International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Intelligent Medicine: How Artificial Intelligence and Machine Learning Are Reshaping Modern Healthcare

From Pattern Recognition to Foundation Models- Capabilities, Evidence, Risks, and the Path to Trustworthy Clinical Deployment

Rahul Chakraborty

Independent Researcher in Healthcare Strategy & MedTech Commercialization, Dubai, United Arab Emirates

Abstract: ***Artificial intelligence (AI) and machine learning (ML) have moved from the periphery of healthcare research to the centre of clinical practice, drug development, and health-system operations. By the end of 2025 regulators had authorised well over a thousand AI-enabled medical devices; large language models had begun to absorb the documentation burden that drives clinician burnout; AI-designed molecules had entered human clinical trials; and the first randomised controlled trial of a generative-AI therapy chatbot had reported meaningful symptom reductions. Yet the same period produced sobering evidence that widely deployed models can fail silently, embed structural inequities, and resist external validation. This review synthesises the state of AI and ML across the principal clinical and operational domains- diagnostic imaging, decision support, generative and conversational AI, drug discovery, precision medicine and genomics, pathology, surgery, mental health, remote monitoring, and administrative workflow- and then examines the cross-cutting determinants of success: bias and equity, evidence and reporting standards, explainability, privacy, regulation, and economics. The central argument is that the binding constraint is no longer algorithmic capability but the surrounding system of validation, governance, and human–machine collaboration. AI in healthcare succeeds not when it is most autonomous, but when it is most accountable.***

Keywords: artificial intelligence; machine learning; deep learning; foundation models; large language models; clinical decision support; medical imaging; drug discovery; algorithmic bias; reporting standards; health-AI regulation; precision medicine

1. Introduction

1.1 A discipline at an inflection point

Medicine has always been an information science wearing the clothes of a craft. Every diagnosis is an inference under uncertainty; every treatment decision is a prediction about a future that has not yet happened. For most of the twentieth century the instruments of that inference were the trained human mind, the printed literature, and a slowly accumulating base of clinical evidence. Over the past decade a new instrument has joined the bench. Artificial intelligence and machine learning — statistical systems that learn patterns from data rather than following hand-written rules — have demonstrated, repeatedly and across specialties, that they can match or exceed expert performance on narrowly defined tasks: reading a retinal photograph, flagging a stroke on a CT scan, drafting a clinical note, predicting the three-dimensional shape of a protein, or even autonomously completing a portion of a surgical procedure on an animal model.

The scale of adoption is now difficult to overstate. By the close of 2025 the United States Food and Drug Administration (FDA) had authorised more than 1,200 AI-enabled medical devices, with radiology alone accounting for roughly three-quarters of those authorisations and surpassing one thousand cleared tools.[1,2] Market analysts, whose forecasts should always be read with caution, broadly converge on a healthcare-AI market expanding at compound annual growth rates near 37–39%, with several estimates placing the global market close to US$190 billion by 2030.[3] Behind those headline figures sits a more consequential shift: AI is no longer confined to research prototypes but is embedded in the electronic health record, the imaging workstation, the laboratory, the operating theatre, and increasingly the patient's wrist and smartphone.

1.2 Why this review, and why now

The arrival of transformer-based foundation models and clinically capable large language models (LLMs) between 2022 and 2025 marks a genuine discontinuity rather than an incremental improvement. Earlier medical AI was overwhelmingly narrow: a model trained to detect diabetic retinopathy could do nothing else. Foundation models, by contrast, are trained on vast and heterogeneous data and can be adapted to many downstream tasks, including tasks their developers never anticipated.[4,5] This generality is precisely what makes the present moment both promising and hazardous. The same model that drafts an excellent discharge summary may fabricate a plausible-sounding but non-existent laboratory value; the same architecture that achieves expert-level scores on licensing examinations may reason poorly about an atypical patient it has never seen.

