WISELEAF SCIENTIFIC VENTURES

SMART HEALTHCARE REVOLUTION: AI STRATEGIES FOR SUPERIOR GLOBAL HEALTH

CHAPTER 1- THE EVOLUTION OF AI IN GLOBAL HEALTH

Sangeeta Gupta

Research Scholar, Department of Chemistry, IFTM University, Lodhipur Rajput, Delhi Road, NH-24, Moradabad, Uttar Pradesh 244102, India

Email: sg0619916@gmail.com

ABSTRACT

Artificial intelligence (AI) has been part of medicine for almost seventy years, yet its history is often told as a recent burst of progress. This chapter traces that longer arc and asks what each period actually delivered for patients. It begins with the 1955 Dartmouth proposal and the probabilistic and rule-based programs that followed, including de Dombal's Bayesian system for abdominal pain, MYCIN, INTERNIST-1/CADUCEUS and DXplain, and explains why expert systems performed well in evaluations but rarely entered routine care. It then describes the shift to data-driven learning and the deep learning inflection that followed the 2012 ImageNet result, with retinal and dermatological image classifiers as the defining demonstrations. Regulatory milestones are examined next, from the April 2018 De Novo authorisation of IDx-DR as the first autonomous diagnostic AI to a United States list of more than 1,600 AI-enabled devices by September 2026. The COVID-19 pandemic is treated as a stress test: systematic reviews found that almost all of hundreds of published models were at high risk of bias and that none of 62 imaging models was ready for clinical use. Foundation models, including Med-PaLM, Med-PaLM 2 and GPT-4, together with AlphaFold 2 and the 2024 Nobel Prize in Chemistry, mark the most recent phase. The chapter then surveys global governance through WHO guidance of 2021 and 2024, the EU AI Act and its amended timetable, India's ICMR guidelines of 2023 and the Ayushman Bharat Digital Mission. Cautionary cases, including racial bias in a population-health algorithm, the external validation of the Epic Sepsis Model and the evidence behind Watson for Oncology, show that accuracy claims made at development often fail to hold at deployment. The chapter closes by arguing that the central task for the coming decade is evidence generation and accountable deployment rather than further gains on benchmarks.

Keywords: artificial intelligence; expert systems; deep learning; large language models; health governance; algorithmic bias

1. Introduction

Every generation of medical computing has been announced as the one that would finally change clinical practice. In 1959 the claim was that symbolic logic and probability could formalise the physician's reasoning [1]. In the 1970s and 1980s it was that encoded expert knowledge could place a specialist consultant beside every bedside. Since 2012 it has been that neural networks trained on large datasets can match or exceed clinicians at narrow perceptual tasks, and since 2022 that general-purpose language models can reason across the whole of medicine. Each claim was partly right. Each also ran into problems that had little to do with algorithms: the cost of collecting knowledge, the messiness of clinical workflows, shifts between the data used for development and the patients met in practice, and questions of liability and trust that no benchmark could settle.

This chapter treats the history of AI in health as a sequence of such encounters between technical capability and the realities of care. The emphasis falls on evidence, meaning what was measured, in whom, and with what result, rather than on a list of inventions. Later chapters in this volume deal in depth with cardiovascular prediction, surgical robotics, mental health, federated learning, rare diseases and deployment in low-resource settings. Here those topics appear only where they illustrate a turn in the larger story.

Three questions run through the account. The first is why systems that performed well in laboratory or retrospective evaluation so often failed to change outcomes once deployed. The second concerns who sets the rules, and the gradual move from professional self-regulation to binding law in some jurisdictions and soft guidance in others. The third concerns equity: for whom the technology works, and whether populations in India and other low- and middle-income countries (LMICs) are participants in its development or merely late recipients of it.

Figure 1.1 places the main milestones on a single timeline. The long gap between the mid-1990s and 2012 is deliberate. Statistical learning advanced steadily in that period, but few individual medical systems from those years reached a wide clinical audience, and the field drew far less public attention than it had in the expert-system years or would after 2012.

2. The Symbolic Era: From Dartmouth to Expert Systems

2.1 Origins

The term artificial intelligence entered the scientific vocabulary through a proposal dated 31 August 1955, in which John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon sought support for a summer study at Dartmouth College [2]. The proposal conjectured that every feature of learning or intelligence could in principle be described precisely enough for a machine to simulate it. Medicine was not mentioned, but the ambition was soon carried into it.

Four years later, Robert Ledley and Lee Lusted published in *Science* an analysis of how physicians reason [1]. They argued that diagnosis could be decomposed into symbolic logic, probability and value theory: logic to relate symptoms to diseases, probability to weigh the likelihood of each candidate, and value theory to choose actions given uncertain outcomes. The paper is often cited as the intellectual starting point of medical decision support, and its three-part structure still describes much of what modern systems attempt.



Figure 1.1: Selected milestones in the development and governance of AI for health, 1955 to 2026. Dates are taken from the primary sources cited in Sections 2 to 8.

The first convincing clinical demonstration came from probability rather than logic. At Leeds, F. T. de Dombal and colleagues built a program that applied Bayes' theorem to findings recorded from patients with acute abdominal pain. In a prospective, real-time comparison involving 304 patients, the computer reached a diagnostic accuracy of 91.8 per cent, against 79.6 per cent for the most senior clinician who saw each case [3]. The result was striking, and it also exposed a limitation that would recur for decades. The program depended on structured findings collected by trained staff using standard forms, and its probabilities were estimated from local data. Its performance was tied to the setting in which it was built.

2.2 Rule-Based Consultants: MYCIN, INTERNIST-1 and DXplain

The 1970s brought a different idea. Instead of estimating probabilities from data, researchers tried to capture the knowledge of experts as explicit rules. MYCIN, developed at Stanford, advised on the selection of antimicrobial therapy for serious bacterial infections using a large set of if-then rules elicited from infectious disease specialists [4]. Because experts rarely reason with exact probabilities, Shortliffe and Buchanan introduced certainty factors, a calculus for combining partial belief and disbelief that was simpler to elicit than conditional probabilities [5]. The design traded mathematical rigour for something experts could articulate and inspect.

MYCIN's best-known evaluation was a blinded comparison published in 1979. Eight infectious disease specialists from outside Stanford rated therapy recommendations for ten cases of meningitis, without knowing whether each came from MYCIN or from a human prescriber. MYCIN's recommendations were judged acceptable in 65 per cent of ratings, compared with 62.5 per cent for the best-performing Stanford faculty member [4,6]. By the standards of the time this was a strong result. Yet MYCIN did not move into routine care. Its developers themselves noted that its acceptability to practising physicians and its effect on patient care remained to be tested in a clinical environment [4]. In an era before networked hospital records, every case also had to be entered by hand, and nobody had settled who would be responsible if the program's advice were wrong.

INTERNIST-1, built at the University of Pittsburgh by Harry Pople and the internist Jack Myers, attempted something far broader: diagnosis across the whole of internal medicine. Its knowledge base came to include more than 4,100 manifestations of disease, written in INTERLISP and run on the mainframe computers of the day [7]. A 1982 evaluation in the *New England Journal of Medicine* tested the program on 19 clinicopathological cases from the Massachusetts General Hospital series. Its performance was qualitatively similar to that of the hospital clinicians but inferior to that of the case discussants, and the authors concluded that it was not yet reliable enough for clinical use [8]. They listed specific weaknesses: poor reasoning about anatomy and time, difficulty constructing differentials when several problems coexisted, occasional misattribution of findings, and no ability to explain its conclusions [8].

INTERNIST-1 evolved into CADUCEUS and later into Quick Medical Reference (QMR), which offered a friendlier interface for entering findings [7].

