PLOS Digital Health

REVIEW

Clinical artificial intelligence applications of vision-language foundation models

Arun James Thirunavukarasu[1,2*], **Siyou Li**[3], **Pengyao Qin**[4], **Dong Nie**[5], **Rohan Sanghera**[1,6], **Ernest Lim**[7,8,9], **Juntao Yu**[3], **Le Zhang**[4]

1 Department of Clinical Neurosciences, Medical Sciences Division, University of Oxford, Oxford, United Kingdom, **2** International Centre for Eye Health, London School of Hygiene and Tropical Medicine, London, United Kingdom, **3** School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom, **4** School of Engineering, College of Engineering and Physical Sciences, University of Birmingham, Birmingham, United Kingdom, **5** Meta AI, Meta Platforms Inc., Menlo Park, California, United States of America, **6** Heidi Health, Melbourne, Australia, **7** Institute for Safe Autonomy, University of York, York, United Kingdom, **8** Ufonia Ltd., Oxford, United Kingdom, **9** Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom

* ajt205@cantab.ac.uk

Citation: Thirunavukarasu AJ, Li S, Qin P, Nie D, Sanghera R, Lim E, et al. (2026) Clinical artificial intelligence applications of vision-language foundation models. PLOS Digit Health 5(6): e0001453. https://doi.org/10.1371/journal.pdig.0001453

Editor: Nadav Rappoport, Ben-Gurion University of the Negev, ISRAEL

Published: June 11, 2026

Funding: This project was supported by the National Institute for Health and Care Research (ACF-2025-20-001 to AJT). The funders had no role in the conceptualisation, development, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: AJT reports

Abstract

Vision-language models (VLMs) represent a transformative advance in generative artificial intelligence (AI), using multimodal data processing to enhance clinical decision-making and workflow efficiency. Built on transformer architectures, VLMs excel in tasks like image interpretation, report generation, and visual question-answering, with emerging applications in radiology, pathology, and broader clinical practice. Their potential extends to automating documentation, improving medical education, and assisting with clinical decision-making in real-time. However, successful integration requires rigorous validation to address challenges such as bias, interpretability, and safety concerns. Prospective clinical trials, health economic evaluations, and stakeholder engagement are essential to ensure equitable and effective deployment. Regulatory frameworks must evolve to accommodate VLM functionality while maintaining accountability and protecting patient safety. By balancing innovation with robust oversight, VLMs hold promise in reducing clinician workload, expanding access to expert care, and advancing precision medicine—ushering in a new era of AI-augmented healthcare.

Introduction

The advent of generative artificial intelligence (AI) with remarkable abilities to respond appropriately and flexibly to diverse queries from human users has generated interest in potential clinical applications [1]. Interest was initially piqued by large language models (LLMs), as exhibited by uptake of chatbot applications such as ChatGPT, Google Gemini, and Claude [2]. Further development has generated models that can process multimodal data, including images, expanding the functionality and potential

research funding from HealthSense to support research related to medical applications of large language models, and grant funding from Théa Pharmaceuticals for medical equipment. EL is an employee of Ufonia Ltd., which uses artificial intelligence to automate clinical conversations. DN is an employee of Meta AI, who develop large language models and vision-language models. RS is an employee of Heidi Health, who develop clinical scribing tools using generative artificial intelligence.

applications of these systems [3]. AI models that can process and produce images and text are known as vision-language models (VLMs) and include the latest generations of multimodal chatbot applications.

Many terms are used to describe generative AI models, which can cause ambiguity and confusion. 'Foundation model' is a broad umbrella term capturing AI that is pre-trained on large volumes of unlabelled data before fine-tuning on labelled data (*e.g.,* conversational input and output text, or annotated images). Foundation models exclusively trained on text data are known as LLMs, and represent the earliest widely successful application of this pre-training and fine-tuning schema [2]. As these schemata have begun to incorporate multimodal data (*e.g.,* text, images, tables, sound, and video), reference has been made to 'multimodal LLMs', but this is an oxymoron [4]. For foundation models that process language and images, VLM is a preferable term [5].

Clinical practice is highly multimodal, involving information gleaned from spoken conversation and physical examination, captured through laboratory and imaging investigations, and documented in free text and tabular formats. Through their inherent multimodality, VLMs offer opportunities to reduce the workload borne by clinicians and even to expand the capabilities and function of healthcare professions [3]. Extensive work evaluating the potential of text-based LLMs has already been undertaken, and multimodal applications are likely to offer broader functionality and utility in healthcare settings [6,7].

In this narrative review, the technical underpinnings of VLMs are discussed and their potential applications in healthcare are explored. The validation pathways that VLM researchers might use to gain acceptance in clinical practice are outlined, with an emphasis on robust clinical research to justify clinical interventions. Barriers to development and implementation are considered, with specific assessment of diverse stakeholders' perspectives. As VLM applications remain nascent in routine clinical practice, examples from adjacent technologies are drawn upon in the discussion. VLM developers and clinicians working together are well placed to engineer interventions that can improve the provision of healthcare worldwide; this review provides an informative overview that can help guide the work of interested researchers.

Technical overview of vision-language models

VLMs leverage 'transformer architectures' to integrate textual and audiovisual data to enable tasks like conversational interaction, image interpretation or reporting, visual question-answering, and text-to-image generation [8]. While longer standing architectures can be ensembled in applications that interpret and produce multimodal information—such as image classifiers, generative adversarial networks, and determinative chatbots—these rules-based models do not perform 'vision-language modelling'. Vision-language modelling entails combining text and image data into a unified representation, which permits automation of more advanced tasks than with applications comprised of discrete modules that process different modalities.

Transformer architectures and training

VLM training aims to confer image and text processing abilities with an appropriate association between data formats. A wide variety of training tasks are used—corresponding to the architectural paradigms in Fig 1—which challenge the VLM to appropriately process matched multimodal information, such as words from passages of free text and pixels from images. As VLM performance improves on these tasks, they develop an ability to produce descriptive text in response to images, or *vice versa* [9,10]. With sufficient training, these capabilities generalise beyond the images within the pre-training dataset, leading to useful capabilities without specific training ('zero-shot performance') [11]. Emergence of zero-shot performance is ultimately dependent on scale, with up to hundreds of millions of images and enormous computational resource requirements [9]. Interrogation of these foundation models indicates that their abilities stem from genuine abilities to associate, reason, and plan—rather than mere pattern recognition or matching to previously encountered material [12].

Transformer-based architectures may be grouped into four broad categories, although these architectural and training paradigms are not mutually exclusive and many models exhibit combination and overlap of these methods [8,13]. The first category is contrastive vision-language modelling (Fig 1a). Contrastive VLMs usually consist of a text encoder and an image encoder. These encoders convert inputs (such as free text or images) into 'embeddings': representations of information within a unified (multimodal) data-space [14,15]. This facilitates combined processing of text and visual data, which permits direct comparisons of similarity between words and images. Examples within medicine include CXR-CLIP and CT-CLIP, which produce textual reports in response to chest X-ray (CXR) or chest computed tomography (CT) images [16,17]. These models are extensions of CLIP, a popular pre-trained neural network that is trained to map related

