination history, which could be decades old. In a multimodal setting, radiological imaging of a patient's lungs may be labelled as asbestosis via knowledge of an occupational exposure to asbestos 25 years earlier.[31] Though AI has achieved human level performance in many areas, current leading AI models have mixed performance on interpreting context in longer data inputs, such as a patient's entire electronic health record. For many large language models, even those designed to receive long context inputs, context association is found to be strong at the beginning and end of the data input, but context relevance is often 'lost in the middle' of the data input.[32] Recent advances have been made in addressing this problem, such as Google's Gemini 1.5[11] and retrieval-augmented generation models,[33] however they still significantly underperform compared to human long context gap retrieval. In a recent assessment humans were found to be over 10% more accurate in the detection of relevant context in large data inputs than state-of-the-art models.[34] The consequence of misattribution of context in a clinical setting may be dire, including widening healthcare inequalities, inappropriate public health measures or research into flawed treatments. A key role of the clinician in current AI development is thus to interpret the long context in clinical data and provide accurate, long context gap association labels. Such labelling tasks are essential if we are to close the gap between AI and human performance and develop translatable multimodal long context gap models.

Data processing

Data processing refers to the conversion of raw data into a format that can be efficiently and reliably parsed by a model. There are many facets to data processing, including standardisation (particularly if data is acquired from many different sources), removal of noisy features, dealing with missing data (for instance by imputation or excision), and feature weighting (to prioritise those aspects of the data that are most informative). All these steps rely on assumptions concerning the data itself and/or the problem at hand, and clinician involvement is vital to ensure that these assumptions are appropriate. For example, in digital pathology, image-size reduction is often used to exclude irrelevant parts of an image and enable efficient model training. AI models for characterising malignancy often do this by employing a 'white-space' filter, to remove regions of a digital H&E slide that do not contain dark-staining tumour material.[35] In this case, the white space is excluded since it is not relevant to the problem at hand. But this is not always the case. For instance, the same white-space filter can be used to detect adipocytes in biopsies of non-alcoholic fatty liver disease, since adipocytes do not take up colour stains from H&E.[36] In this case, the white space is clinically informative. Clinician involvement in data processing is needed to ensure that assumptions account for such nuances. This issue is relevant for clinical AI generally but is particularly important when considering multimodal clinical AI, since false assumptions concerning one data modality can propagate and become embedded in the larger model workflows in opaque ways.

Model selection

Once training data has been obtained, labelled, and processed, a model must be constructed to associate data with clinical outcome(s). There are a very wide variety of ways of constructing such models, from traditional statistical methods such as logistic regression, to decision trees, ensemble methods, and deep neural networks.[37] Multimodal models additionally require a so-called data fusion step in which data of different modes are combined for learning (see Box 2). In general, while there is freedom in the selection of model architecture, the choice of data fusion approach can have significant impact on the model's clinical utility and the extent to which it is trusted by clinicians. Because no current data fusion methodologies adequately mirror the dynamic personalised nature of patient journeys, it is perhaps here that clinician involvement in the multimodal AI development pipeline is most needed. To adequately mirror clinical workflows, data fusion must be highly flexible, able to accommodate missing data modalities and identify which added modalities will

be most informative to guide decision making, given clinical constraints and in a patient specific manner. Development of new tools that can dynamically update model predictions in light of new evidence, taking account of clinical constraints, will require very deep engagement between data scientists and clinicians. Indeed, because the challenges are both technical and clinical, addressing them will, we believe, require proactive community building and training of a new generation of clinician-scientists that have both computational and clinical expertise.