DXplain, developed at Massachusetts General Hospital and described in 1987, took a more modest line [9]. It generated ranked differential diagnoses from entered findings and explained why each disease was suggested. It was presented as an aid to memory and education, not as a substitute for judgement. That modest framing anticipated the role that later evaluations would assign to every program of its kind.

2.3 Why Expert Systems Stalled

By the early 1990s the limits of the approach were visible. In 1994 Berner and colleagues published in the *New England Journal of Medicine* a head-to-head comparison of four computer-based diagnostic systems on a common set of challenging cases [10]. The study framed the programs as aids that might prompt a clinician to consider overlooked possibilities, judged by how well their lists matched expert opinion. That framing was itself telling. Two decades after INTERNIST-1 had set out to act as a consultant, the most careful evaluations treated these programs as reminders rather than as autonomous diagnosticians.

The reasons were structural rather than incidental. The first was the knowledge acquisition bottleneck. Every rule and every disease profile had to be written by hand, checked by clinicians and maintained as medicine changed, a burden that grew with each disease added [7]. The second was brittleness. Systems performed well within the scope their authors anticipated and poorly at its edges, with no awareness of where those edges lay. The third was workflow. Before electronic records, every finding had to be typed in by a clinician who had already done the intellectual work of eliciting it. The fourth was accountability: regulators, insurers and courts had no framework for software that offered clinical advice.

The lesson most relevant today is that good performance in a retrospective or simulated evaluation did not translate into clinical use. A system had to fit the way care was delivered, and someone had to be accountable for it. Both points would return in the deep learning era with very little modification.

3. The Statistical Turn

The gradual replacement of hand-written rules by models estimated from data was less a revolution than a change in where effort went. De Dombal’s Bayesian program had already shown that probabilities learned from local cases could outperform clinicians in a narrow task [3], and even MYCIN’s certainty factors were an attempt to handle uncertainty that pure logic could not [5]. Through the 1990s and 2000s, logistic regression, Bayesian networks, decision trees, support vector machines and ensemble methods were applied to risk prediction, triage,

laboratory interpretation and signal processing. Electronic health records, digitised images and genomic data gradually supplied the raw material.

The strengths of the statistical approach were precisely the weaknesses of the expert systems. Knowledge did not have to be elicited rule by rule; it was estimated from examples. Models could be refitted when practice changed. Performance could be quantified with calibration curves and areas under the receiver operating characteristic curve (AUC), which made comparison across systems possible.

The cost was a new kind of fragility. A model learned whatever regularities its training data contained, including those that reflected how care was delivered rather than the biology of disease. Because models were fitted to particular populations, they could fail quietly when moved elsewhere. These problems, later grouped under the terms dataset shift and shortcut learning, were recognised in the statistical literature but became prominent only when deep learning made models both far more powerful and far harder to inspect.

Two features of this period shaped what came after. One was the choice of prediction target. A statistical model optimises whatever outcome it is given, and in health records the outcomes most easily available are administrative: admissions, billing codes, costs, prescriptions. Each is a proxy for the clinical state of interest, and each carries the habits and inequities of the system that generated it. The consequences of that choice, which looked like a technical convenience, became visible only years later in algorithms deployed at national scale (Section 9.1).

The other was the tension between interpretability and performance. A logistic regression with a dozen variables can be read and argued over by clinicians; an ensemble of thousands of trees cannot. Expert systems had at least offered a trace of the rules they fired, and DXplain had made explanation a design goal [9]. As models grew more accurate they also grew more opaque, and the question of whether a clinician can be expected to act on a recommendation whose basis cannot be examined was left largely unanswered when deep learning arrived.

4. The Deep Learning Inflection

4.1 ImageNet and the Return of Neural Networks

Neural networks had been studied since the 1950s and had repeatedly fallen out of favour. That changed in 2012, when a deep convolutional network trained on graphics processors by Krizhevsky, Sutskever and Hinton won the ImageNet Large Scale Visual Recognition Challenge with a top-5 test error of 15.3 per cent, against 26.2 per cent for the next best entry [11]. The network had about 60 million parameters and 650,000 neurons. Its significance lay less in the architecture than in the demonstration that large labelled datasets, cheap parallel

computation and learned rather than hand-engineered features could together produce a step change in perceptual tasks.

Medicine is full of perceptual tasks. Radiologists, pathologists, dermatologists and ophthalmologists spend much of their time classifying images, and many such images were already archived digitally. Within four years the first high-profile medical applications appeared.

4.2 Retina and Skin: The Defining Demonstrations

In December 2016, Gulshan and colleagues at Google reported a deep learning algorithm for detecting referable diabetic retinopathy in retinal fundus photographs [12]. The network was developed on 128,175 images graded by ophthalmologists and validated on two independent datasets. On EyePACS-1 it achieved an AUC of 0.991, and on Messidor-2 an AUC of 0.990; at a high-specificity operating point, sensitivity was 90.3 per cent and specificity 98.1 per cent on EyePACS-1 [12]. An accompanying editorial argued that diabetic retinopathy, a leading cause of vision loss worldwide, met the standard criteria for a screening target [13]. As later commentary recalled, it also flagged the need for validation in other populations, the difficulty of fitting the algorithm into clinical workflows and the problem of persuading clinicians and patients to trust a black box [14].

A year later, Ting and colleagues in Singapore reported a system trained and validated on retinal images from multiethnic populations, using one primary validation dataset from the national screening programme and ten external datasets from Singapore, China, Hong Kong SAR, the United States, Australia and Mexico [15]. For referable diabetic retinopathy in the primary validation set the AUC was 0.936, with sensitivity of 90.5 per cent and specificity of 91.6 per cent [15]. The lower figures than in the Google study are instructive. Real screening populations contain more poor-quality images and more ambiguous cases than curated research sets, and a more demanding validation tends to produce more modest but more credible numbers.

In February 2017, Esteva and colleagues at Stanford trained a convolutional network on 129,450 clinical images covering 2,032 skin diseases and compared its classifications with those of 21 board-certified dermatologists [16]. On biopsy-proven test images, the network matched the dermatologists on two binary tasks, keratinocyte carcinoma versus benign seborrhoeic keratosis and malignant melanoma versus benign naevus [16]. The paper’s title, with its claim of dermatologist-level classification, did much to fix public expectations of what medical AI could do.

Table 1.1 compares the main eras of medical AI, and Table 1.2 summarises the landmark imaging studies alongside later evidence from deployment.

Table 1.1: Eras of medical AI, their defining systems and their characteristic limitations

Era (approx.)	Dominant paradigm	Defining systems or results	Typical evidence	Characteristic limitation
1955-1970	Formal logic and probability	Dartmouth proposal [2]; Ledley and Lusted's framework [1]	Conceptual analyses	Little computing power or clinical data
1970-1995	Probabilistic programs and rule-based expert systems	de Dombal's abdominal-pain program [3]; MYCIN [6]; INTERNIST-1/CADUCEUS [8]; DXplain [9]	Retrospective or simulated case comparisons; one prospective study [3]	Knowledge acquisition bottleneck; brittleness; manual data entry; unclear accountability [10]
1995-2012	Statistical machine learning	Regression, Bayesian networks, tree ensembles applied to records and signals	Internal validation, some external validation	Dependence on local data; dataset shift
2012-2021	Deep learning on images and signals	ImageNet result [11]; retinopathy [12,15]; skin lesions [16]; IDx-DR [17]	Large retrospective test sets; first prospective pivotal trials [17]	Opaque models; shortcut learning; few outcome trials [18]
2022 onward	Foundation models and large language models	Med-PaLM and Med-PaLM 2 [19,20]; GPT-4 [21]; AlphaFold 2 [22]	Benchmarks and examination questions; expert rating studies	Fabrication, bias, unclear intended use, weak outcome evidence [23]

4.3 From Retrospective Accuracy to Prospective Evidence

Retrospective accuracy on archived images answers a narrow question: whether the algorithm agrees with a reference standard on data like those it was built from. It does not show that patients benefit when the algorithm is used. Two kinds of evidence are needed for that: prospective studies in the intended setting and, ideally, trials with clinical or process outcomes.