This article therefore takes a deliberately balanced stance. It is neither a catalogue of triumphs nor a recitation of dangers, but an attempt to map where the evidence is strong, where it is thin, and where the field is being carried by enthusiasm ahead of proof. The guiding thesis is stated here so that it can be tested against the evidence that follows: the binding constraint on AI in medicine is no longer raw model capability, but the surrounding infrastructure of validation, governance, equity, and clinical integration.

1.3 Scope and Structure

Section 2 establishes the conceptual foundations- the families of methods, the meaning of "learning," the data modalities medicine offers, and the privacy-preserving techniques that make learning from sensitive data possible. Sections 3 through 12 examine the principal application domains in turn: diagnostic imaging; clinical decision support and predictive analytics; generative and conversational AI; drug discovery and development; precision medicine, genomics, and oncology; digital pathology; surgery and robotics; mental and behavioural health; remote monitoring and wearables; and administrative and operational AI. Sections 13 through 18 address the cross-cutting determinants of success: algorithmic bias and health equity; evidence, validation, and reporting standards; explainability, trust, and human factors; privacy, security, and data governance; regulation, governance, and ethics; and economics, workforce, and health systems. Section 19 offers a synthesis and forward-looking recommendations. Two appendices- a timeline of medical AI and a glossary- and a full reference list follow.

Throughout, claims are anchored to peer-reviewed studies, regulatory data, and authoritative reporting wherever possible. Where the evidence is contested or preliminary, this is stated plainly. The intended reader is the informed generalist- a clinician, executive, policymaker, or scientifically literate member of the public- who wants neither marketing copy nor impenetrable jargon, but an honest account of what these systems can and cannot yet do.

1.4 A note on sources and method

This is a narrative review rather than a systematic one. Its aim is synthesis and interpretation across a fast-moving and deliberately broad field, not exhaustive enumeration within a narrow one. Sources were selected to favour primary evidence- peer-reviewed clinical studies, regulatory documents, and consensus reporting guidelines- supplemented by reputable secondary reporting for the most recent developments, which by their nature have not yet entered the peer-reviewed literature. Three principles guided the selection. First, landmark studies that shaped the field's understanding are preferred to the larger mass of confirmatory work. Second, claims of clinical benefit are weighted by the rigour of their evidence, so that a prospective external validation counts for more than a retrospective single-centre result, and a company's own efficiency figures are reported as such. Third, where the technology has outrun the evidence- as it frequently has- this gap is treated as a finding in its own right rather than smoothed over. Market projections, which vary widely by source and methodology, are cited only to convey direction and scale, never precision. The reference list is numbered, and in-text superscript markers point to it.

2. Foundations: The Methods Behind Medical AI

2.1 From expert systems to learning machines

The first wave of medical AI, in the 1970s and 1980s, was built on hand-coded rules. Systems such as MYCIN encoded the knowledge of human experts as long chains of if-then statements. They were transparent and occasionally impressive, but brittle: every new piece of knowledge had to be entered by hand, and the systems could not generalise beyond their explicit rules. The modern paradigm inverts this logic. Rather than telling the machine the rules, we give it examples and let it infer the rules statistically. This is **machine learning**: the construction of mathematical models whose parameters are tuned by exposure to data so that the model performs well on a task without being explicitly programmed for every case. The shift is profound because it relocates the source of knowledge from the human expert to the data- with all the promise, and all the hidden bias, that the data carry.

2.2 Three flavours of learning

Supervised learning dominates clinical AI. The model is shown labelled examples- chest X-rays marked "pneumonia" or "normal," pathology slides annotated by expert pathologists- and learns to map inputs to outputs. The quality of the labels is decisive: a model is only as good as the ground truth it is trained against, and much of the difficulty in medical AI lies in the fact that medical "truth" is itself uncertain, contested, and sometimes biased. Inter-observer disagreement among expert clinicians is the rule rather than the exception, and a model trained to imitate one institution's consensus may not transfer to another's.