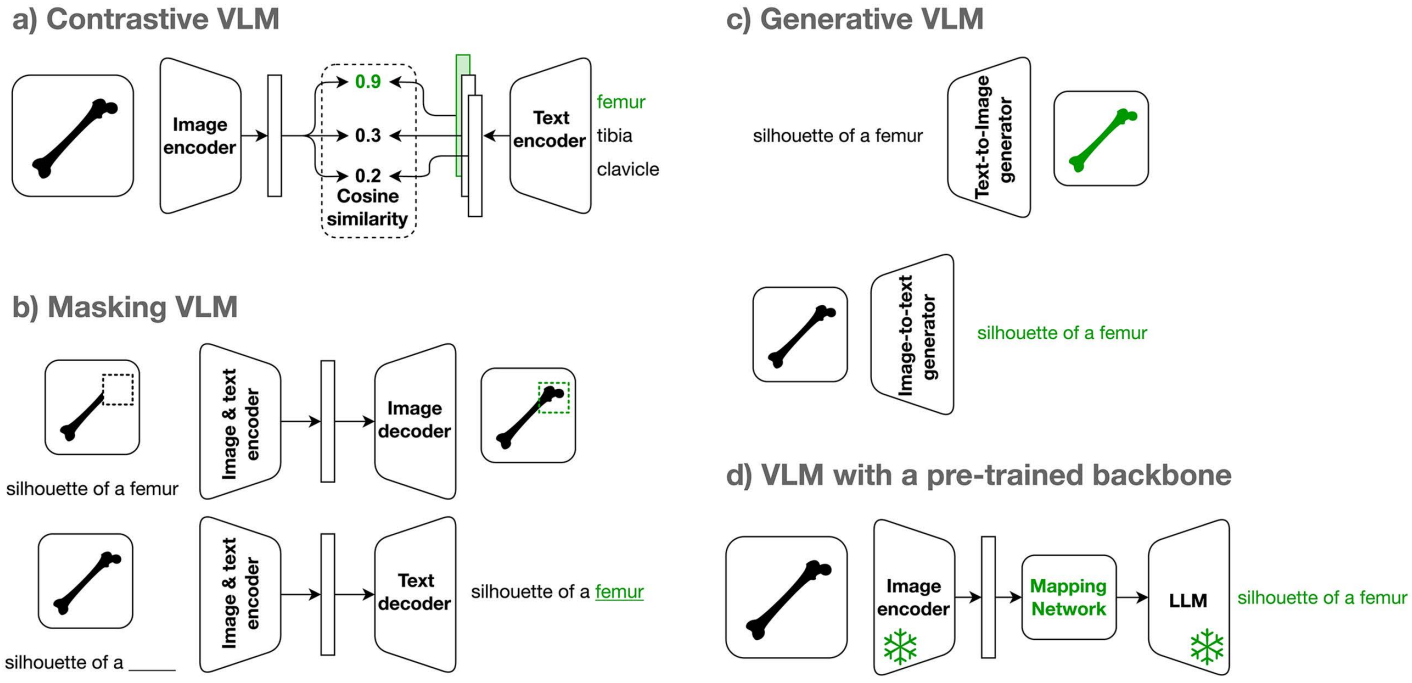

Fig 1. Four architectural paradigms of vision-language modelling. These architectures are not mutually exclusive, and aspects of each are frequently combined with one another. **(A)** Contrastive vision-language models (VLMs) convert image and text inputs into embeddings that are represented in a multimodal representation space, allowing for combination and comparison to fulfil user-defined tasks. **(B)** Masking VLMs are trained to reconstruct images or text with missing portions when provided with multimodal data as a reference. **(C)** Generative VLMs are trained to generate output whole, such as through sequential production of tokens (words) and pixels, or through iterative denoising that can be directed by textual or image prompts. **(D)** VLMs with pre-trained backbones take advantage of existing models—such as LLMs trained and fine-tuned at scales beyond most research teams—to take advantage of innate knowledge and reasoning abilities, which can be combined with multimodal input data.

https://doi.org/10.1371/journal.pdig.0001453.g001

image-text pairs closer together to facilitate downstream recognition of novel but related objects or scenes without further training [9].

The second category is VLMs trained with masking-based tasks. In this approach, models are challenged to reconstruct masked (hidden or obscured) portions of images when provided with textual captions, or *vice versa*: reconstructing obscured captions when provided an unmasked image (Fig 1b); similar to the pre-training of many LLMs [2]. Masked autoencoders typify this approach: already well established in developing applications that can reconstruct images without textual guidance, and used in a variety of visual foundation models in medicine such as RETFound [18,19]. Once trained, masked autoencoders can be fine-tuned on a wide variety of tasks that do not involve reconstructing images, such as classifying pathology on provided images [19]. Broadening training and fine-tuning to include text-based and other data formats broadens the capabilities of these model, such as by enabling generation of appropriate text in response to images [20]. FLAVA is a frontier VLM that was developed using masking strategies that required less than a fifth of the data used to train CLIP, with a curated but open source set of images and text [21].

Generative vision-language modelling is a relatively recent third category (Fig 1c). While earlier schemata depend on abstract representations produced by encoder components that map images, text, and other data onto a multimodal latent data-space, generative VLMs generate text and/or images more directly. These generative VLMs are trained to produce complete images or phrases in the form of sequences of tokens representing pixels or word fragments—frequently requiring higher computational costs for development [22]. CM3Leon is an instructive example that facilitates image-to-text and text-to-image generation, both from scratch as well as controllable editing at user-defined levels of abstraction [23,24]. Diffusion models are a widespread subset of generative VLMs, which are trained to reverse the process of decomposition into random noise; thereby producing realistic text, images, and other data from scratch [25]. Generative VLM applications include Stable Diffusion, which combines a U-Net diffusion model (for denoising) with a variational autoencoder (for image embedding) and CLIP text encoder (for text-embedding) to facilitate image synthesis in response to textual prompts [22,26,27].

Finally, VLMs can be built using a pre-trained backbone (Fig 1d). As with contrastive VLMs, an image encoder is used to enable representation of images in a multimodal space that can also accommodate textual information. Interfacing with an LLM is used to promote association between images and relevant text [28,29]. The LLM can be fine-tuned for a variety of use cases including report generation, visual question-answering, or segmenting requested organs or lesions—as exhibited by M3D-LaMed which works with three-dimensional imaging modalities such as CT [30]. Other examples include Frozen and MiniGPT, which map representations of images to pre-trained LLMs to guide text production [31,32]. These LLMs offer unparalleled text-embedding capability and contain useful knowledge and reasoning abilities that may transfer to clinical tasks [33–35]. Most research teams do not have the necessary resources to develop comparable models from scratch, making incorporation of a pre-trained LLM attractive. While this entails reliance on an external model with vulnerability to unannounced changes or discontinuation, a growing number of open source models mitigate these risks and exhibit competitive performance [28,29,36].

Improving alignment through fine-tuning

Once a VLM is trained to generate appropriate-seeming text in response to images, or *vice versa*, further fine-tuning may be undertaken to improve the usefulness of generated content. The aim of fine-tuning is to promote VLM outputs that align with users' requirements or towards a particular distribution of desired behaviours, and fine-tuning processes can be modified based on the aims of development. Many VLM applications are instruction-tuned: exposed to instructions and other inputs (*e.g.,* free text queries and images) and desired outputs (*e.g.,* image interpretation) [37]. While input/output pairs may be produced by humans, the general requirement for large datasets for fine-tuning has stimulated research into synthetic (generative AI-produced) data, with results exceeding the performance of previous VLMs [38].

Further progress has followed from pursuing similar approaches to those exemplified by chatbot applications [2]. By training an extraneous 'reward model' using human ratings of output text or images, a VLM can be autonomously fine-tuned to optimise predicted human ratings of its outputs in response to a large number of input challenges [39]. This 'reinforcement learning from human feedback' has been further augmented by using frontier LLMs or VLMs to generate ratings rather than humans, facilitating 'reinforcement learning from AI feedback' at an even greater pace [40]. Relying on a reward model that scores an abstract concept of quality is a convenient means of developing VLMs with useful responsiveness and reasoning capabilities, without defining an exhaustive list of metrics or qualities that distinguish good outputs from bad [41].

Existing and potential applications of vision-language models in medicine

Most commonly, VLMs are applied to imaging data combined with textual or tabular EPR data, although other combinations can include free text from clinicians or patients, omics data (*e.g.,* genetic sequencing), and time-series data including significant clinical events [42]. VLM applications can assist clinicians that specialise in imaging interpretation—such as radiologists and pathologists—or in a wider array of specialities where clinicians conduct, interpret, and act upon imaging as a smaller component of their responsibilities. Considering realistic and useful applications of VLMs—rather than what is currently possible with available data—is essential to direct innovation that leads to positive and impactful change [43]. However, as VLM applications are not yet widely used in healthcare, examples from adjacent technologies are used to illustrate the discussion.

Radiological and pathological applications

A wide variety of VLM applications are being developed and validated in radiology and pathology, where much of the clinical workload relates to asynchronous image interpretation [44,45]. These applications fit into a variety of models of 'computer-assisted detection' (CAD) [46]. The most obvious use cases for VLMs in radiology and pathology relate to replicating the tasks undertaken by specialists, either autonomously or by assisting supervising clinicians [46]. These functions could act to increase productivity and broaden accessibility, thereby reducing wait times and increasing the proportion of patients that receive expert care.

Successful incorporation of non-doctor imaging interpretation—such as radiographers reporting basic X-ray investigations—illustrates where generative AI could feasibly fit within care pathways in acceptable roles to clinicians and patients [47,48]. The latest forms of VLM—using transformer architectures—exhibit unprecedented performance, which could be of greater use than classical deep learning classification systems in clinical environments [49]. In retrospective studies, non-expert doctors using generative AI exhibit comparable performance to expert radiologists in assessing chest radiographs, and are also faster than unaided clinicians [50]. Moreover, VLMs tasked with reporting chest radiographs can be indistinguishable from on-site radiologists and superior to teleradiology physicians when rated by board-certified physicians [51]. Models are emerging which broaden capabilities to more data-intensive cross-sectional imaging such as CT and MRI scans: examples include CT-CLIP and M3D-LaMed, which preserve three-dimensional imaging and generate appropriate textual reports that detail and localise abnormalities [17,30]. However, it is worth noting that these impressive results are generally reported from retrospective studies, often small in scale; performance may well degrade in prospective or multicentre trials [52,53].