Model training and validation

Model training and validation refers to the process by which labelled data is used to train the chosen model (i.e., tune its parameters to suit the problem of interest) and test its robustness and/or generalisability. As described above, these processes can be confounded by evolving clinical definitions and best practices. Similarly, the training data itself may also evolve, a process known as dataset drift or shift. Such data changes can occur gradually—for example, due to public health strategies aimed at modifying behaviours, like smoking, diet, and exercise[38,39]—in which case they are known as dataset drift; or can occur suddenly—for example, after widespread challenges, as during the COVID-19 pandemic, or among refugees experiencing traumatic displacement[40,41]—in which case they are known as dataset shift. Models trained on out-of-date data are naturally less accurate, and such accuracy degradation due to dataset drift or shift is likely in the clinical setting.[42]

Detecting dataset shift/drift is classically considered a problem for data scientists, rather than clinicians and is typically achieved by collecting an additional dataset after model training, testing for distributional differences between this new data and the training data, and assessing the impact of these differences on model performance. There are a number of practical questions which must be considered in this process, including: (1) When should one collect the new dataset and pause model use (continual data collection would be highly expensive)? (2) How much data should one collect before testing for shift/drift (even large datasets may be underpowered to detect small but significant shifts)? (3) What kind of data should be used to retrain the model to account for shift/drift, can any historical data be reused? While these practical concerns appear on first glance considerations for data scientists alone, clinicians can provide important guidance to help address each, allowing detection and correction for dataset shift/drift efficiently and safely.

For example, consider a model used to guide clinical decisions in practice, under emerging legislation, such as the EU AI Act, model performance must be monitored, allowing the overseeing clinician to understand the level of confidence they should have in the model prediction[43] (Testing and deployment). These metrics (unlike full datasets) are cheap to collect and store, and can be autonomously assessed for real time shifts in model performance. Once a shift in model performance over time is detected (by overseeing data scientists), this can be discussed with clinicians to determine its clinical relevance. It may be that a small drop in model performance is acceptable for AI assisting with very low risk clinical tasks, and the cost of taking the model offline to adjust for shift/drift would be more detrimental to the patient population than continuing its use. For higher risk clinical tasks, however, this may not be the case and the clinician-data scientist team may decide that it is appropriate to pause model use, and collect data to assess shift/drift, addressing the above concern (1).

Deciding on the amount of data one must collect to detect shift/drift (concern (2)) again will benefit from clinician input. This requires a 'power calculation', which data scientists employ to estimate the amount of data needed to detect shift/drift. This calculation requires knowledge of the expected size of the shift/drift and an acceptable false negative rate, which the data scientist must choose. Generally speaking the smaller the size of the shift/drift and the smaller the false negative rate, the more data is required for detection. However, it is clinicians who have the strongest insight into both of these free parameters. For example, a pathologist might notice a roughly 10% increase of tumour infiltrating lymphocytes in biopsies treated with a given immunotherapy–providing an estimate for drift size. The acceptable false negative rate for shift/drift detection will also be determined by clinical concerns. For example, when detecting a drift in the number of metastatic deposits for a given tumour type, clinicians will likely request a smaller acceptable false negative rate, than detecting a drift in vitamin D levels in healthy individuals.

Once drift/shift is detected, clinicians are again required, to advise as to the reason for the dataset drift/shift, using their domain expertise. Drift/shift may be due to new public health measures, new treatments, updated clinical guidelines, new pathogens etc., or the reason may be less clear. Regardless, the clinicians' insight can guide collection of the most appropriate dataset for efficient retraining (concern (3) above). For example, if the shift is due to a new treatment, data collection should be biased towards patients receiving this treatment; if the shift is due to guideline changes, historical data may be re-used for model retraining following modification of outcome labels to reflect the latest guidelines.

Because data changes can occur in many ways, dataset shift/drift is likely to be exacerbated when learning from multimodal data, and model re-training will likely be necessary if multimodal clinical AI is ever to have stable clinical utility. Yet, retraining is (at least currently) a highly expensive process, requiring not

only costly computational resources, but also updated data collection and labelling. In order to address this problem in the safest and most efficient manner, clinicians and data scientists must work closely together.

Testing and deployment

Once a model has been trained, it must be deployed into clinical practice. A clinician's working day is typically a busy one, and seamless introduction of a new technology is challenging. Previous efforts to embed AI into clinical practice have often failed. Interestingly, lack of consideration for the clinical user is often a more important factor than model performance.[44,45] Clinician involvement is thus essential in the earliest planning stages of clinical AI deployment, both to ensure that tools are easy to use and meeting genuine clinical needs, as well as to educate users of the benefits. Importantly, it has been shown that clinician seniority is negatively related to the number of diagnostic investigations performed (i.e., senior clinicians tend to order fewer investigations) yet positively associated with improved outcomes,[46] suggesting that clinicians with different levels of experience may engage with multimodal AI in different ways. This highlights the importance of consulting a wide range of clinicians in the deployment of multimodal AI.