The first came in 2018. Abràmoff and colleagues conducted a pivotal trial of IDx-DR, an autonomous system that analyses retinal photographs and returns a screening result without a clinician interpreting the images [17]. The trial enrolled 900 adults with diabetes at ten primary care sites in the United States, of whom 819 had complete data. Against a reading-centre reference standard, sensitivity was 87.2 per cent and specificity 90.7 per cent after correction for the enrichment of the sample, with observed values of 87.4 and 89.5 per cent [17,24]. These figures are lower than those reported in the retrospective studies, yet in some respects they mean more, because they were obtained prospectively, in primary care settings, against pre-specified endpoints.

Commentators were careful about what the trial did and did not show. Keane and Topol welcomed it as the first prospective evaluation of a commercially available autonomous diagnostic AI, while noting its modest size, its enrichment strategy, its confidence intervals that

straddled the superiority endpoint and its exclusion of low-quality images. Their central point was that diagnostic accuracy does not establish clinical effectiveness [25]. Nsoesie made a related argument about a primary care evaluation of retinal screening in Western Australia: curated training data hide weaknesses such as poor image quality that appear only in real clinics, and published clinical evaluations of AI diagnostic tools were then scarce [18].

Randomised evidence on outcomes has been slow to accumulate, but some now exists. In the ACCESS trial at Johns Hopkins, 164 young people with diabetes were randomised either to point-of-care autonomous AI screening or to standard referral to an eye specialist. Within six months, all 81 participants in the AI arm completed the diabetic eye examination, compared with 22 per cent in the control arm [26]. The outcome is a process measure, not vision saved, but it shows that the main benefit of autonomous screening may come from removing a referral step that many patients never complete rather than from superior accuracy.

Evidence from India is part of this story. A clinical validation study at Aravind Eye Hospital in Madurai and Sankara Nethralaya in Chennai compared the Google algorithm with manual grading in 3,049 patients. For moderate or worse retinopathy the algorithm had a sensitivity of 88.9 per cent and specificity of 92.2 per cent at Aravind, and 92.1 and 95.2 per cent at Sankara, figures broadly comparable to those of retina specialists and trained graders at the same sites, with a different balance of sensitivity and specificity [27]. A later evaluation drew on a programme that had screened more than 600,000 patients across 45 Aravind-affiliated sites in southern India, reporting a sensitivity of 97.0 per cent and specificity of 96.4 per cent for severe non-proliferative or proliferative disease in a random sample of 4,537 screened patients [28]. Chapter 10 examines what such programmes require in practice. For the purposes of this history, the point is that some of the most substantial real-world evidence on autonomous screening now comes from LMIC settings.

Table 1.2: Landmark deep learning studies in medical imaging and the evidence that followed

Study	Task and setting	Design and size	Key reported performance	What it did not show
Gulshan et al., 2016 [12]	Referable diabetic retinopathy; US and French image sets	Retrospective; 128,175 development images	AUC 0.991 (EyePACS-1), 0.990 (Messidor-2)	Performance in routine screening; effect on outcomes
Ting et al., 2017 [15]	Retinopathy and related eye disease; multiethnic populations	Retrospective; primary plus 10 external validation sets	AUC 0.936; sensitivity 90.5%; specificity 91.6%	Prospective use; clinical benefit
Esteva et al., 2017 [16]	Skin lesion classification	Retrospective; 129,450 images, 2,032 diseases; 21 dermatologists	Comparable to dermatologists on two binary tasks	Performance on photographs taken in practice; diverse skin types
Abràmoff et al., 2018 [17]	Autonomous retinopathy	Prospective pivotal trial; 900 enrolled, 819 analysed	Sensitivity 87.2%; specificity 90.7%	Patient outcomes; cost-effectiveness [25]

Study	Task and setting	Design and size	Key reported performance	What it did not show
	screening; 10 US primary care sites			
Gulshan et al., 2019 [27]	Retinopathy screening at two eye hospitals in India	Clinical validation; 3,049 patients	Sensitivity 88.9-92.1%; specificity 92.2-95.2%	Use outside specialist hospitals
Wolf et al., 2024 [26]	Point-of-care screening in youth with diabetes	Randomised controlled trial; 164 randomised	Exam completion 100% vs 22%	Long-term visual outcomes
Brant et al., 2025 [28]	Deployed screening across 45 sites in southern India	Random sample of 4,537 from a programme screening over 600,000 patients	Sensitivity 97.0%; specificity 96.4% (severe NPDR or PDR)	Prospective outcome data

NPDR, non-proliferative diabetic retinopathy; PDR, proliferative diabetic retinopathy.

5. Regulation Catches Up

On 11 April 2018 the United States Food and Drug Administration (FDA) granted a De Novo request for IDx-DR [24]. The decision created a new device classification for retinal diagnostic software. It was the first AI device authorised to make diabetic retinopathy screening decisions without a clinician interpreting the images [18]. The De Novo pathway mattered. It is used for novel devices that have no existing equivalent on the market, and the classification it creates can serve as a reference point for later products. The FDA also reduced risk by narrowing the intended use: the device was authorised for screening adults without visual symptoms, with referral of those found to have more than mild retinopathy [14].

Since then the number of authorised devices has grown rapidly. The FDA maintains a public list of AI-enabled medical devices and states that, as of September 2026, it had authorised more than 1,600 such devices for marketing in the United States [29]. The FDA updates the list periodically [29].

These counts need careful reading. Windecker and colleagues analysed the 903 devices on the list up to August 2024. They found that 97.1 per cent had been cleared through the 510(k) pathway and 76.6 per cent were radiology devices [30]. Clinical performance studies were reported for only 55.9 per cent of devices. Of those 505 studies, 41 were prospective and 12 randomised [30]. The 510(k) route rests on substantial equivalence to a device already on the market and seldom requires new clinical trials.

The concentration in radiology reflects the history traced in Section 4: imaging was where deep learning first worked, where digital archives already existed and where a model's output could be framed as a second read rather than an autonomous decision. It also means that the regulatory record is a poor guide to the needs of most health systems. By implication, only a modest share of authorised products addresses primary care, maternal and child health or infectious diseases, which account for much of the burden of illness in LMICs. A country that

relies on the United States list as a proxy for what is safe and useful will find much of it irrelevant and some of it untested in comparable populations. The narrowing of intended use that made IDx-DR acceptable to the FDA, restricting it to asymptomatic adults and referral decisions [14], is a reminder that authorisation always attaches to a specific use in a specific setting. A count of authorisations is therefore a measure of commercial and regulatory activity. It says little about clinical benefit: in the Windecker analysis, prospective studies had been reported for fewer than one device in twenty [30].

Regulation also had to confront a problem that rule-based systems never posed: models that change. A static algorithm can be tested once, but a model that is retrained on new data may behave differently after authorisation, and one that is left unchanged may still degrade as the population or equipment around it shifts. Both possibilities push regulators toward oversight across the whole product life cycle rather than a single decision at the point of market entry. The FDA's answer was the predetermined change control plan (PCCP), set out in final guidance posted in December 2024 [31]. In a PCCP, the manufacturer describes in its marketing submission the modifications it plans, the protocol and acceptance criteria by which they will be validated, and an assessment of their impact. Changes that stay within the plan can then be made without a new submission [32]. The governance layers that now surround these devices are discussed in Section 8.