Image-based pathology tasks lend themselves to similar approaches to automation as well demonstrated in radiology. Foundation models used 'out-of-the-box' or with relatively simple modifications exhibit some ability to interpret pathological features, particularly those corresponding to more common conditions [54]. Many domain-specific foundation models have been developed to improve accuracy and expand functionality, using large datasets of pathology images and corresponding descriptions or reports [55–57]. Tasks that VLMs can assist with range from simple classification (categorising images based on the pathology they feature) to more sophisticated report generation, segmentation or highlighting regions of

interest, and producing synthetic images and text for educational or research purposes [58]. A growing number of AI pathology applications have gained regulatory approval, although these tend to be narrow in scope—frequently limited to a single disease group or tissue type [45]. Further work is anticipated to lead to VLM applications with more general capabilities that can offer more assistance to pathologists to mitigate the burden of a growing volume of investigations [59].

Medical and surgical applications

In medicine and surgery, VLM applications relate more to synchronous patient interaction. Administration and documentation provide a potential low-risk domain for AI intervention. VLMs can automate documentation tasks through summative capabilities, whilst incorporating interpretations from imaging modalities present in the patient's EHR, consequently improving the representation of these investigations in downstream documentation and decision-making [60]. A growing number of AI scribes and other documentation tools are being used in day-to-day clinical work already [61]. While currently generally limited to audio and text processing, these applications will improve their functionality with capacity to process images. Similarly, clinical education offers another lower-risk environment, and VLMs have demonstrated abilities in question-answering, annotating images, and synthetic data generation which can support development and learning of healthcare professionals [62]. Through explainable visual grounding and simulated data, VLMs can have a substantial positive impact on training through interactive scenarios and dynamic feedback [63].

VLMs also have significant potential to contribute more actively to clinical decision-making. Early evaluations of their potential have used aptitude tests that doctors are expected to pass as part of their training, which require an ability to incorporate clinical images, medical knowledge, contextual information, and semantic clinical reasoning [60,64–66]. These applications establish a technical basis for deploying VLMs as clinical co-pilots to assist with real-time decision-making. Beyond direct question-answering, VLMs could be used in parallel with clinicians as safety nets to mitigate medical errors and provide real-time feedback to clinicians to improve their decisions [67]. Some early studies in clinical settings even suggest that AI plans and suggestions can be superior to clinicians working with the same information [68]. However, it is important to note that most early evaluations rely on unvalidated surrogates of actual clinical utility and safety, and further validation in real-world settings is necessary [69].

Rather than imitating the work of image-based specialties such as radiology and pathology, VLMs could offer new functionality in specialties where clinicians review images themselves. In resource-constrained environments, it may be unfeasible for subspecialist experts to review every fundus photograph, dermatoscopy picture, or endoscopy recording. However, VLMs trained on a set of labelled images that could be feasibly produced by these experts could provide a resource to offer expert interpretation of images to inform decision-making—much like how physicians and surgeons frequently rely upon reports from radiologists and pathologists (Fig 2B). This could improve efficiency (addressing the growing demand for healthcare) and broaden access to expert-level care [70]. Early examples of applications demonstrating potential in this domain include ophthalmology VLMs for fundus photography and optical coherence tomography (OCT), cardiology VLMs for echocardiograms and electrocardiograms (ECGs), and neurology VLMs for electroencephalogram analysis [71–76].

The proliferation of automated ECG interpretation (generally relying on more basic algorithms than vision-language foundation modelling) serves as a useful analogy; clinician error is concerningly frequent, and while overreliance on imperfect computer-generated reports can lead to mistakes, their ability to highlight nonnormal traces makes them a useful safety net for inexpert clinicians [77,78]. Moreover, the ability of VLMs to output multimodal data—such as bounding boxes or highlighted regions indicating pathological features—could improve the explainability and usefulness of image interpretation for clinicians [79].

Training and fine-tuning with data other than expert labels could empower VLMs to work as risk calculators and treatment response predictors. Unimodal LLMs have already been used to predict clinical events and future diagnoses with remarkable accuracy after training on patient record data, and performance is likely to improve with access to multimodal data including

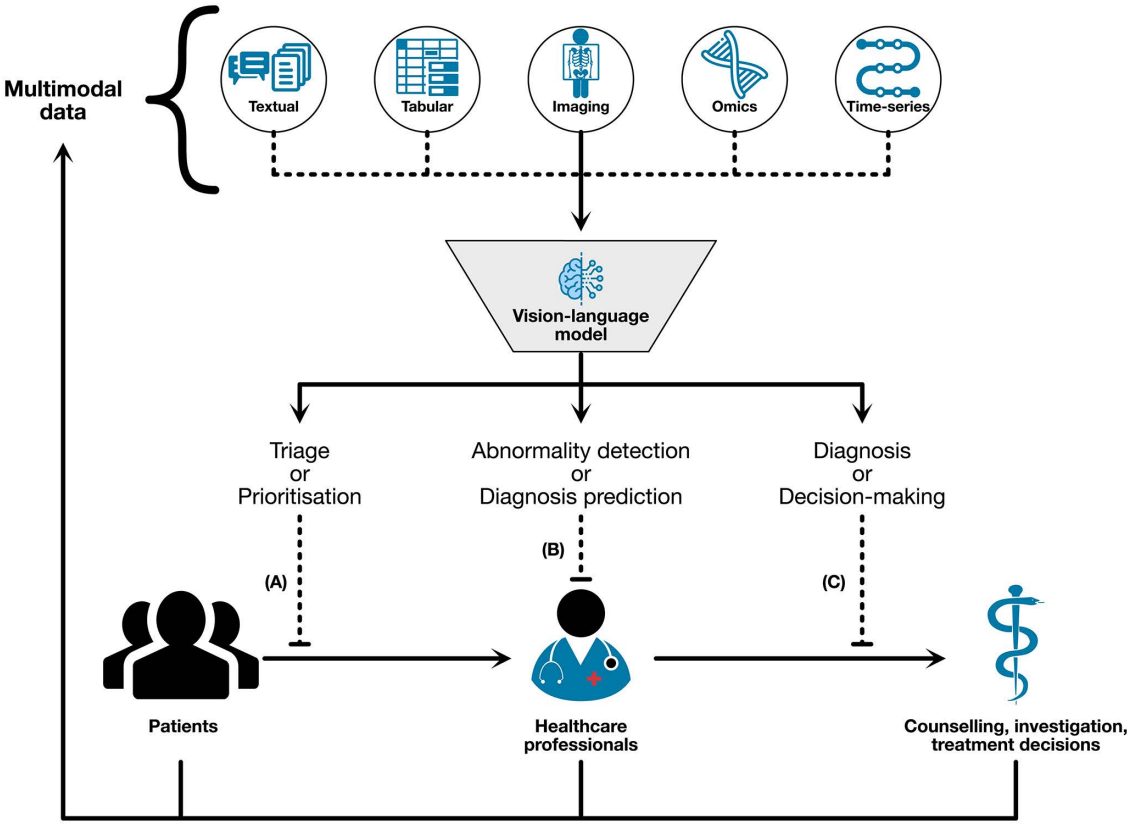

Fig 2. The potential roles of vision-language models in healthcare settings. Artificial intelligence could impact healthcare work through a wide variety of applications, but these may be conceptualised as one of three interactions with clinicians: **(A)** Direction of clinician efforts by prioritising patients based on the predicted severity or urgency of their investigation results; **(B)** Assistance of clinician efforts by augmenting investigation results with predicted diagnoses or risk scores based on recognised pathological features; or **(C)** Replacement of clinician efforts by automating aspects of the clinical workflow, such as by providing diagnoses, patient advice, or making decisions about further investigation, referral, and treatment without direct oversight.

https://doi.org/10.1371/journal.pdig.0001453.g002

imaging and genomics [80,81]. Predictive VLMs could improve healthcare decisions by identifying patients at higher risk of complications and deterioration, or by suggesting where treatments are likely to have greater or lesser effects. This could facilitate 'precision medicine'—tailoring treatment to patients' unique phenotypes, but requires careful study in prospective trials to ensure that patients are not adversely affected by bias or inaccuracy in decisions suggested by VLM output [82].

Finally, VLM applications could interface with patients rather than clinicians to provide clinical advice and reduce the demand for specialist services. A growing number of studies have evaluated the ability of AI chatbots to provide evidence-based medical advice and guidance [6]. Comparisons of an LLM application with physicians in textual real-time conversations with patient actors suggest that AI can provide superior diagnostic accuracy and conversational quality [83]. As many patient triage services rely on inexpert providers applying rigid algorithms—such as call centre professionals in the NHS 111 triage service or receptionists in primary care—there is great scope to expand access to more flexible and expert conversationalists. Dora R1, an autonomous AI chatbot, is successfully used for autonomous telephone follow-up of ophthalmology patients after surgery, and demonstrates how validated applications can reduce clinicians' workload—perhaps with great potential to impact if able to incorporate image data as well as audio and text [84]. These applications could prioritise access to appointments and direct referrals (Fig 2A), or even initiate investigation and treatment for conditions where the risks related to false positive and false negative classification are acceptable (Fig 2C).