As regulation over the use of high risk AI tools emerges, such as the EU AI Act,[43] it is apparent that clinical AI will require human oversight measures built in to their design. These measures are required to provide the overseeing clinician an understanding of the confidence they can place in AI model predictions, as well as an insight into the rationale behind AI recommendations. Developing these measures is currently the domain of data scientists, using tools from the fields of personalised uncertainty quantification[47] and explainable AI[48] respectively. However, these fields are vast and there are many approaches, most of which have been developed for AI applied to non-clinical domains. Clinicians, as the overseers of clinical AI and the users of these measures, must be involved in their development, to ensure they are fit for purpose. Clinician-data scientist teams have recently identified conformal prediction[49] as a personalised uncertainty quantification tool and concept level explainable AI[50] as oversight measures well suited to clinical AI models.

A two-way collaboration between a diverse team of clinical end users and developers is thus needed to ensure that multimodal AI is useful to the full clinical team, who know to expect improved outcomes and/or increased efficiency and can feedback to the developers to facilitate model improvements and ongoing integration of AI into clinical practice.[45]

Building multidisciplinary communities

We have argued that the production of translatable, multimodal clinical AI requires clinician involvement throughout development. However, such engagement is not yet standard practice.[13]

At present, the most common approach to clinical AI development separates tool building (the first 5 steps in Fig. 3) in which clinicians are rarely involved in an integrated sense, from tool deployment (step 6) in which clinician involvement often occurs by default, yet may be too late for impact. This not only fails to leverage the domain knowledge of clinicians in AI development,[45] but may foster mistrust of AI in the clinical community.[51] Currently, most clinician involvement in AI also relies on interview, questionnaire, and survey feedback.[13] Though these techniques provide an easy means for developers to collect information, a dynamic dialogue between the clinician and developer is rarely achieved, preventing the clinician from being an active stakeholder.

Facilitating the clinician as an active participant in multimodal AI development is a problem without a simple solution, and will require a range of long-term approaches. These include strategies to encourage clinicians into AI research, such as academic funding bodies stipulating an expectation of clinician involvement throughout AI development. Complementarily, AI developers should be incorporated as observing members of the clinical multidisciplinary team to deepen their understanding of clinical workflows. This strategy, in particular, is a recognised but as yet unrealised goal of Health Education England.[52] In addition, wider education of healthcare professionals about AI, via large scale targeted programs such as PathLAKE for computational pathology,[53] can reduce mistrust and direct clinicians towards active involvement in AI development, encouraging them to help fix the issues with AI that they see.

Whilst these strategies are important, it remains a fact that the AI developer community and the clinical community are culturally distinct. At present, the two groups have different philosophies, speak different languages, and have different goals. Fostering integration of these distinct groups cannot therefore be a task-specific process, focused on the technical challenge of model development or the practical challenge of model deployment. Instead, formal structures are needed to develop an integrated community of clinicians and AI developers, with common philosophy and shared challenges. The development and fostering of these groups cannot be a passive process and will require application of stakeholder management techniques, to ensure that clinicians are fully integrated throughout multimodal AI development.[54] The role of Research Community Managers (RCM) is growing in popularity and recognition and they are well placed to enact this. RCMs broad set of skills encompass communication, engagement, strategic planning, and technical expertise through activities such as participation guidelines for projects, stakeholder mapping, organising knowledge sharing events and

workshops, and maintaining and supporting technical documentation.[55] RCMs can therefore play a leading role in fostering collaboration, transparency, and community-based approaches in interdisciplinary projects between data scientists and clinicians, ensuring these projects and groups can function as a strongly connected ecosystem and achieve improved outcomes and innovation in a healthcare environment.