6. COVID-19 as a Stress Test

The pandemic created conditions that seemed ideal for AI: an urgent need, a flood of new data and an international research community willing to share. It produced instead one of the clearest demonstrations of how weak much of the published evidence was.

Wynants and colleagues began a living systematic review of prediction models for COVID-19 diagnosis and prognosis in early 2020 and updated it as the literature grew [33]. In a later update covering prognostic models, the review identified 310 studies describing 606 models or validations. Of these, 545 were judged at high risk of bias and 32 at unclear risk; only 29 were at low risk [33]. The review did identify a small number of models suitable for particular uses, among them QCovid for risk stratification in the general population and the ISARIC4C Deterioration model for hospital prognosis, while still recommending further validation [33].

Roberts and colleagues focused on machine learning models that used chest radiographs or computed tomography [34]. From 2,212 studies identified, 415 remained after initial screening and 62 were included after quality screening. None of the 62 was judged to be of potential clinical use [34]. The failures were often elementary. Many papers used a pneumonia dataset of images from young children as non-COVID controls, so that a model could succeed merely by telling children from adults. Some algorithms were trained and tested on identical or overlapping data. Most development datasets were small, and 19 of 32 papers used fewer than

2,000 data points. So-called Frankenstein datasets, assembled from other collections and redistributed under new names, carried hidden duplicates [34]. Figure 1.2 summarises both reviews.

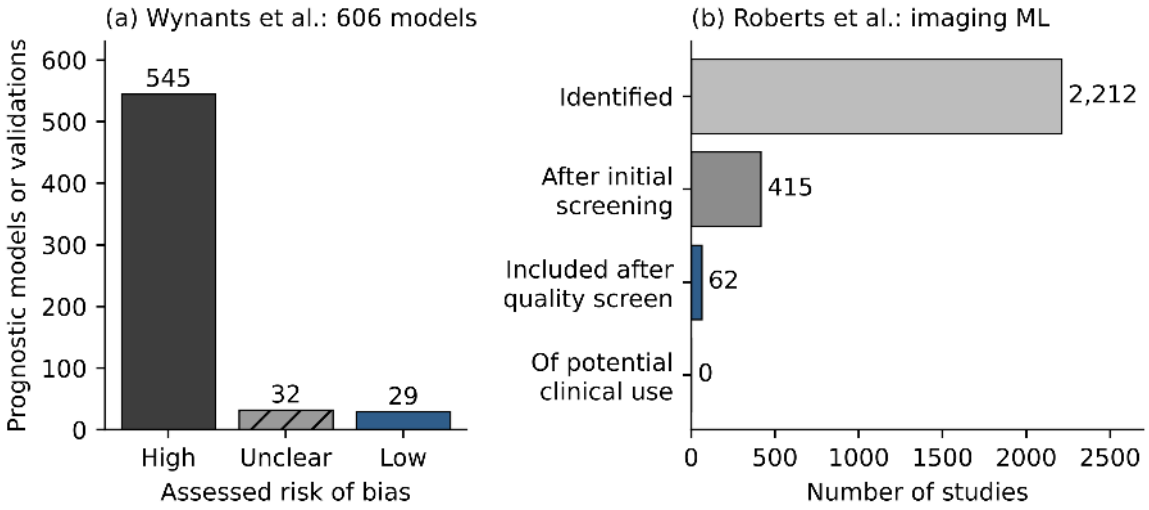

Figure 1.2: COVID-19 as a stress test for clinical prediction. (a) Assessed risk of bias among 606 prognostic models or validations in a later update of the Wynants living review [33]. (b) Study selection in the Roberts review of machine learning models using chest radiographs and CT, in which none of 62 included models was judged of potential clinical use [34].

The pandemic thus served as a natural experiment. When incentives rewarded speed and publication, the field produced hundreds of models, most of them unfit for use. The weaknesses that the reviews documented were not new; they were the familiar ones of small and unrepresentative data, absent external validation and evaluation detached from the clinical decision the model was meant to inform. What changed after 2020 was that these weaknesses could no longer be treated as teething problems.

A second lesson concerns data. The pandemic showed how quickly clinical data can be shared when there is political will, and how little that sharing is worth without documentation of where data came from and how cases were labelled. Merged public collections allowed hundreds of groups to train models within weeks, but the same collections carried the confounders and duplicates that made most of those models meaningless [34]. Curation, not volume, turned out to be the scarce resource.

Reporting standards began to catch up in the same period. In 2020 the SPIRIT-AI and CONSORT-AI extensions added 15 and 14 AI-specific items to the established guidelines for trial protocols and trial reports [35]. Among them are the version of the algorithm used and whether the AI intervention or its code can be accessed [36]. A review of 65 randomised trials of AI interventions published after CONSORT-AI found a median concordance of 90 per cent.

Only 10 trials explicitly reported using the guideline, the algorithm version was stated in 20 per cent, and only 3 of 52 journals endorsed or mandated it [36]. Standards exist, in other words, but adherence to the items most specific to AI remains weak.

7. Foundation Models and Large Language Models

7.1 From Narrow Classifiers to General Models

Every system discussed so far was built for one task. Foundation models invert that logic. A single very large model is pretrained on broad data, usually text, images or both, and then adapted to many tasks with little task-specific training. Large language models (LLMs) are the most visible kind.

The medical capabilities of LLMs were first assessed mainly through examination questions. Singhal and colleagues introduced MultiMedQA, a benchmark combining six existing question-answering datasets with a new set of consumer health queries, HealthSearchQA [19]. On MedQA, which uses questions in the style of the United States Medical Licensing Examination (USMLE), the Flan-PaLM model reached 67.6 per cent accuracy, more than 17 percentage points above the previous best result of 50.3 per cent [19]. Human evaluation, however, revealed gaps in the long-form answers, and the authors developed Med-PaLM through instruction prompt tuning to narrow them [19]. Its successor, Med-PaLM 2, scored up to 86.5 per cent on MedQA. In a pairwise ranking study of 1,066 consumer medical questions, physicians preferred Med-PaLM 2 answers to physician-written answers on eight of nine clinical axes, though answers from specialists were still preferred overall in a pilot with real-world specialist questions [20].

General-purpose models soon matched or exceeded these specialised systems. In March 2023 Nori and colleagues reported that GPT-4, without specialised prompting, exceeded the USMLE passing score by more than 20 points and outperformed both GPT-3.5 and Med-PaLM [21]. Later that year, a prompting strategy called Medprompt took GPT-4 past 90 per cent on MedQA [37]. Scores rose so quickly that the examination benchmark began to lose its power to separate leading models.

7.2 What the Benchmarks Do Not Measure

Examination performance is a weak proxy for clinical value. Multiple-choice questions present curated information in a fixed format, whereas clinical work involves eliciting history, handling incomplete and contradictory data, and taking responsibility for decisions over time. LLMs can generate fluent text that is wrong, can reproduce biases in their training data and can be confidently mistaken. WHO's 2023 statement on LLMs warned that responses can sound authoritative while being inaccurate, that training data may have been used without consent, and that such models can be used to generate convincing health disinformation [38].

A recent test makes the gap concrete. Hunt, Dai and Abràmoff evaluated commercial multimodal models, including GPT-4o, Gemini and Grok, on diabetic eye screening, the task on which narrow autonomous AI first won regulatory authorisation [39]. The best configuration, GPT-4o with background prompting, reached a sensitivity of 88 per cent but a specificity of only 63 per cent, with an AUC of 0.83, well below the AUC of about 0.94 estimated for retina specialists and short of the thresholds used to authorise autonomous devices [39]. Breadth of capability had not produced reliability on a specific, well-defined clinical task.