Unsupervised learning finds structure in unlabelled data- clustering patients into subtypes, or learning compact representations of complex inputs. It is central to discovering previously unrecognised disease phenotypes and to the self-supervised pre-training that underpins foundation models. **Reinforcement learning**, in which an agent learns by trial and error to maximise a reward, has shown promise in sequential decision problems such as dosing, mechanical-ventilation management, and treatment scheduling, though its clinical use remains largely experimental because the cost of a wrong action in medicine is not a lost game but a harmed patient, and because learning from real patients by trial and error is ethically impossible. Most medical reinforcement learning therefore happens offline, learning from historical records rather than live experimentation.

2.3 Deep learning and neural networks

The breakthroughs of the past decade rest on **deep learning**- neural networks with many layers that learn hierarchical representations of data. In imaging, convolutional neural networks (CNNs) learn to detect edges, then textures, then anatomical structures, then pathology, each layer building on the last. It was a CNN that, in 2016, was trained on 128,175 retinal fundus photographs graded by 54 ophthalmologists and achieved an area under the receiver-operating-characteristic curve above 0.99 for referable diabetic retinopathy- a landmark result that helped launch the modern era of clinical deep learning.[17] The power of deep learning is its capacity to learn relevant features directly from raw data, dispensing with the laborious hand-engineering of features that characterised earlier approaches. Its weakness is the corollary: the features it learns are opaque, sometimes spurious, and not guaranteed to be the clinically meaningful ones.

2.4 Transformers and the foundation-model era

The architecture that now underpins the most capable systems is the **transformer**, introduced in 2017, whose attention mechanism allows a model to weigh the relevance of every part of an input to every other part.[23] Transformers scale remarkably well: trained on enormous corpora, they yield **foundation models-** large, general-purpose models that can be adapted to many tasks. In medicine this manifests in two related families. **Large language models (LLMs)** process and generate text, and have been adapted to clinical documentation, question-answering, summarisation, and coding. **Multimodal foundation models** combine modalities- text with images, or histopathology with genomics- and are beginning to deliver genuinely integrative reasoning across the fragmented data of a patient's record.[5,20] The defining feature of this era is generality: a single pre-trained model can be fine-tuned, prompted, or retrieval-augmented to serve a multitude of clinical purposes, dramatically lowering the data and effort required to build each new application.

2.5 The data medicine offers- and the biases it carries

AI is only as useful as the data available to it, and medicine produces an extraordinary variety: structured electronic health record (EHR) fields, free-text clinical notes, radiological and pathological images, genomic sequences, continuous physiological waveforms from monitors and wearables, and administrative and claims data. Each modality carries its own biases. Claims data reflect what was billed rather than what was wrong- a distinction that, as Section 13 shows, has produced some of the most consequential failures of fairness in the field.[15] Clinical notes encode the documenting clinician's assumptions and the institution's practice patterns. Images carry the fingerprint of the scanner and protocol that produced them. None of these data sources is a neutral window onto biological truth; each is a socially and technically constructed artefact, and a model trained on it inherits its distortions.

2.6 Learning without exposing the data: federated and privacy-preserving ML

A central obstacle to medical AI is that the data are sensitive, siloed, and legally protected. A model trained at a single hospital sees a narrow slice of humanity; pooling data across institutions improves generalisability but collides with privacy law and patient trust. **Federated learning** offers a partial resolution: rather than moving patient data to a central server, the model is sent to each institution, trained locally, and only the learned parameters- not the raw data- are aggregated. Techniques such as **differential privacy** add carefully calibrated noise so that individual patients cannot be re-identified from the shared updates. Recent work has applied these methods to fine-tune language models for sensitive domains such as mental health while remaining compatible with HIPAA in the United States and the GDPR in Europe.[40] These approaches are not a panacea- federated systems are harder to coordinate and can still leak information- but they point toward a future in which models can learn from the world's clinical data without that data ever leaving the institutions that hold it.