Validation pathways for vision-language models

Preclinical validation

Reliance on similar data to that used in training and fine-tuning is insufficient due to risks of overfitting—performance optimised for a single dataset may not generalise in new contexts where idiosyncratic differences in imaging technique and population may generate subtly different distributions of features [85]. External validation—using data to which models have not been previously exposed—is therefore essential [86,87]. A growing number of datasets are publicly available to facilitate validation, but further work is needed and developers may need to gather new data specific to their proposed use case, particularly for complex multimodal interventions or applications where augmentation of data (such as bounding boxes around pathological features) is needed [88,89]. Although patient data is frequently difficult to gather due to privacy concerns, researchers have gathered imaging and text data from novel internet sources such as social media, which has been sufficient for development and robust validation exercises [55,90].

The type of application should inform the design of validation study and evaluation metrics used to quantify performance. Frequently, linguistic metrics borrowed from computer science have been used for their convenience, but these fail to capture clinical accuracy and usefulness [91]. Examination performance has also been used frequently, but this requires contextualisation and has not been shown to predict usefulness in real-world settings [33]. Many studies evaluating AI advice fail to define a reference standard or use subjective judgement without referring to evidence-based guidelines [6]. Ideally, performance evaluation should capture the success of VLM applications concisely and intuitively, and should also specifically explore the frequency and potential consequences of failure and edge-cases [92,93]. To mitigate the problems of *p*-hacking and publication bias, researchers should be encouraged to define their evaluation and analytical plan prospectively and disseminate their findings regardless of their significance (such as through a preprint server). Making models open for others to use can allow external researchers to undertake independent validation studies to verify generalisability of performance.

Clinical validation

Once an application has shown good potential with a clear conceptualisation of how it may be implemented in healthcare settings, clinical validation is essential. Close oversight may be especially necessary during validation to ensure that the risk of harm to patients is minimised. For interventions determining diagnosis, investigation, and treatment, randomised clinical trials may be required to provide objective evidence of the effectiveness and safety of the system [69]. Previous trials of AI interventions have been small in scale (often limited to single centres), relied on nonclinical or unvalidated surrogate endpoints, and lacked comprehensive demographic data–factors that make it challenging to assess the generalisability of results [52,94]. Conducting larger studies with clinical primary endpoints based on morbidity and mortality, as well as improved transparency in reporting, would provide more robust evidence supporting the deployment of GAI in clinical settings. To ensure that systems do not generate harm, which may not be captured in trials powered to detect differences in primary outcomes that are more frequent than adverse safety events, structured revalidation and ongoing monitoring for adverse effects may be organised, analogous to stage 4 clinical trials [95].

For nonclinical interventions designed to enhance clinicians' productivity or overall work experience, formal clinical trials may not always be essential [69]. However, prospective randomisation remains the most reliable approach to establishing the causal effects of new interventions, and A/B testing has long been applied outside medicine [96,97]. Reliably assaying the benefit or lack thereof of digital health interventions is essential because previous changes, such as electronic patient records have led to inefficiency, lower-quality documentation, and clinician burnout despite their intended advantages [98,99].

Beyond validation of clinical and operational benefits, health economic analysis is a key step before deployment to ensure that limited resources are directed towards interventions that confer maximal gain to patients and practitioners

[100]. This is especially important for systems that require significant resource investment due to opportunity costs; comparisons with alternative interventions can help ensure that changes maximise benefit to patients [101]. Illustrative examples from previous generations of AI include high performance diabetic retinopathy screening, which exhibited optimal cost effectiveness with a 'human-in-the-loop' rather than autonomous use in a Singaporean study [100]. Additionally, ethical concerns such as bias and fairness must be addressed, and several initiatives have been established to support clinicians, researchers, and policymakers in actively mitigating these challenges [102–104]. Through standardised, high-quality work in these areas, the field can continue progressing toward greater fairness in AI-driven healthcare solutions.

Barriers to validation and deployment

Utilitarian concerns

The primary aim of clinical interventions is to improve healthcare accessibility and effectiveness, thereby optimising patient outcomes and minimising harm. These utilitarian concerns require careful validation and monitored implementation of VLM applications, as plausible changes to healthcare can nevertheless worsen outcomes, either directly due to poor real-world performance or indirectly by representing an opportunity cost that prevents more useful interventions being pursued [105]. Reporting guidelines have been established to promote rigorous and transparent research to maximise generalisability of AI validation studies, and include CHART (for generative AI providing clinical advice), DECIDE-AI (for decision-support systems) as well as TRIPOD-LLM and TRIPOD-AI (for prediction models) [93,106–108]. Supplements to quality assessment tools to improve critical appraisal in systematic review or other evidence synthesis have also been developed, such as PROBAST-AI (for prediction models) [92].

Research has identified a particularly concerning bias in VLMs, where ethnic or gender disparities in diagnostic accuracy lead to the underdiagnosis of underserved patient populations [109–111]. Studies have revealed significant true positive rate (TPR) disparities across various protected attributes, including race, gender, and socioeconomic factors, with real-world implications that can lead to inequitable and unfair outcomes for patients [112]. It is possible that biological differences define differential performance ceilings even with optimised performance of human-delivered or AI-powered healthcare, but further work nevertheless remains necessary to address these disparities.

Beyond bias, the use of VLMs in clinical practice raises several other safety concerns, including the potential for erroneous clinical decisions. Textual output from VLMs, often generated by component LLMs, can generate plausible-sounding but factually incorrect or fabricated information ('hallucinations'), which can lead to inappropriate medical decisions [2,113]. Additionally, shortcut learning, where the AI system relies on superficial or irrelevant features in the data, can severely impact diagnostic reliability and lead to poor performance in real-world clinical settings [114]. Even with accurate and reliable VLMs, the misinterpretation of model outputs by clinicians who over-rely on predictions without understanding their limitations can result in suboptimal clinical decisions [115].

Addressing these utilitarian concerns and developing strategies to mitigate bias and safety issues are crucial for the fair and responsible integration of VLMs into healthcare.

Deontological concerns

Deontological concerns stem from risks of breaching moral duties such as transparency, accountability, and respect for human autonomy. The 'black-box' nature of frontier VLM architectures complicates verification of AI suggestions [116]. This opacity conflicts with the principle of informed clinical judgment, directed by accountable clinicians and patients' priorities and values. For instance, a VLM may diagnose a rare condition based on subtle imaging features, but without providing a decision pathway that can be traced to establish plausibility. In addition, accountability for actions taken on the basis of recommendations that may be erroneous has not been established, although the simplest default recourse is for clinicians to retain all responsibility when using tools such as VLM applications [117]. This responsibility should come with

autonomy to disregard model suggestions if they are not felt to be useful, as clinicians otherwise risk losing their ability to practice independently and having their performance inextricably linked to AI [118].

Current consent processes rarely address the role of AI in decision-making, and need to be updated [119]. Patients have a right to know if their care is informed by VLMs, and especially if their data is planned to be used for further model development and validation. Ideally, patients should be offered the chance to opt-out of AI-informed care or use of their data for training. Some stakeholders go further, advocating for explicit opt-in schema requiring patients to give informed consent for any use of their data [120,121]. Conversely, others advocate for a civic duty to provide anonymised data to advance development and produce more useful models to improve healthcare for all [122]. Opt-out schemata may represent an acceptable middle ground that preserve patient autonomy while maximising availability of data to promote innovation [123].

Stakeholders' concerns

Ultimately, AI applications must be acceptable to users and appropriately regulated with oversight that addresses deontological and utilitarian concerns. Users may be clinicians using applications for decision-support or improved efficiency, or patients using tools that improve access to prompt advice and care. Most clinicians do not currently use AI tools in their daily practice, and perspectives on AI interventions vary considerably [124]. While AI is felt to offer significant potential value in improving education and training, automating repetitive tasks, and improving patient outcomes, concerns persist regarding opaque decision-making processes as many models do not offer interpretability of how or why their outputs are generated [125]. In addition, clinical stakeholders highlight inadequate workflow integration and potential to increase rather than alleviate clinical workloads, particularly if models are encouraged to pursue a maximalist approach to investigation and treatment in the name of safety—for fear of criticism that tends to be levelled at omitting active management rather than over-investigating or overtreating [105,126]. Questions around clinical liability, potential impacts on service demand, and long-term implications for workforce planning also remain unresolved [124]. Hype surrounding LLM chatbots has led to discussions about the prospect of clinicians being replaced by AI, although this seems unlikely to happen suddenly with currently available technology [127].