The authors went further, speculating that if general models improve, regulators might one day consider a task-specific licensing route loosely modelled on the licensing of physician assistants, while stating explicitly that they did not endorse the idea [39]. The speculation points to a real difficulty. Existing device regulation assumes a defined intended use, a fixed or controlled algorithm and a manufacturer who can describe both. A general-purpose model that can be prompted to perform almost any clinical task fits none of these assumptions. WHO's response has been to recommend that general models used in health be assessed and approved by a designated regulator, that laws require them to meet ethical and human rights standards regardless of their risk level, and that their real-world effects be audited after release [40]. How such recommendations will be implemented, and by whom, remains one of the open questions of the field as of late 2026.

7.3 AlphaFold and Scientific Discovery

Some of the clearest successes of the foundation-model period lie outside the clinic. AlphaFold 2, developed by DeepMind, predicted protein three-dimensional structures with a median backbone accuracy of 0.96 Å in the CASP14 assessment, against 2.8 Å for the next best method [22]. The team went on to predict structures for almost all of the roughly 200 million known proteins, and by October 2024 more than two million people from 190 countries had used it [41]. In 2024 the Nobel Prize in Chemistry was divided between David Baker, for computational protein design, and Demis Hassabis and John Jumper, for protein structure prediction [41].

For health, the significance of AlphaFold is indirect but substantial. Structure prediction supports target identification and structure-based drug design, and machine learning now features throughout the computational pipeline that screens and optimises candidate molecules [42]. The same caution applies here as at the bedside, however. A predicted structure or a computationally prioritised molecule is a hypothesis, and the path from a predicted structure to an approved medicine still runs through synthesis, assays, animal studies and clinical trials.

Table 1.3: Selected milestones of foundation models in medicine and what they measured

System	Year	Evaluation	Reported result	Limitation of the evidence
Flan-PaLM / Med-PaLM [19]	2022-2023	MultiMedQA; MedQA (USMLE-style)	67.6% on MedQA, about 17 points above prior best	Multiple-choice benchmark; human raters found gaps in long-form answers
GPT-4 [21]	2023	USMLE practice materials; MultiMedQA	More than 20 points above USMLE pass mark	Preprint; examination format only
GPT-4 with Medprompt [37]	2023	MedQA	Above 90%	Prompting strategy; benchmark saturation
Med-PaLM 2 [20]	2023-2025	MedQA; physician rating of 1,066 consumer questions	Up to 86.5%; preferred on 8 of 9 axes	Specialist answers still preferred overall; no patient outcomes
AlphaFold 2 [22,41]	2021	CASP14 blind protein structure prediction	Median backbone error 0.96 Å vs 2.8 Å	Structure is not function; drug development still needed
Commercial multimodal LLMs [39]	2026	Diabetic eye screening images	Best AUC 0.83; specificity 63%	Below specialist performance and autonomous-device thresholds

8. Global Governance

8.1 WHO Guidance

WHO entered the field with *Ethics and governance of artificial intelligence for health*, published on 28 June 2021 [43]. The guidance set out six principles: protect autonomy; promote human well-being, human safety and the public interest; ensure transparency, explainability and intelligibility; foster responsibility and accountability; ensure inclusiveness and equity; and promote AI that is responsive and sustainable [38,43]. The principles were deliberately general so that countries with very different legal traditions could adopt them.

Two further documents made the guidance more operational. In October 2023 WHO, with the International Telecommunication Union (ITU), published a document setting out regulatory considerations on AI for health, addressed to regulators and developers [44]. On 18 January 2024 WHO released guidance on large multimodal models, which identified five broad areas of application: diagnosis and clinical care, patient-guided use, clerical and administrative work, medical and nursing education, and scientific research and drug development [23,40]. It made more than 40 recommendations. Among them, governments were urged to designate a regulator to assess and approve these models for health use, to require independent post-release audits and impact assessments for large-scale deployments, with results broken down by user characteristics such as age, race or disability, and to fund public or not-for-profit computing infrastructure and datasets [40]. Developers were asked to involve clinicians, researchers and patients from the earliest design stages and to build models for well-defined tasks [40].

WHO guidance is not binding. Its influence depends on uptake by national regulators and professional bodies, and on initiatives such as the Global Initiative on AI for Health, established with other United Nations agencies [44].

8.2 The EU AI Act

The European Union adopted the first comprehensive horizontal law on AI. Regulation (EU) 2024/1689, the Artificial Intelligence Act, was published in the Official Journal on 12 July 2024 and entered into force on 1 August 2024 [45,46]. It classifies AI systems by risk. Practices judged unacceptable are prohibited, high-risk systems must meet requirements for risk management, data governance, technical documentation, transparency, human oversight, accuracy and cybersecurity, and general-purpose AI models carry their own obligations [45]. AI that is a medical device, or a safety component of one, subject to third-party conformity assessment under EU medical device law falls into the high-risk category through Annex I of the Act.

The Act applies in phases. Prohibitions and AI literacy obligations applied from 2 February 2025, and rules for general-purpose AI models, governance and penalties from 2 August 2025 [46,47]. The original text made the Act generally applicable from 2 August 2026, with high-risk obligations for Annex I products, including medical devices, from 2 August 2027. These dates were changed in 2026. Under the so-called Digital Omnibus on AI, a political agreement reached in May 2026 was adopted by the European Parliament and the Council in June and published as Regulation (EU) 2026/1744 on 24 July 2026 [46,47]. High-risk obligations for stand-alone Annex III systems now apply from 2 December 2027, and those for Annex I product-embedded systems, the category that covers AI medical devices, from 2 August 2028 [46,47]. As of November 2026, therefore, manufacturers of AI medical devices in the EU remain governed mainly by the Medical Device and In Vitro Diagnostic Regulations, with AI Act obligations for their products almost two years away.

8.3 India: ICMR Guidelines and the Digital Mission

India has so far chosen guidance over horizontal AI legislation. In 2023 the Indian Council of Medical Research (ICMR) issued *Ethical guidelines for application of artificial intelligence in biomedical research and healthcare*, prepared by the DHR-ICMR Artificial Intelligence Cell [48]. The guidelines set out ten principles: autonomy; safety and risk minimisation; trustworthiness; data privacy; accountability and liability; optimisation of data quality; accessibility, equity and inclusiveness; collaboration; non-discrimination and fairness; and validity [48]. They also address ethics committee review, informed consent and the separate stages of development, validation and deployment, and they describe themselves as a living document to be revised as practice evolves [48]. Their force comes from the role of ICMR guidance in research ethics review rather than from statute.

Two provisions deserve emphasis. The guidelines assign accountability according to the cause of harm. Where a malfunction stems from flaws in functionality, the designer, developer or manufacturer may be held responsible. Where harm follows from defective implementation, the end user or organisation may be accountable instead [48]. This departs from the assumption that the treating clinician alone carries responsibility. The guidelines also state that patients must be free to choose or reject AI technologies [48]. Both provisions are easier to state than to implement. A patient cannot meaningfully refuse a tool whose use is invisible, and liability cannot be allocated without records of which algorithm, in which version, informed which decision.

Digital infrastructure is the other half of India's approach. The Ayushman Bharat Digital Mission (ABDM) assigns citizens an Ayushman Bharat Health Account (ABHA) and builds registries of facilities and professionals so that records can follow patients. By May 2026 more than 90 crore (900 million) ABHA accounts had been created and over 100 crore health records linked [49]. Linked, standardised records at this scale could in time provide the data on which AI systems for Indian populations are trained and validated. They also raise questions about consent, security and secondary use, which Chapter 8 takes up in relation to the Digital Personal Data Protection Act, 2023.