One such example of an initiative that has community management coordination is The Alan Turing Institute's Clinical AI Interest Group.[56] The 3 main aims of the group are:

- Sharing: providing an incubator space for clinicians and data scientists to share expertise in the design and methodology of clinical AI.
- Educating: the current focus of this aspect is around developing a curriculum and core training which educates the health sector in Clinical AI to a baseline level.
- Pooling: drawing together the community. Most recently this has been through supra-interest groups that bring together clinicians and AI experts in a focused manner in different clinical specialities.

Other examples of multidisciplinary programmes that unite clinicians and data scientists are the DEMON Network and the InSilico UK network. DEMON focuses on dementia research and provides a range of training, networking opportunities, seminars, and workshops, as well as coordinated engagement with industry for real-world impact. The InSilico UK network was co-created with the Royal Academy of Engineering, the British Standards Institute, the Association of British HealthTech Industries, the Association of the British Pharmaceutical Industry, techUK and Avicenna Alliance. This network also has a regulatory focus and brings together a community to deliver an ecosystem for medical products. Fostering and actively managing these interdisciplinary communities is essential to ensure integrated clinician-data scientist teams guiding the development of translatable multimodal clinical AI.

Conclusion

The current pace of AI development, across a range of diverse applications, has been exceptional. Recent advances in multimodal AI are allowing us to harness information from disparate sources in ways which were not previously possible, such as developing 'co-pilots' for pathology[57] and radiology[58] reporting, as well as multimodal diagnostic tools.[59] In the clinic, where data comes in many varieties, multimodal AI has the potential to provide profound advancements in our understanding of disease and best practice for patient care. Most multimodal AI tools, however, are not designed with the clinical setting in mind, and the high cost of developing this technology means that repurposing of models is commonplace.

Yet, clinical data is acquired in both a temporal and highly individual way that reflects each patient's personal clinical journey. The clinical setting accordingly provides distinct challenges to multimodal AI developers and there is a need for new approaches to multimodal AI that are attuned to these challenges. The involvement of clinicians throughout the multimodal AI development pipeline is therefore essential, but this endeavour is no trivial undertaking. Long term strategies, including active community management, are needed to unify the culturally distinct communities of AI developers and front-line clinicians, enabling us to work together towards the common goal of improving the health of all.

Contributors

CRSB, BD, ABS, CH, VH, TC, BDM wrote and edited the manuscript. CRSB made the figures. CRSB and BDM accessed and verified the data.

Declaration of interests

CRSB, BDM, TC, and VH are supported by the Turing-Roche Strategic partnership. CRSB is additionally supported by Cancer Research UK. CH is an employee of Roche.

Acknowledgements

CRSB, TC, VH, CH, and BDM were supported by the Turing-Roche Strategic Partnership. CRSB was also supported by the CRUK City of London Centre Award [CTRQQR-2021\100004].

References

1 Pichai S, Hassabis D. Introducing Gemini: Google's most capable AI model yet. https://blog.google/technology/ai/google-gemini-ai/#sundar-note; 2023.
2 Summaira J, Muhammad Shoib A, Bourahla O, Songyuan L, Abdul J. Recent advances and trends in multimodal deep learning: a review. https://arxiv.org/abs/2105.11087v1; 2021.
3 Pang L, Zhu S, Ngo CW. Deep multimodal learning for affective analysis and retrieval. *IEEE Trans Multimed*. 2015;17:2008–2020.
4 Zadeh A, Chen M, Poria S, Cambria E, Morency LP. Tensor fusion network for multimodal sentiment analysis. http://arxiv.org/abs/1707.07250; 2017.
5 TCGA Research Network. The cancer genome atlas. https://www.cancer.gov/tcga; 2024.
6 Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779.
7 Sosinsky A, Ambrose J, Cross W, et al. Insights for precision oncology from the integration of genomic and clinical data of 13, 880 tumors from the 100,000 Genomes Cancer Programme. *Nat Med*. 2024;30(1):279–289.
8 Lipkova J, Chen RJ, Chen B, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*. 2022;40(10):1095–1110. https://doi.org/10.1016/j.ccell.2022.09.012.
9 Steyaert S, Pizurica M, Nagaraj D, et al. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nat Mach Intell*. 2023;5(4):351–362.
10 Team G, Anil R, Borgeaud S, et al. Gemini: a family of highly capable multimodal models. https://arxiv.org/abs/2312.11805v1; 2023.
11 Reid M, Savinov N, Teplyashin D, et al. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. https://arxiv.org/abs/2403.05530v1; 2024.
12 McKinzie B, Gan Z, Fauconnier JP, et al. MM1: methods, analysis & insights from multimodal LLM pre-training. https://arxiv.org/abs/2403.09611v3; 2024.
13 Tulk Jesso S, Kelliher A, Sanghavi H, Martin T, Henrickson Parker S. Inclusion of clinicians in the development and evaluation of clinical artificial intelligence tools: a systematic literature review. *Front Psychol*. 2022;13:830345.