For patients, concerns remain around the risk posed by decision errors and algorithmic biases, reduced transparency in explanations and inadequate rationale for AI-derived recommendations, and potential dehumanisation of the clinician-patient relationship. These challenges are especially relevant in mental health and social care contexts, where interpersonal dynamics are central to care delivery [128]. Nevertheless, patients also recognise potential benefits, acknowledging that AI may improve healthcare accessibility, and increase confidence in medical decisions through access to a 'second opinion' [124]. Generative AI applications may lead to progression from existing self-directed research by patients from search engines (the 'Dr. Google' phenomenon). While concerns have been raised about the potential of these applications to induce anxiety and reduce trust in clinicians, survey data suggest that access to information promotes confidence in the therapeutic relationship as well as patient-centred care [129,130].

At an organisational and managerial level, healthcare systems face considerable challenges in the selection and implementation of AI models. Limited interoperability between existing technical systems creates integration barriers, while the specificity of current models—often fine-tuned to specific scenarios—restricts broader application. The complexity of managing multi-agent systems, staff training requirements, and inadequate digital infrastructure further complicates successful adoption. Additionally, evolving regulatory frameworks and legislation introduce uncertainties that may slow implementation timelines.

Governance structures and accountability

VLMs present distinct governance challenges compared to single-purpose AI tools due to their variable and less rigidly defined capabilities—ranging from image interpretation and report generation to clinical decision-support [70]. This versatility makes traditional regulatory approaches, which typically classify medical software by specific functions under

a specific 'intended use', inadequate for VLMs that can dynamically shift between roles even within a single model or implementation.

Currently, the International Medical Device Regulatory Forum (IMDRF) risk classification, which underlies the FDA and EU MDR standards, uses a simple three-by-three matrix for classification of software risks depending on the significance of the information provided to a healthcare professional, and the clinical risk inherent in the clinical situation or patient (Table 1) [131]. Because VLMs can serve multiple distinct functions, clear delineation of when and how their capabilities are deployed—as well as whether functions are autonomous or semi-autonomous—is required. A comprehensive taxonomy of VLM behaviours is urgently needed to help develop governance frameworks that address their multi-functional nature, ideally fitting into existing classification systems [132].

In addition, the adaptability of certain VLMs challenges existing regulatory and assurance paradigms. Foundation models such as Google's Med-PaLM attained near-expert clinician abilities to answer questions about healthcare after being fine-tuned on just 65 question-answer pairs [35]. These models can also improve performance in given tasks after being shown examples of appropriate and successful responses ('in-context learning') [133]. This upends machine learning assurance frameworks which often place a significant emphasis on mitigating biases through assurance of training datasets, rather than runtime assurance in a real-world environment [134,135]. The complex interaction between developers designing these versatile foundation models, healthcare institutions or smaller teams deploying them across multiple use cases, and clinicians who may use different functions within the same VLM requires transparent documentation of model boundaries and limitations across its range of capabilities. Lessons may be learned by LLM applications that gain regulatory approval such as Prof Valmed, which took an iterative approach to first establish response consistency through repeated prompting, then clinical efficacy by testing against expert clinicians, and finally real-world piloting to establish safety. Work by POLARIS-GM (Partnership for Oversight, Leadership, and Accountability in Regulating Intelligent Systems: Generative Models in Medicine) is ongoing to define and mitigate the risks of these flexible systems [136]. Regulators should ideally work directly with stakeholders to ensure that innovation is not stifled by reliance on outdated and unaccommodating guidelines—while robustly protecting patients from harm—and that new frameworks have the flexibility to cope with rapidly shifting technological paradigms.

Conclusion

VLMs provide a technical basis for broadening the capabilities of medical AI to deal with multimodal information, which applies to most areas of clinical practice. Their potential utility has grown since the emergence of transformer architectures that–with sufficient training–can exhibit useful abilities outside tasks they have been exposed to previously. Future

Table 1. Classification of software as a medical device based on the International Medical Device Regulators Forum proposed framework. Classification is based on fixed definitions of the clinical situation the application is intended to be used in, as well as the effect of information generated by the device. VLM applications defy rigid definitions as they may be used in a multitude of scenarios and can have variable effects on clinician action and decision-making.

		Significance of the information provided by the software related to diagnosis, investigation, or treatment		
		High: diagnose or treat *~IMDRF 5.1.1*	Medium: influence management *~IMDRF 5.1.2*	Low: inform management *everything else*
State of clinical situation or patient condition	Critical *~IMDRF 5.2.1*	Class III *Category IV.i*	Class IIb *Category III.i*	Class IIa *Category II.i*
	Serious *~IMDRF 5.2.2*	Class IIb *Category III.ii*	Class Iia Category II.ii	Class IIa *Category II.ii*
	Nonserious *everything else*	Class IIa *Category II.iii*	Class IIa *Category I.iii*	Class IIa *Category I.i*

https://doi.org/10.1371/journal.pdig.0001453.t001

applications may range from report generation to triaging of clinicians' workload and even automated consultation and risk stratification. Validation requirements differ according to the intended use case and degree of autonomy afforded to an AI intervention, but robust study of benefits and risks is essential to ensure that resources are directed efficiently, and that patients and practitioners benefit from change. Ongoing consultation with stakeholders including clinicians, patients, and regulators is necessary to determine how VLMs should be implemented in healthcare, and to develop a permissive environment for innovation.

Author contributions

Conceptualization: Arun James Thirunavukarasu.

Project administration: Arun James Thirunavukarasu.

Supervision: Arun James Thirunavukarasu, Juntao Yu, Le Zhang.

Visualization: Arun James Thirunavukarasu.

Writing – original draft: Arun James Thirunavukarasu, Siyou Li, Pengyao Qin, Dong Nie, Rohan Sanghera, Ernest Lim, Juntao Yu, Le Zhang.

Writing – review & editing: Arun James Thirunavukarasu, Juntao Yu, Le Zhang.

References

1. Teo ZL, Thirunavukarasu AJ, Elangovan K, Cheng H, Moova P, Soetikno B, et al. Generative artificial intelligence in medicine. Nat Med. 2025;31(10):3270–82. https://doi.org/10.1038/s41591-025-03983-2 PMID: 41053447
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–40. https://doi.org/10.1038/s41591-023-02448-8 PMID: 37460753
3. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259–65. https://doi.org/10.1038/s41586-023-05881-4 PMID: 37045921
4. AlSaad R, Abd-Alrazaq A, Boughorbel S, Ahmed A, Renault MA, Damseh R. Multimodal large language models in health care: applications, challenges, and future outlook. J Med Internet Res. 2024;26:e59505. https://doi.org/10.2196/59505
5. Lin J, Yin H, Ping W, Lu Y, Molchanov P, Tao A. VILA: on pre-training for visual language models. arXiv. 2024. https://doi.org/10.48550/arXiv.2312.07533
6. Huo B, Boyle A, Marfo N, Tangamornsuksan W, Steen JP, McKechnie T, et al. Large language models for chatbot health advice studies: a systematic review. JAMA Netw Open. 2025;8(2):e2457879. https://doi.org/10.1001/jamanetworkopen.2024.57879 PMID: 39903463
7. Templin T, Perez MW, Sylvia S, Leek J, Sinnott-Armstrong N. Addressing 6 challenges in generative AI for digital health: a scoping review. PLOS Digit Health. 2024;3(5):e0000503. https://doi.org/10.1371/journal.pdig.0000503 PMID: 38781686
8. Bordes F, Pang RY, Ajay A, Li AC, Bardes A, Petryk S. An introduction to vision-language modeling. arXiv. 2024. https://doi.org/10.48550/ARXIV.2405.17247
9. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S. Learning transferable visual models from natural language supervision. arXiv. 2021. https://doi.org/10.48550/arXiv.2103.00020
10. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. arXiv. 2015. https://doi.org/10.48550/arXiv.1512.03385
11. Jia C, Yang Y, Xia Y, Chen Y-T, Parekh Z, Pham H. Scaling up visual and vision-language representation learning with noisy text supervision. arXiv. 2021. https://doi.org/10.48550/arXiv.2102.05918
12. Lindsey J, Gurnee W, Ameisen E, Chen B, Pearce A, Turner NL. On the biology of a large language model. Anthropic; 2025. Available from: https://transformer-circuits.pub/2025/attribution-graphs/biology.html
13. Bai Y, Cheng H, Zhou Y, Zhou J, Thirunavukarasu A, Ke Y. EVLF-FM: explainable vision language foundation model for medicine. arXiv. 2025. https://doi.org/10.48550/arXiv.2509.24231
14. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. 2019. https://doi.org/10.48550/arXiv.1810.04805
15. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T. An image is worth 16x16 words: transformers for image recognition at scale. arXiv. 2021. https://doi.org/10.48550/arXiv.2010.11929
16. Chen Z, Varma M, Xu J, Paschali M, Veen DV, Johnston A. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. arXiv. 2024. https://doi.org/10.48550/arXiv.2401.12208