Table 1.4 summarises the main instruments, and Figure 1.3 shows how they relate to one another and to the life cycle of an AI product.

Table 1.4: Key governance instruments for AI in health

Instrument	Issuer and year	Scope	Binding status	Notable provisions
Ethics and governance of AI for health [43]	WHO, 2021	Global; all AI for health	Non-binding guidance	Six ethical principles
Regulatory considerations on AI for health [44]	WHO with ITU, 2023	Global; regulators and developers	Non-binding	Considerations for regulating AI-based health products
Guidance on large multimodal models [23,40]	WHO, 2024	Global; generative models in health	Non-binding	Over 40 recommendations; regulatory approval and post-release audits
Artificial Intelligence Act, Regulation (EU) 2024/1689 [45,46]	European Union, 2024	Horizontal; all sectors, risk-based	Binding in EU Member States	Prohibitions from 2025; high-risk requirements; medical devices high-risk via Annex I
Digital Omnibus amendments, Regulation (EU) 2026/1744 [46,47]	European Union, 2026	Amends AI Act timetable	Binding	Annex III high-risk from 2 Dec 2027; Annex I (incl. medical devices) from 2 Aug 2028
FDA AI-enabled device authorisations [24,29]	US FDA, 2018 onward	Medical devices marketed in the USA	Binding regulatory decisions	De Novo, 510(k) and premarket approval pathways; public device list; PCCP guidance 2024 [31]

Instrument	Issuer and year	Scope	Binding status	Notable provisions
Ethical guidelines for AI in biomedical research and healthcare [48]	ICMR, India, 2023	Research and healthcare in India	Guidance; applied through ethics review	Ten principles; ethics committee review; consent
Ayushman Bharat Digital Mission [49]	Government of India	National digital health infrastructure	Programme, not AI-specific law	ABHA identifiers; registries; linked records

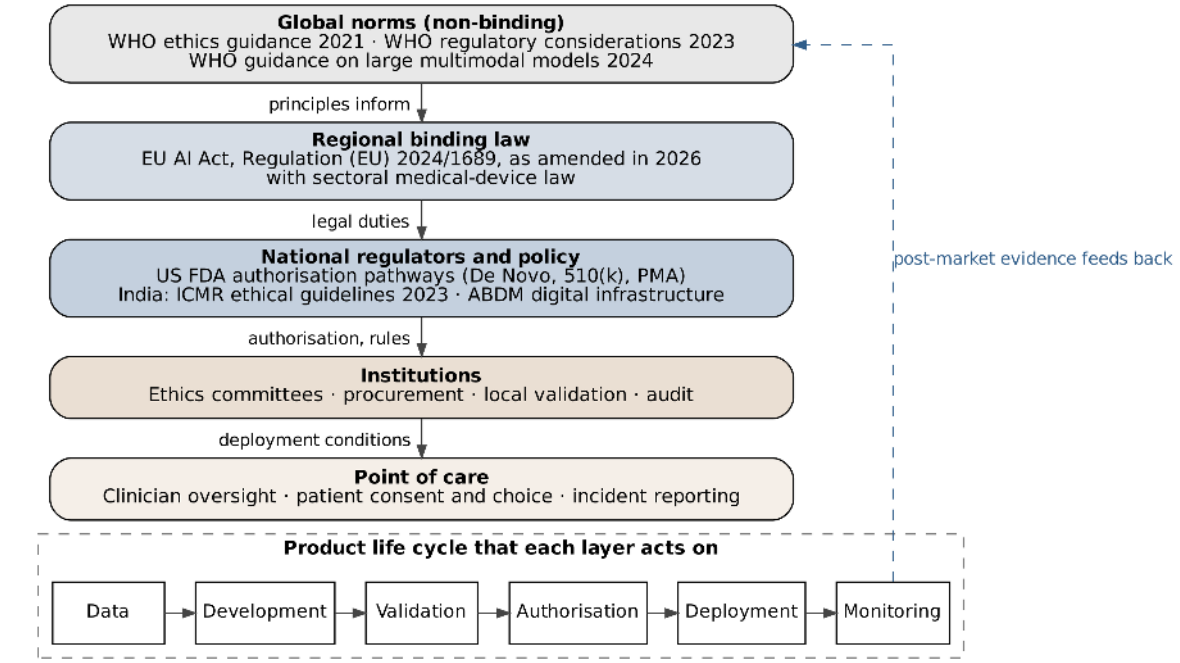

Figure 1.3: Layers of governance for AI in health and the product life cycle on which they act. Global norms inform regional and national law, which in turn shape institutional and point-of-care practice; post-market evidence should feed back to every layer. Constructed from the instruments discussed in Section 8 [23,29,43-45,47-49].

9. Equity, Bias and Cautionary Tales

9.1 Bias Built into the Target

The most influential demonstration of algorithmic bias in health involved neither images nor deep learning. Obermeyer and colleagues examined a commercial algorithm used by health systems in the United States to identify patients for care-management programmes [50]. The algorithm predicted future health care costs and used them as a proxy for health need. Because less money had historically been spent on Black patients with the same level of illness, the algorithm systematically underestimated their needs. At a given risk score, Black patients were considerably sicker than white patients. Correcting the disparity would have raised the proportion of Black patients identified for additional help from 17.7 per cent to 46.5 per cent

[50]. The source of the bias lay in the label the algorithm was asked to predict, not in its inputs, so changing the label was what removed it [50].

The lesson generalises well beyond race or the United States. An algorithm can be accurate in the statistical sense, predicting exactly what it was trained to predict, and still be unjust because the target encodes unequal access to care. In settings where use of services reflects income, distance, gender or caste as much as clinical need, any model trained on utilisation data risks the same error. This is why both the WHO and ICMR frameworks give equity and non-discrimination the status of principles rather than technical details [43,48].

9.2 Watson for Oncology

IBM's Watson for Oncology was marketed as a system that could recommend cancer treatment by drawing on literature and expert training. A prominent strand of the published evidence for it consisted of concordance studies, which measured how often the system's recommendation matched that of a local tumour board. At Manipal Comprehensive Cancer Center in Bengaluru, for example, the system's recommendations for 638 breast cancers treated between 2014 and 2016 agreed with those of the multidisciplinary board in 93 per cent of cases [51].

Concordance is not the same thing as benefit. A system can agree with clinicians without improving on them, and agreement depends on the stage mix of cases, local guidelines, drug availability and how the comparison is defined. Concordance studies reveal nothing about survival, toxicity or cost. The Watson experience illustrates a problem that applies to many commercial systems: evidence designed to support adoption is often weaker than evidence designed to test benefit, and buyers, particularly in health systems with limited evaluation capacity, may not be able to tell the two apart.

9.3 The Epic Sepsis Model

The Epic Sepsis Model was a proprietary prediction tool distributed with one of the most widely used electronic health record systems in the United States [52]. By 2021 it was in use at hundreds of hospitals, yet it had been developed and validated only by its vendor, and no independent validation had been published [52]. In 2021 Wong and colleagues evaluated it at Michigan Medicine in a cohort of 27,697 patients and 38,455 hospitalisations, of which 2,552 involved sepsis [52]. The hospitalisation-level AUC was 0.63. At the alert threshold used in practice, the model had a sensitivity of 33 per cent and a positive predictive value of 12 per cent. It generated alerts for 18 per cent of all hospitalisations, yet failed to identify 1,709 of the patients with sepsis, 67 per cent of the total [52].