14 Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med*. 2021;4(4):31.
15 Bergquist M, Rolandsson B, Gryska E, et al. Trust and stakeholder perspectives on the implementation of AI tools in clinical radiology. *Eur Radiol*. 2024;34(1):338–347.
16 Umer A, Mattila J, Liedes H, et al. A decision support system for diagnostics and treatment planning in traumatic brain injury. *IEEE J Biomed Health Inform*. 2019;23(3):1261–1268.
17 Arias-Londono JD, Gomez-Garcia JA, Moro-Velazquez L, Godino-Llorente JI. Artificial intelligence applied to chest X-ray images for the automatic detection of COVID-19. A thoughtful evaluation approach. *IEEE Access*. 2020;8:226811.
18 Olsson H, Kartasalo K, Mulliqi N, et al. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. *Nat Commun*. 2022;13(13):7761.
19 Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172–180.
20 Kirillov A, Mintun E, Ravi N, et al. Segment anything. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2023:3992–4003. https://doi.org/10.1109/ICCV51070.2023.00371.
21 Lu MY, Chen B, Williamson DF, et al. A foundational multimodal vision language AI assistant for human pathology. https://arxiv.org/abs/2312.07814v1; 2023.
22 Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. *Brief Bioinform*. 2022;23(2):bbab569.
23 Ng MY, Kapur S, Blizinsky KD, Hernandez-Boussard T. The AI life cycle: a holistic approach to creating ethical AI for health decisions. *Nat Med*. 2022;28(11):2247–2249.
24 Dugan L, Leech L, Speroni KG, Corriher J. Factors affecting hemolysis rates in blood samples drawn from newly placed IV sites in the emergency department. *J Emerg Nurs*. 2005;31(4):338–345.
25 Howard FM, Dolezal J, Kochanny S, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nat Commun*. 2021;12(1):4423.
26 NHS England and NHS Improvement. Diagnostic imaging dataset. https://www.longtermplan.nhs.uk/online-version/chapter-3-further-progress-on-care-quality-and-outcomes; 2020.
27 Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med*. 2022;5(1):194.
28 Kuo RYL, Harrison C, Curran TA, et al. Artificial intelligence in fracture detection: a systematic review and meta-analysis. *Radiology*. 2022;304(1):50–62.
29 van Dooijeweert C, van Diest PJ, Willems SM, et al. Significant inter- and intra-laboratory variation in grading of invasive breast cancer: a nationwide study of 33,043 patients in the Netherlands. *Int J Cancer*. 2020;146(3):769–780.
30 Clark E, Donovan EF, Schoettker P. From outdated to updated, keeping clinical guidelines valid. *Int J Qual Health Care*. 2006;18(3):165–166.
31 Frost G. The latency period of mesothelioma among a cohort of British asbestos workers (1978–2005). *Br J Cancer*. 2013;109(7):1965–1973.
32 Liu NF, Lin K, Hewitt J, et al. Lost in the middle: how language models use long contexts. *Trans Assoc Comput Linguist*. 2024;12(5):157–173.
33 Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst*. 2020;33:9459–9474.
34 Laban P, Fabbri AR, Xiong C, Wu CS, Salesforce AI Research. Summary of a haystack: a challenge to long-context LLMs and RAG systems. https://arxiv.org/abs/2407.01370; 2024.
35 Faghani S, Codipilly DC, Moassefi M, Iyer PG, Erickson BJ. Optimizing storage and computational efficiency: an efficient algorithm for whole slide image size reduction. *Mayo Clin Proc Digit Health*. 2023;1(3):419–424.
36 Vanderbeck S, Bockhorst J, Komorowski R, Kleiner DE, Gawrieh S. Automatic classification of white regions in liver biopsies by supervised machine learning. *Hum Pathol*. 2014;45(4):785–792.