17. Hamamci IE, Er S, Almas F, Simsek AG, Esirgun SN, Dogan I. Developing generalist foundation models from a multimodal dataset for 3D computed tomography. arXiv. 2024. https://doi.org/10.48550/arXiv.2403.17834
18. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. arXiv. 2021. https://doi.org/10.48550/arXiv.2111.06377
19. Zhou Y, Chia MA, Wagner SK, Ayhan MS, Williamson DJ, Struyven RR, et al. A foundation model for generalizable disease detection from retinal images. Nature. 2023;622(7981):156–63. https://doi.org/10.1038/s41586-023-06555-x PMID: 37704728
20. Geng X, Liu H, Lee L, Schuurmans D, Levine S, Abbeel P. Multimodal masked autoencoders learn transferable representations. arXiv. 2022. https://doi.org/10.48550/arXiv.2205.14204
21. Singh A, Hu R, Goswami V, Couairon G, Galuba W, Rohrbach M. FLAVA: a foundational language and vision alignment model. arXiv. 2022. https://doi.org/10.48550/arXiv.2112.04482
22. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. arXiv. 2022. https://doi.org/10.48550/arXiv.2112.10752
23. Yu L, Shi B, Pasunuru R, Muller B, Golovneva O, Wang T. Scaling autoregressive multi-modal models: pretraining and instruction tuning. arXiv. 2023. https://doi.org/10.48550/arXiv.2309.02591
24. Chameleon Team. Chameleon: mixed-modal early-fusion foundation models. arXiv. 2024. https://doi.org/10.48550/arXiv.2405.09818
25. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. arXiv. 2020. https://doi.org/10.48550/arXiv.2006.11239
26. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. arXiv. 2015. https://doi.org/10.48550/arXiv.1505.04597
27. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv. 2013. https://doi.org/10.48550/ARXIV.1312.6114
28. Llama Team A@ M. The Llama 3 herd of models. Available from: https://llama.meta.com/; 2024.
29. Guo D, Yang D, Zhang H, Song J, Zhang R, DeepSeek-AI. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. arXiv. 2025. https://doi.org/10.48550/ARXIV.2501.12948
30. Bai F, Du Y, Huang T, Meng MQ-H, Zhao B. M3D: advancing 3D medical image analysis with multi-modal large language models. arXiv. 2024. https://doi.org/10.48550/arXiv.2404.00578
31. Tsimpoukelli M, Menick J, Cabi S, Eslami SMA, Vinyals O, Hill F. Multimodal few-shot learning with frozen language models. arXiv. 2021. https://doi.org/10.48550/arXiv.2106.13884
32. Zhu D, Chen J, Shen X, Li X, Elhoseiny M. MiniGPT-4: enhancing vision-language understanding with advanced large language models. arXiv. 2023. https://doi.org/10.48550/arXiv.2304.10592
33. Thirunavukarasu AJ, Mahmood S, Malem A, Foster WP, Sanghera R, Hassan R, et al. Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: a head-to-head cross-sectional study. PLOS Digit Health. 2024;3(4):e0000341. https://doi.org/10.1371/journal.pdig.0000341 PMID: 38630683
34. Dou ZY, Kamath A, Gan Z, Zhang P, Wang J, Li L. Coarse-to-fine vision-language pre-training with fusion in the backbone. arXiv. 2022. https://doi.org/10.48550/arXiv.2206.07643
35. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–80. https://doi.org/10.1038/s41586-023-06291-2 PMID: 37438534
36. Yang A, Yang B, Hui B, Zheng B, Yu B, Zhou C. Qwen2 technical report. arXiv. 2024. https://doi.org/10.48550/arXiv.2407.10671
37. Li C, Ge Y, Li D, Shan Y. Vision-language instruction tuning: a review and analysis. arXiv. 2023. https://doi.org/10.48550/arXiv.2311.08172
38. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. arXiv; 2023. https://doi.org/10.48550/arXiv.2304.08485
39. Sun Z, Shen S, Cao S, Liu H, Li C, Shen Y, et al. Aligning large multimodal models with factually augmented RLHF. arXiv. 2023. https://doi.org/10.48550/arXiv.2309.14525
40. Wang Y, Sun Z, Zhang J, Xian Z, Biyik E, Held D. RL-VLM-F: reinforcement learning from vision language foundation model feedback. arXiv. 2024. https://doi.org/10.48550/arXiv.2402.03681
41. Kumar K, Ashraf T, Thawakar O, Anwer RM, Cholakkal H, Shah M. LLM post-training: a deep dive into reasoning large language models. arXiv. 2025. https://doi.org/10.48550/arXiv.2502.21321
42. Kline A, Wang H, Li Y, Dennis S, Hutch M, Xu Z, et al. Multimodal machine learning in precision health: a scoping review. NPJ Digit Med. 2022;5(1):171. https://doi.org/10.1038/s41746-022-00712-8 PMID: 36344814
43. Balagopalan A, Baldini I, Celi LA, Gichoya J, McCoy LG, Naumann T, et al. Machine learning for healthcare that matters: reorienting from technical novelty to equitable impact. PLOS Digit Health. 2024;3(4):e0000474. https://doi.org/10.1371/journal.pdig.0000474 PMID: 38620047
44. Milam ME, Koo CW. The current status and future of FDA-approved artificial intelligence tools in chest radiology in the United States. Clin Radiol. 2023;78(2):115–22. https://doi.org/10.1016/j.crad.2022.08.135 PMID: 36180271
45. Matthews GA, McGenity C, Bansal D, Treanor D. Public evidence on AI products for digital pathology. NPJ Digit Med. 2024;7(1):300. https://doi.org/10.1038/s41746-024-01294-3 PMID: 39455883

46. McNamara SL, Yi PH, Lotter W. The clinician-AI interface: intended use and explainability in FDA-cleared AI devices for medical image interpretation. NPJ Digit Med. 2024;7(1):80. https://doi.org/10.1038/s41746-024-01080-1 PMID: 38531952
47. Ekpo EU, Egbe NO, Akpan BE. Radiographers' performance in chest X-ray interpretation: the Nigerian experience. Br J Radiol. 2015;88(1051):20150023. https://doi.org/10.1259/bjr.20150023 PMID: 25966290
48. Tonks A, Jimenez Y, Gray F, Ekpo E. A stake in the game: can radiographer image interpretation improve X-ray quality? A scoping review. Radiography (Lond). 2024;30(2):641–50. https://doi.org/10.1016/j.radi.2024.01.017 PMID: 38340575
49. Khader F, Müller-Franzes G, Wang T, Han T, Tayebi Arasteh S, Haarburger C, et al. Multimodal deep learning for integrating chest radiographs and clinical parameters: a case for transformers. Radiology. 2023;309(1):e230806. https://doi.org/10.1148/radiol.230806 PMID: 37787671
50. Anderson PG, Tarder-Stoll H, Alpaslan M, Keathley N, Levin DL, Venkatesh S, et al. Deep learning improves physician accuracy in the comprehensive detection of abnormalities on chest X-rays. Sci Rep. 2024;14(1):25151. https://doi.org/10.1038/s41598-024-76608-2 PMID: 39448764
51. Huang J, Neill L, Wittbrodt M, Melnick D, Klug M, Thompson M. Generative artificial intelligence for chest radiograph interpretation in the emergency department. JAMA Netw Open. 2023;6:e2336100. https://doi.org/10.1001/jamanetworkopen.2023.36100
52. Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P. Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. Lancet Digit Health. 2024;6(5):e367–73. https://doi.org/10.1016/S2589-7500(24)00047-5 PMID: 38670745
53. Busch F, Hoffmann L, Dos Santos DP, Makowski MR, Saba L, Prucker P, et al. Large language models for structured reporting in radiology: past, present, and future. Eur Radiol. 2025;35(5):2589–602. https://doi.org/10.1007/s00330-024-11107-6 PMID: 39438330
54. Omar M, Ullanat V, Loda M, Marchionni L, Umeton R. ChatGPT for digital pathology research. Lancet Digit Health. 2024;6(8):e595–600. https://doi.org/10.1016/S2589-7500(24)00114-6 PMID: 38987117
55. Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J. A visual-language foundation model for pathology image analysis using medical Twitter. Nat Med. 2023;29(9):2307–16. https://doi.org/10.1038/s41591-023-02504-3 PMID: 37592105
56. Ahmed F, Sellergren A, Yang L, Xu S, Babenko B, Ward A. PathAlign: a vision-language model for whole slide images in histopathology. arXiv. 2024. https://doi.org/10.48550/arXiv.2406.19578
57. Lu MY, Chen B, Williamson DFK, Chen RJ, Zhao M, Chow AK, et al. A multimodal generative AI copilot for human pathology. Nature. 2024;634(8033):466–73. https://doi.org/10.1038/s41586-024-07618-3 PMID: 38866050
58. Lu MY, Chen B, Williamson DFK, Chen RJ, Liang I, Ding T, et al. A visual-language foundation model for computational pathology. Nat Med. 2024;30(3):863–74. https://doi.org/10.1038/s41591-024-02856-4 PMID: 38504017
59. Cui M, Zhang DY. Artificial intelligence and computational pathology. Lab Invest. 2021;101(4):412–22. https://doi.org/10.1038/s41374-020-00514-0 PMID: 33454724
60. Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J, et al. A generalist vision-language foundation model for diverse biomedical tasks. Nat Med. 2024;30(11):3129–41. https://doi.org/10.1038/s41591-024-03185-2 PMID: 39112796
61. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, Kipnis P. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. In: Catal Non-Issue Content. 2024;5:CAT.23.0404. https://doi.org/10.1056/CAT.23.0404
62. Ng FYC, Thirunavukarasu AJ, Cheng H, Tan TF, Gutierrez L, Lan Y, et al. Artificial intelligence education: an evidence-based medicine approach for consumers, translators, and developers. Cell Rep Med. 2023;4(10):101230. https://doi.org/10.1016/j.xcrm.2023.101230 PMID: 37852174
63. Huemann Z, Church S, Warner JD, Tran D, Tie X, McMillan AB. Vision-language modeling in pet/ct for visual grounding of positive findings. arXiv. 2025. https://doi.org/10.48550/arXiv.2502.00528
64. Hu X, Gu L, Kobayashi K, Liu L, Zhang M, Harada T, et al. Interpretable medical image visual question answering via multi-modal relationship graph learning. Med Image Anal. 2024;97:103279. https://doi.org/10.1016/j.media.2024.103279 PMID: 39079429
65. Moor M, Huang Q, Wu S, Yasunaga M, Zakka C, Dalmia Y, et al. Med-flamingo: a multimodal medical few-shot learner. arXiv. 2023. https://doi.org/10.48550/arXiv.2307.15189
66. Yang Z, Yao Z, Tasmin M, Vashisht P, Jang WS, Wang B. Performance of multimodal GPT-4V on USMLE with image: potential for imaging diagnostic support with explanations. medRxiv. 2023:2023.10.26.23297629. https://doi.org/10.1101/2023.10.26.23297629
67. McDuff D, Schaekermann M, Tu T, Palepu A, Wang A, Garrison J, et al. Towards accurate differential diagnosis with large language models. Nature. 2025;642(8067):451–7. https://doi.org/10.1038/s41586-025-08869-4 PMID: 40205049
68. Zeltzer D, Kugler Z, Hayat L, Brufman T, Ilan Ber R, Leibovich K, et al. Comparison of initial Artificial Intelligence (AI) and final physician recommendations in AI-assisted virtual urgent care visits. Ann Intern Med. 2025;178(4):498–506. https://doi.org/10.7326/ANNALS-24-03283 PMID: 40183679
69. Thirunavukarasu AJ. How can the clinical aptitude of AI assistants be assayed? J Med Internet Res. 2023;25:e51603. https://doi.org/10.2196/51603
70. Rao VM, Hla M, Moor M, Adithan S, Kwak S, Topol EJ, et al. Multimodal generative AI for medical image interpretation. Nature. 2025;639(8056):888–96. https://doi.org/10.1038/s41586-025-08675-y PMID: 40140592
71. Mishra A, Shukla S, Torres J, Gwizdka J, Roychowdhury S. Thought2Text: text generation from EEG signal using Large Language Models (LLMs). arXiv. 2024. https://doi.org/10.48550/arXiv.2410.07507