These figures describe a model that missed most of the cases it was designed to find while generating enough alerts to cause alarm fatigue. The case also shows how commercial distribution can outrun evidence. A tool that arrives bundled with a record system can spread

faster than the evidence for it, and few hospitals have the capacity, or the access to a proprietary model's internals, to validate it locally. Performance also varies from site to site. Across nine hospitals in one US network, the model's C-statistic ranged from 0.55 to 0.73. It was lowest at the largest hospital and was inversely related to sepsis incidence and comorbidity burden [53].

Table 1.5: Cautionary cases in medical AI and their lessons

Case	Claim or expectation	What independent evaluation found	Lesson for governance
INTERNIST-1 [8]	Consultant-level diagnosis in internal medicine	Similar to hospital clinicians but inferior to expert discussants; judged not reliable for clinical use	Simulated case performance does not establish clinical readiness
Population-health risk algorithm [50]	Neutral identification of high-need patients	Cost proxy underestimated need among Black patients; share flagged would rise from 17.7% to 46.5% if corrected	Audit the choice of target, not only accuracy
Watson for Oncology [51]	Expert-level treatment recommendation	Evidence largely concordance with local tumour boards (93% in one Indian centre)	Concordance is not outcome evidence
COVID-19 prediction models [33,34]	Rapid support for triage and diagnosis	545 of 606 prognostic models at high risk of bias; none of 62 imaging models clinically useful	Speed without methodological standards wastes effort and risks harm
Epic Sepsis Model [52,53]	Early warning of sepsis	AUC 0.63; sensitivity 33%; 67% of sepsis cases missed; C-statistic 0.55-0.73 across 9 hospitals	Independent external validation before wide deployment
Multimodal LLMs for eye screening [39]	General models can do specialist tasks	Specificity 63%; below specialist and regulatory thresholds	Breadth does not substitute for task-specific validation

9.4 Equity at the Global Scale

Equity problems arise between countries as well as within them. Evaluation at one site does not guarantee generalisability, tools may need tailoring for low-resource regions, and performance can depend on the imaging device or on the over-representation of particular demographic groups in training data [18]. An algorithm validated in Singapore or California will not necessarily perform in Moradabad or Lagos, and the best evidence that it does must come from local evaluation, as the Indian retinal studies show [27,28]. There is also an asymmetry of power. Countries that supply data, or that serve as sites for real-world evaluation, do not always share in the intellectual property or the revenue that follow. WHO's call for public infrastructure, open datasets and audits disaggregated by population group is in part a response to this asymmetry [40].

Language adds a further layer. Clinical encounters in India take place in dozens of languages and in mixtures of them, while the text on which most large language models are trained is dominated by a few widely written languages. A model that answers fluently in English may perform worse, or fail in ways that are harder to detect, when a patient describes symptoms in

a regional language or a health worker documents a visit in a mixture of scripts. Benchmarks built from licensing examinations in English cannot reveal such gaps. Evaluation in the languages and settings where a tool will actually be used is therefore part of what equity requires, not an optional refinement.

10. Conclusion and Future Directions

Seventy years of medical AI suggest a pattern. Each era produced systems that performed impressively in the setting where they were built and less impressively elsewhere. Each underestimated the work needed to fit technology into clinical practice, and each found that questions of accountability, trust and fairness could not be postponed until the technology matured. De Dombal's program depended on local data and standard forms [3]. MYCIN performed as well as experts and did not enter routine use [4,6]. Deep learning classifiers achieved near-perfect AUC on curated data and lower figures in prospective trials [12,17]. Language models passed licensing examinations and then fell short on a narrow screening task [21,39].

The pattern does not imply that progress is illusory. Autonomous retinal screening has moved from a retrospective study to regulatory authorisation, randomised evidence and programmes screening hundreds of thousands of patients in India [17,26,28]. Protein structure prediction has changed how laboratories approach biology [22,41]. WHO, the United States, the European Union and India now have regulatory or ethical frameworks that did not exist a decade ago [29,43,45,48].

Several directions seem likely to define the next decade. The first is a shift in evidence standards from accuracy to outcomes. Trials such as ACCESS, which measure whether patients actually receive care, should become the norm for any system that claims clinical benefit [26]. The second is post-market monitoring. Models drift as populations, equipment and practice change, and the Epic experience shows that local validation cannot be assumed [52]. Life-cycle regulation, whether through the FDA's change control plans or the EU AI Act's requirements for high-risk systems, will only be as good as the monitoring data that feed it [31,45]. The third concerns language and multimodal models, which will need forms of evaluation that test behaviour in realistic clinical conversations and documentation tasks rather than examination scores, together with clear assignment of responsibility when they err [23]. The fourth is equity by design: local validation, audits disaggregated by population group and the inclusion of LMIC institutions as developers rather than only as test sites [40,48].

For India the opportunity is considerable. A digital health infrastructure covering most of the population, national ethical guidelines and an established record of rigorous local validation in retinal screening together provide foundations that many countries lack [27,48,49]. Whether these become an AI ecosystem that serves Indian patients depends on investment in evaluation

capacity, on clear regulatory pathways for software as a medical device and on protection of the people whose data make such systems possible. The history traced in this chapter suggests that the decisive work will lie less in building more capable models than in proving, setting by setting, that they help.

REFERENCES

1. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science. 1959;130(3366):9-21.
2. McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. AI Mag. 2006;27(4):12-4. Available from: https://ojs.aaai.org/index.php/aimagazine/article/view/1904
3. de Dombal FT, Leaper DJ, Staniland JR, McCann AP, Horrocks JC. Computer-aided diagnosis of acute abdominal pain. Br Med J. 1972;2(5804):9-13. https://doi.org/10.1136/bmj.2.5804.9
4. Buchanan BG, Shortliffe EH, editors. Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Reading (MA): Addison-Wesley; 1984. Available from: https://people.dbmi.columbia.edu/shortliffe/Buchanan-Shortliffe-1984/
5. Shortliffe EH, Buchanan BG. A model of inexact reasoning in medicine. Math Biosci. 1975;23:351-79.
6. Yu VL, Fagan LM, Wraith SM, Clancey WJ, Scott AC, Hannigan J, et al. Antimicrobial selection by a computer. A blinded evaluation by infectious diseases experts. JAMA. 1979;242(12):1279-82.
7. Myers JD. The background of INTERNIST I and QMR. In: Proceedings of the ACM Conference on History of Medical Informatics. New York: Association for Computing Machinery; 1987. p. 195-7. https://doi.org/10.1145/41526.41543
8. Miller RA, Pople HE Jr, Myers JD. INTERNIST-1, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med. 1982;307(8):468-76.
9. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain. An evolving diagnostic decision-support system. JAMA. 1987;258(1):67-74. https://doi.org/10.1001/jama.258.1.67
10. Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, et al. Performance of four computer-based diagnostic systems. N Engl J Med. 1994;330(25):1792-6. https://doi.org/10.1056/NEJM199406233302506
11. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84-90. https://doi.org/10.1145/3065386
12. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402-10. https://doi.org/10.1001/jama.2016.17216
13. Wong TY, Bressler NM. Artificial intelligence with deep learning technology looks into diabetic retinopathy screening. JAMA. 2016;316(22):2366-7. https://doi.org/10.1001/jama.2016.17563