37 Sarker IH. Machine learning: algorithms, real-world applications and research directions. *SN Comput Sci*. 2021;2(3):160. https://doi.org/10.1007/s42979-021-00592-x.
38 Duckworth C, Chmiel FP, Burns DK, et al. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. *Sci Rep*. 2021;11(1):23017.
39 Kore A, Abbasi Bavil E, Subasri V, et al. Empirical data drift detection experiments on real-world medical imaging data. *Nat Commun*. 2024;15(1):1887.
40 Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med*. 2021;385(3):283–286.
41 Pandey A, Wells CR, Stadnytskyi V, et al. Disease burden among Ukrainians forcibly displaced by the 2022 Russian invasion. *Proc Natl Acad Sci U S A*. 2023;120(8):e2215424120.
42 Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*. 2020;21(2):345–352.
43 EU Member States. The EU AI act. https://artificialintelligenceact.eu/; 2024.
44 Yang Q, Steinfeld A, Zimmerman J. Unremarkable AI: fiting intelligent decision support into critical, clinical decision-making processes. In: *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery; 2019. https://doi.org/10.1145/3290605.3300468.
45 Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR Med Inform*. 2018;6(2):e24. https://doi.org/10.2196/medinform.8912.
46 Li CJ, Syue YJ, Tsai TC, Wu KH, Lee CH, Lin YR. The impact of emergency physician seniority on clinical efficiency, emergency department resource use, patient outcomes, and disposition accuracy. *Medicine (Baltimore)*. 2016;95(6):e2706.
47 Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. *Nat Mach Intell*. 2019;1:20–23. https://doi.org/10.1038/s42256-018-0004-1.
48 Saranya A, Subhashini R. A systematic review of Explainable Artificial Intelligence models and applications: recent developments and future trends. *Decis Anal J*. 2023;7:100230.
49 Banerji CRS, Chakraborti T, Harbron C, Macarthur BD. Clinical AI tools must convey predictive uncertainty for each individual patient. *Nat Med*. 2023;29(12):2996–2998.
50 Chanda T, Hauser K, Hobelsberger S, et al. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nat Commun*. 2024;15(1):524.
51 Jacobs M, He J, Pradier MF. Designing ai for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In: *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery; 2021. https://doi.org/10.1145/3411764.3445385.
52 NHS England. NHS England: digital transformation, Chapter 4: workforce transformation. https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning/developing-healthcare-workers-confidence-in-ai/chapter-4-workforce-transformation/ai-multi-disciplinary-teams-mdts; 2023.
53 PathLAKE Consortium. PathLAKE: pathology image data lake for analytics knowledge & education. pathlake.org; 2023.
54 Roloff J. Learning from multi-stakeholder networks: issue-focussed stakeholder management. *J Bus Ethics*. 2008;82:233–250.
55 Sharan M, Karoune E, Hellon V, et al. Professionalising community management roles in interdisciplinary research projects. http://arxiv.org/abs/2409.00108; 2024.
56 The Alan Turing Institute. The Alan Turing Institute clinical AI group. https://www.turing.ac.uk/research/interest-groups/clinical-ai; 2024.
57 Lu MY, Chen B, Williamson DFK, et al. A multimodal generative AI copilot for human pathology. *Nature*. 2024;634(8033):466–473.
58 Zhou HY, Chen X, Zhang Y, Luo R, Wang L, Yu Y. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nat Mach Intell*. 2022;4(1):32–40.
59 Kaczmarczyk R, Wilhelm TI, Martin R, Roos J. Evaluating multimodal AI in medical diagnostics. *NPJ Digit Med*. 2024;7(1):205.