72. Christensen M, Vukadinovic M, Yuan N, Ouyang D. Vision-language foundation model for echocardiogram interpretation. Nat Med. 2024;30(5):1481–8. https://doi.org/10.1038/s41591-024-02959-y PMID: 38689062
73. Wang M, Lin T, Yu K, Lin A, Peng Y, Wang L. Common and rare fundus diseases identification using vision-language foundation model with knowledge of over 400 diseases. arXiv. 2024. https://doi.org/10.48550/arXiv.2406.09317
74. Liu R, Bai Y, Yue X, Zhang P. Teach multimodal LLMs to comprehend electrocardiographic images. arXiv. 2024. https://doi.org/10.48550/arXiv.2410.19008
75. Holland R, Taylor TRP, Holmes C, Riedl S, Mai J, Patsiamanidi M, et al. Specialist vision-language models for clinical ophthalmology. arXiv. 2024. https://doi.org/10.48550/ARXIV.2407.08410
76. Antaki F, Chopra R, Keane PA. Vision-language models for feature detection of macular diseases on optical coherence tomography. JAMA Ophthalmol. 2024;142(6):573–6. https://doi.org/10.1001/jamaophthalmol.2024.1165 PMID: 38696177
77. Martínez-Sellés M, Marina-Breysse M. Current and future use of artificial intelligence in electrocardiography. J Cardiovasc Dev Dis. 2023;10:175. https://doi.org/10.3390/jcdd10040175
78. Schläpfer J, Wellens HJ. Computer-interpreted electrocardiograms: benefits and limitations. J Am Coll Cardiol. 2017;70(9):1183–92. https://doi.org/10.1016/j.jacc.2017.07.723 PMID: 28838369
79. Tanida T, Müller P, Kaissis G, Rueckert D. Interactive and explainable region-guided radiology report generation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023. pp. 7433–7442. https://doi.org/10.1109/CVPR52729.2023.00718
80. Kraljevic Z, Yeung JA, Bean D, Teo J, Dobson RJ. Large language models for medical forecasting -- foresight 2. arXiv. 2024. https://doi.org/10.48550/arXiv.2412.10848
81. Chen YM, Hsiao TH, Lin CH, Fann YC. Unlocking precision medicine: clinical applications of integrating health records, genetics, and immunology through artificial intelligence. J Biomed Sci. 2025;32:16. https://doi.org/10.1186/s12929-024-01110-w
82. Johnson KB, Wei W, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci. 2021;14: 86–93. https://doi.org/10.1111/cts.12884
83. Tu T, Schaekermann M, Palepu A, Saab K, Freyberg J, Tanno R, et al. Towards conversational diagnostic artificial intelligence. Nature. 2025;642(8067):442–50. https://doi.org/10.1038/s41586-025-08866-7 PMID: 40205050
84. Meinert E, Milne-Ives M, Lim E, Higham A, Boege S, de Pennington N, et al. Accuracy and safety of an autonomous artificial intelligence clinical assistant conducting telemedicine follow-up assessment for cataract surgery. EClinicalMedicine. 2024;73:102692. https://doi.org/10.1016/j.eclinm.2024.102692 PMID: 39050586
85. Cabitza F, Campagner A, Soares F, García de Guadiana-Romualdo L, Challa F, Sulejmani A, et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. Comput Methods Programs Biomed. 2021;208:106288. https://doi.org/10.1016/j.cmpb.2021.106288 PMID: 34352688
86. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2020;14: 49–58. https://doi.org/10.1093/ckj/sfaa188
87. Thirunavukarasu AJ, Elangovan K, Gutierrez L, Li Y, Tan I, Keane PA. Democratizing artificial intelligence imaging analysis with automated machine learning: tutorial. J Med Internet Res. 2023;25:e49949. https://doi.org/10.2196/49949
88. Khan SM, Liu X, Nath S, Korot E, Faes L, Wagner SK, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. Lancet Digit Health. 2021;3(1):e51–66. https://doi.org/10.1016/S2589-7500(20)30240-5 PMID: 33735069
89. Bai J, Bai S, Yang S, Wang S, Tan S, Wang P, et al. Qwen-VL: a frontier large vision-language model with versatile abilities. arXiv. 2023. https://doi.org/10.48550/arXiv.2308.12966
90. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589–96. https://doi.org/10.1001/jamainternmed.2023.1838 PMID: 37115527
91. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, Thieme A, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. NPJ Digit Med. 2024;7(1):82. https://doi.org/10.1038/s41746-024-01074-z PMID: 38553625
92. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. https://doi.org/10.1136/bmj-2024-082505 PMID: 40127903
93. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378 PMID: 38626948
94. Lam TYT, Cheung MFK, Munro YL, Lim KM, Shung D, Sung JJY. Randomized controlled trials of artificial intelligence in clinical practice: systematic review. J Med Internet Res. 2022;24:e37188. https://doi.org/10.2196/37188
95. Thynne TR, Gabb GM. Limitations of randomised controlled trials as evidence of drug safety. Aust Prescr. 2023;46(2):22–3. https://doi.org/10.18773/austprescr.2023.005 PMID: 38053569