Table 1: The principal machine-learning paradigms and their characteristic roles in healthcare. Foundation models blur these categories by combining self-supervised pre-training with task-specific adaptation; federated methods address the data-governance constraints that limit all of them

Learning paradigm	What it does	Representative clinical use
Supervised	Maps labelled inputs to outputs	Image classification; readmission and mortality prediction
Unsupervised / self-supervised	Finds structure; learns representations at scale	Patient subtyping; foundation-model pre-training
Reinforcement (offline)	Learns action policies from historical data	Dosing, ventilation, treatment scheduling (experimental)
Federated + differential privacy	Trains across sites without sharing raw data	Multi-institution models under HIPAA / GDPR

2.7 A note on terminology

Public discussion of medical AI is hampered by loose vocabulary, and precision here repays the effort. "Artificial intelligence" is the broad umbrella: any system performing tasks that would ordinarily require human intelligence. "Machine learning" is the dominant modern subset, in which systems learn from data. "Deep learning" is a subset of machine learning using multi-layer neural networks, and it powers nearly all the recent breakthroughs. "Generative AI" refers to systems that produce new content - text, images, molecules- rather than merely classifying or predicting, and large language models are its most prominent instance. These distinctions matter clinically because the risk profile differs sharply: a narrow classifier that outputs a probability is a fundamentally different object from a generative model that can fabricate fluent falsehoods, and conflating them under the single word "AI" obscures exactly the differences that should drive how each is validated and governed. Throughout this article the terms are used in these specific senses.

3. Diagnostic Imaging: The Proving Ground

3.1 Why radiology came first

No specialty has been transformed by AI as visibly as radiology, and the reasons are structural. Radiology digitised early, its images are standardised and abundant, its ground truth is comparatively well defined, and its tasks- detection, classification, segmentation- map neatly onto the strengths of deep learning. The first FDA authorisation for an AI imaging tool dates to 1998, for a mammography computer-aided detection system.[2] The pace then accelerated dramatically: a systematic review in JAMA Network Open found that of 950 AI/ML devices authorised by the FDA through late 2024, 723- some 76%- were radiology devices, and that approvals had grown from a trickle before 2016 to 221 in the single year 2023.[1]

By the FDA's own running tally, AI-enabled radiology authorisations passed the one-thousand mark in late 2025, with imaging accounting for roughly 77% of all AI medical-device authorisations since tracking began. GE HealthCare led individual companies with around 115 radiology

authorisations- a figure inflated by acquisitions- followed by Siemens Healthineers, Philips, Canon, and United Imaging, a roster that doubles as a map of consolidation in the sector.[2] Within radiology, the most active product category by late 2025 was computer-assisted detection and diagnosis, reflecting AI's central role in helping radiologists find and characterise abnormalities rather than replacing their judgement.

Although radiology dominates, image-based AI now reaches well beyond it. Dermatology classifiers distinguish malignant from benign skin lesions from photographs; ophthalmic AI extends past diabetic retinopathy to glaucoma and age-related macular degeneration; cardiology models read electrocardiograms and automate measurements from echocardiograms; gastroenterology systems flag polyps during live colonoscopy; and the same techniques are migrating into pathology, as Section 8 describes. The common thread is that wherever a clinical decision rests on the interpretation of an image or a waveform, deep learning has proven able to assist- and the breadth of that applicability, rather than performance on any single task, is what makes imaging AI structurally important to the future of the field.

3.2 The autonomous-diagnosis milestone

The single most important regulatory event in clinical AI may be the April 2018 authorisation of IDx-DR (later rebranded LumineticsCore), the first system cleared to make a diagnostic decision *without* a clinician interpreting the result.[16,27] In its pivotal trial of 900 patients across ten primary-care sites, the system detected more-than-mild diabetic retinopathy with a sensitivity of 87.2% and specificity of 90.7% against a rigorous reading-centre reference standard.[16] The clinical significance is not the accuracy figure alone but the model of care it enables: specialty-level screening delivered in a primary-care office, by a technician, for a disease that is a leading cause of preventable blindness. Reaching that milestone required roughly US$22 million in investment and eight years of validation work with the regulator.[28] That ratio- a short technical effort followed by a long, expensive validation- is the defining economic feature of autonomous medical AI, and it explains why so few systems have followed.