14. Stead WW. Clinical implications and challenges of artificial intelligence and deep learning. JAMA. 2018;320(11):1107-8. https://doi.org/10.1001/jama.2018.11029
15. Ting DSW, Cheung CY, Lim G, Tan GSW, Quang ND, Gan A, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA. 2017;318(22):2211-23. https://doi.org/10.1001/jama.2017.18152
16. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-8. https://doi.org/10.1038/nature21056
17. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med. 2018;1:39. https://doi.org/10.1038/s41746-018-0040-6
18. Nsoesie EO. Evaluating artificial intelligence applications in clinical settings. JAMA Netw Open. 2018;1(5):e182658. https://doi.org/10.1001/jamanetworkopen.2018.2658
19. 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
20. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med. 2025;31(3):943-50. https://doi.org/10.1038/s41591-024-03423-7
21. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems [Preprint]. arXiv:2303.13375. 2023. https://doi.org/10.48550/arXiv.2303.13375
22. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9. https://doi.org/10.1038/s41586-021-03819-2
23. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: WHO; 2024. ISBN 9789240084759. Available from: https://www.who.int/publications/i/item/9789240084759
24. US Food and Drug Administration. De Novo classification request for IDx-DR (DEN180001): decision summary. Silver Spring (MD): FDA; 2018 [cited 2026 Nov 3]. Available from: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf
25. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med. 2018;1:40. https://doi.org/10.1038/s41746-018-0048-y
26. Wolf RM, Channa R, Liu TYA, Zehra A, Bromberger L, Patel D, et al. Autonomous artificial intelligence increases screening and follow-up for diabetic retinopathy in youth: the ACCESS randomized control trial. Nat Commun. 2024;15:421. https://doi.org/10.1038/s41467-023-44676-z
27. Gulshan V, Rajan RP, Widner K, Wu D, Wubbels P, Rhodes T, et al. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. JAMA Ophthalmol. 2019;137(9):987-93. https://doi.org/10.1001/jamaophthalmol.2019.2004

28. Brant A, Singh P, Yin X, Yang L, Nayar J, Jeji D, et al. Performance of a deep learning diabetic retinopathy algorithm in India. JAMA Netw Open. 2025;8(3):e250984. https://doi.org/10.1001/jamanetworkopen.2025.0984
29. US Food and Drug Administration. Artificial intelligence-enabled medical devices [Internet]. Silver Spring (MD): FDA; content current as of 2026 Sep 22 [cited 2026 Nov 3]. Available from: https://www.fda.gov/medical-devices/digital-health-center-excellence/artificial-intelligence-enabled-medical-devices
30. Windecker D, Baj G, Shiri I, Mohammadi Kazaj P, Kaesmacher J, Gräni C, et al. Generalizability of FDA-approved AI-enabled medical devices for clinical use. JAMA Netw Open. 2025;8(4):e258052. https://doi.org/10.1001/jamanetworkopen.2025.8052
31. US Food and Drug Administration. Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: final guidance [Internet]. Silver Spring (MD): FDA; 2024 Dec [cited 2026 Nov 3]. Announced in: FDA Roundup: December 3, 2024. Available from: https://www.fda.gov/news-events/press-announcements/fda-roundup-december-3-2024
32. US Food and Drug Administration. Webinar: final guidance on marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions [slides]. Silver Spring (MD): FDA; 2025 Jan 14 [cited 2026 Nov 3]. Available from: https://www.fda.gov/media/185050/download
33. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020;369:m1328. https://doi.org/10.1136/bmj.m1328
34. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021;3(3):199-217. https://doi.org/10.1038/s42256-021-00307-0
35. Ibrahim H, Liu X, Cruz Rivera S, Moher D, Chan AW, Sydes MR, et al. Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines. Trials. 2021;22:11. https://doi.org/10.1186/s13063-020-04951-6
36. Martindale APL, Llewellyn CD, de Visser RO, Ng B, Ngai V, Kale AU, et al. Concordance of randomised controlled trials for artificial intelligence interventions with the CONSORT-AI reporting guidelines. Nat Commun. 2024;15:1619. https://doi.org/10.1038/s41467-024-45355-3
37. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine [Preprint]. arXiv:2311.16452. 2023. https://doi.org/10.48550/arXiv.2311.16452
38. World Health Organization. WHO calls for safe and ethical AI for health [news release]. Geneva: WHO; 2023 May 16 [cited 2026 Nov 3]. Available from: https://www.who.int/news/item/16-05-2023-who-calls-for-safe-and-ethical-ai-for-health
39. Hunt MS, Dai T, Abràmoff MD. Evaluating commercial multimodal AI for diabetic eye screening and implications for an alternative regulatory pathway. NPJ Digit Med. 2026;9:42. https://doi.org/10.1038/s41746-025-02216-7

40. World Health Organization. WHO releases AI ethics and governance guidance for large multi-modal models [news release]. Geneva: WHO; 2024 Jan 18 [cited 2026 Nov 3]. Available from: https://www.who.int/news/item/18-01-2024-who-releases-ai-ethics-and-governance-guidance-for-large-multi-modal-models
41. Nobel Prize Outreach. The Nobel Prize in Chemistry 2024: popular information [Internet]. Stockholm: The Royal Swedish Academy of Sciences; 2024 [cited 2026 Nov 3]. Available from: https://www.nobelprize.org/prizes/chemistry/2024/popular-information/
42. Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. Nature. 2023;616:673-85. https://doi.org/10.1038/s41586-023-05905-z
43. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: WHO; 2021. ISBN 9789240029200. Available from: https://www.who.int/publications/i/item/9789240029200
44. World Health Organization. Harnessing artificial intelligence for health [Internet]. Geneva: WHO; 2026 [cited 2026 Nov 3]. Available from: https://www.who.int/teams/digital-health-and-innovation/harnessing-artificial-intelligence-for-health
45. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union. 2024;L 2024/1689 (12 July 2024). Available from: http://data.europa.eu/eli/reg/2024/1689/oj
46. Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 113: entry into force and application, consolidated text as amended by Regulation (EU) 2026/1744 [Internet]. Springlex; 2026 [cited 2026 Nov 3]. Available from: https://www.springlex.eu/en/packages/ai-act/ai-act-regulation/article-113/
47. Future of Privacy Forum. The AI Act implementation timeline: what changes under the AI Omnibus? [Internet]. Washington (DC): FPF; 2026 [cited 2026 Nov 3]. Available from: https://fpf.org/blog/the-ai-act-implementation-timeline-what-changes-under-the-ai-omnibus/
48. Indian Council of Medical Research. Ethical guidelines for application of artificial intelligence in biomedical research and healthcare. New Delhi: ICMR; 2023. ISBN 978-93-5811-343-3. Available from: https://www.icmr.gov.in/ethical-guidelines-for-application-of-artificial-intelligence-in-biomedical-research-and-healthcare
49. Press Information Bureau. Ayushman Bharat Digital Mission: India's digital health backbone [backgrounder]. New Delhi: PIB, Government of India; 2026 Jul 6 [cited 2026 Nov 3]. Available from: https://static.pib.gov.in/WriteReadData/specificdocs/documents/2026/jul/doc202676912801.pdf
50. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53. https://doi.org/10.1126/science.aax2342
51. Somashekhar SP, Sepúlveda MJ, Puglielli S, Norden AD, Shortliffe EH, Rohit Kumar C, et al. Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. Ann Oncol. 2018;29(2):418-23. https://doi.org/10.1093/annonc/mdx781

52. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. 2021;181(8):1065-70. https://doi.org/10.1001/jamainternmed.2021.2626

53. Lyons PG, Hofford MR, Yu SC, Michelson AP, Payne PRO, Hough CL, et al. Factors associated with variability in the performance of a proprietary sepsis prediction model across 9 networked hospitals in the US. JAMA Intern Med. 2023;183(6):611-2. https://doi.org/10.1001/jamainternmed.2022.7182