96. Quin F, Weyns D, Galster M, Silva CC. A/B testing: a systematic literature review. J Syst Softw. 2024;211:112011. https://doi.org/10.1016/j.jss.2024.112011
97. Austrian J, Mendoza F, Szerencsy A, Fenelon L, Horwitz LI, Jones S, et al. Applying A/B testing to clinical decision support: rapid randomized controlled trials. J Med Internet Res. 2021;23(4):e16651. https://doi.org/10.2196/16651 PMID: 33835035
98. Hill RG Jr, Sears LM, Melanson SW. 4000 clicks: a productivity analysis of electronic medical records in a community hospital ED. Am J Emerg Med. 2013;31(11):1591–4. https://doi.org/10.1016/j.ajem.2013.06.028 PMID: 24060331
99. Ober KP, Applegate WB. The electronic health record: are we the tools of our tools? In: The Pharos of Alpha Omega Alpha-Honor Medical Society. 2015. pp. 9–14.
100. Xie Y, Nguyen QD, Hamzah H, Lim G, Bellemo V, Gunasekeran DV, et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. Lancet Digit Health. 2020;2(5):e240–9. https://doi.org/10.1016/S2589-7500(20)30060-1 PMID: 33328056
101. Sculpher M, Claxton K, Pearson SD. Developing a value framework: the need to reflect the opportunity costs of funding decisions. Value Health. 2017;20(2):234–9. https://doi.org/10.1016/j.jval.2016.11.021 PMID: 28237201
102. Alderman JE, Palmer J, Laws E, McCradden MD, Ordish J, Ghassemi M, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. Lancet Digit Health. 2025;7(1):e64–88. https://doi.org/10.1016/S2589-7500(24)00224-3 PMID: 39701919
103. Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ. 2025;388:e081554. https://doi.org/10.1136/bmj-2024-081554
104. Ning Y, Liu X, Collins GS, Moons KGM, McCradden M, Ting DSW, et al. An ethics assessment tool for artificial intelligence implementation in healthcare: CARE-AI. Nat Med. 2024;30(11):3038–9. https://doi.org/10.1038/s41591-024-03310-1 PMID: 39424948
105. Mandrola J, Cifu A, Prasad V, Foy A. The case for being a medical conservative. Am J Med. 2019;132(8):900–1. https://doi.org/10.1016/j.amjmed.2019.02.005 PMID: 30851263
106. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904. https://doi.org/10.1136/bmj-2022-070904 PMID: 35584845
107. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60–9. https://doi.org/10.1038/s41591-024-03425-5 PMID: 39779929
108. CHART Collaborative. Reporting guidelines for chatbot health advice studies: explanation and elaboration for the Chatbot Assessment Reporting Tool (CHART). BMJ. 2025;390:e083305. https://doi.org/10.1136/bmj-2024-083305 PMID: 40750271
109. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med. 2021;27(12):2176–82. https://doi.org/10.1038/s41591-021-01595-0 PMID: 34893776
110. Koçak B, Ponsiglione A, Stanzione A, Bluethgen C, Santinha J, Ugga L, et al. Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. Diagn Interv Radiol. 2025;31(2):75–88. https://doi.org/10.4274/dir.2024.242854 PMID: 38953330
111. Yang Y, Liu Y, Liu X, Gulhane A, Mastrodicasa D, Wu W, et al. Demographic bias of expert-level vision-language foundation models in medical imaging. Sci Adv. 2025;11(13):eadq0305. https://doi.org/10.1126/sciadv.adq0305 PMID: 40138420
112. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: fairness gaps in deep chest X-ray classifiers. Pac Symp Biocomput. 2021;26:232–43. https://doi.org/10.1142/9789811232701_0022 PMID: 33691020
113. Agarwal V, Jin Y, Chandra M, Choudhury MD, Kumar S, Sastry N. MedHalu: hallucinations in responses to healthcare queries by large language models. arXiv. 2024. https://doi.org/10.48550/arXiv.2409.19492
114. Banerjee I, Bhattacharjee K, Burns JL, Trivedi H, Purkayastha S, Seyyed-Kalantari L. Shortcuts causing bias in radiology artificial intelligence: causes, evaluation, and mitigation. J Am Coll Radiol. 2023;20:842–51. https://doi.org/10.1016/j.jacr.2023.06.025
115. Behzad S, Tabatabaei SMH, Lu MY, Eibschutz LS, Gholamrezanezhad A. Pitfalls in interpretive applications of artificial intelligence in radiology. Am J Roentgenol. 2024;223:e2431493. https://doi.org/10.2214/AJR.24.31493
116. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15. https://doi.org/10.1038/s42256-019-0048-x PMID: 35603010
117. Dennstädt F, Hastings J, Putora PM, Schmerder M, Cihoric N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. NPJ Digit Med. 2025;8(1):143. https://doi.org/10.1038/s41746-025-01476-7 PMID: 40050366
118. Lawton T, Morgan P, Porter Z, Hickey S, Cunningham A, Hughes N, et al. Clinicians risk becoming "liability sinks" for artificial intelligence. Future Healthc J. 2024;11(1):100007. https://doi.org/10.1016/j.fhj.2024.100007 PMID: 38646041
119. Price WN 2nd, Cohen IG. Privacy in the age of medical big data. Nat Med. 2019;25(1):37–43. https://doi.org/10.1038/s41591-018-0272-7 PMID: 30617331
120. Ploug T. In Defence of informed consent for health record research - why arguments from 'easy rescue', 'no harm' and 'consent bias' fail. BMC Med Ethics. 2020;21:75. https://doi.org/10.1186/s12910-020-00519-w

121. Wiertz S, Boldt J. Ethical, legal, and practical concerns surrounding the implemention of new forms of consent for health data research: qualitative interview study. J Med Internet Res. 2024;26:e52180. https://doi.org/10.2196/52180 PMID: 39110970

122. Müller S. Is there a civic duty to support medical AI development by sharing electronic health records? BMC Med Ethics. 2022;23(1):134. https://doi.org/10.1186/s12910-022-00871-z PMID: 36496427

123. de Man Y, Wieland-Jorna Y, Torensma B, de Wit K, Francke AL, Oosterveld-Vlug MG, et al. Opt-in and opt-out consent procedures for the reuse of routinely recorded health data in scientific research and their consequences for consent rate and consent bias: systematic review. J Med Internet Res. 2023;25:e42131. https://doi.org/10.2196/42131 PMID: 36853745

124. Scott IA, Carter SM, Coiera E. Exploring stakeholder attitudes towards AI in clinical practice. BMJ Health Care Inform. 2021;28(1):e100450. https://doi.org/10.1136/bmjhci-2021-100450 PMID: 34887331

125. Hogg HDJ, Al-Zubaidy M, Talks J, Denniston AK, Kelly CJ, Malawana J. Stakeholder perspectives of clinical artificial intelligence implementation: systematic review of qualitative evidence. J Med Internet Res. 2023;25:e39742. https://doi.org/10.2196/39742

126. Rogers WA. Avoiding the trap of overtreatment. Med Educ. 2014;48(1):12–4. https://doi.org/10.1111/medu.12371 PMID: 24330111

127. Thirunavukarasu AJ. Large language models will not replace healthcare professionals: curbing popular fears and hype. J R Soc Med. 2023;116(5):181–2. https://doi.org/10.1177/01410768231173123 PMID: 37199678

128. Thirunavukarasu AJ, O'Logbon J. The potential and perils of generative artificial intelligence in psychiatry and psychology. Nat Mental Health. 2024;2(7):745–6. https://doi.org/10.1038/s44220-024-00257-7

129. Van Riel N, Auwerx K, Debbaut P, Van Hees S, Schoenmakers B. The effect of Dr Google on doctor-patient encounters in primary care: a quantitative, observational, cross-sectional study. BJGP Open. 2017;1(2):bjgpopen17X100833. https://doi.org/10.3399/bjgpopen17X100833 PMID: 30564661

130. Jutel A. "Dr. Google" and his predecessors. Diagnosis (Berl). 2017;4(2):87–91. https://doi.org/10.1515/dx-2016-0045 PMID: 29536917

131. International Medical Device Regulators Forum. Software as a medical device: possible framework for risk categorization and corresponding considerations. International Medical Device Regulators Forum; 2014. Available from: https://www.imdrf.org/documents/software-medical-device-possible-framework-risk-categorization-and-corresponding-considerations

132. Lim E, Thirunavukarasu A, He YV, Monkhouse H, Fu DJ, de Pennington N, et al. Building a code of conduct for AI-driven clinical consultations. Nat Med. 2026;32(2):400–3. https://doi.org/10.1038/s41591-025-04068-w PMID: 41495407

133. Ferber D, Wölflein G, Wiest IC, Ligero M, Sainath S, Ghaffari Laleh N, et al. In-context learning enables multimodal large language models to classify cancer pathology images. Nat Commun. 2024;15(1):10104. https://doi.org/10.1038/s41467-024-51465-9 PMID: 39572531

134. International Organization for Standardization. ISO/IEC TR 24027:2021. International Organization for Standardization; 2021. Available from: https://www.iso.org/standard/77607.html

135. Fuller JG. Run-time assurance: a rising technology. In: 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC), 2020. pp. 1–9. https://doi.org/10.1109/DASC50938.2020.9256425

136. Ong JCL, Ning Y, Collins GS, Bitterman DS, Beecy AN, Chang RT, et al. International partnership for governing generative artificial intelligence models in medicine. Nat Med. 2025;31(9):2836–9. https://doi.org/10.1038/s41591-025-03787-4 PMID: 40588674