3.3 Triage and the value of time

A second well-evidenced use case is acute triage. Viz.ai's stroke-detection algorithm, the first FDA-cleared AI tool for neurovascular imaging, flags suspected large-vessel-occlusion strokes and alerts the stroke team directly, compressing the interval between scan and intervention. By 2025 the company reported deployment across more than 1,600 hospitals, with documented time savings that matter acutely in a condition where "time is brain."[29] Triage tools illustrate an important principle: AI can add value not by being more accurate than a radiologist but by being faster at directing human attention to the cases that cannot wait. The model does not need to be a better doctor; it needs to be a better dispatcher.

3.4 Screening, workload, and the search for the normal

A growing line of work asks whether AI can safely absorb the high volume of normal studies that consume radiologist time. Personalised breast-cancer-risk models and automated triage of normal chest radiographs are active areas, and in 2025 the FDA granted a de novo authorisation for a software tool that estimates future breast-cancer risk from mammograms- explicitly a risk-prediction aid rather than a diagnostic device.[29] The appeal is obvious: if a trustworthy system could confidently set aside the genuinely normal studies, scarce expert attention could be concentrated where it matters. The hazard is equally obvious: defining "normal" is harder than it looks, and a missed finding in a study triaged as normal is a serious harm. This is a domain where the validation bar must be exceptionally high.

3.5 The validation gap hiding behind the approvals

These successes must be read alongside an uncomfortable finding from the same JAMA Network Open review. Of the radiology devices with available submission documentation, only about 5% had undergone prospective testing, only 8% incorporated a human-in-the-loop evaluation, and fewer than a third reported any clinical testing at all; the overwhelming majority- some 97%- reached market through the 510(k) pathway by demonstrating substantial equivalence to an existing device rather than through new clinical evidence.[1] In other words, the impressive headline count of cleared devices substantially overstates the volume of rigorous clinical evidence underpinning them. This tension- abundant authorisation, scarce prospective validation- recurs throughout this review and is examined directly in Section 14.

Table 2: Selected landmark events in medical-imaging AI, illustrating the progression from narrow detection aids to autonomous diagnosis and multimodal foundation models

Milestone	Year	Specialty	Significance
Image Checker mammography CAD	1998	Radiology	First FDA-authorised AI imaging tool
Deep-learning DR algorithm (Gulshan et al.)	2016	Ophthalmology	AUC > 0.99; launched modern clinical deep learning
IDx-DR / Luminetics Core	2018	Ophthalmology	First autonomous diagnostic AI in any field
Viz.ai stroke triage	2018	Neurology	First FDA-cleared neurovascular triage AI
Virchow pathology foundation model	2024	Pathology	Pan-cancer detection, AUC 0.95

4. Clinical Decision Support and Predictive Analytics

4.1 The promise of early warning

If imaging AI answers the question "what is in this picture?", predictive analytics answers "what is about to happen to this patient?" Models built on EHR data attempt to forecast deterioration, sepsis, readmission, mortality, and complications hours or days before they become clinically obvious. The appeal is intuitive: sepsis alone accounts for nearly a million hospitalisations annually in the United States

and carries in-hospital mortality of roughly 12–15%, and earlier recognition allows earlier, life-saving treatment.[13] The same logic applies to readmission prevention, deterioration on general wards, and post-operative complications: in each case a few hours of warning can change an outcome.

4.2 A cautionary tale: the Epic Sepsis Model

No case better illustrates the gap between deployment and validation than the Epic Sepsis Model (ESM), a proprietary tool embedded in one of the most widely used EHR systems and active in hundreds of US hospitals. In 2021 an external validation by Wong and colleagues at the University of Michigan examined 38,455 hospitalisations and found that the model performed far worse in the real world than its reputation implied: an area under the curve of just 0.63, a sensitivity of only 33%, and a positive predictive value of 12%, while generating alerts on a large fraction of patients and contributing to alert fatigue.[12,13]

The finding was not a one-off. A 2024 external validation across two county emergency departments, examining nearly 146,000 encounters, found a sensitivity of just 14.7% within a six-hour window, again underscoring that performance reported by a vendor in development does not transfer automatically to a new setting.[14] The episode crystallised a principle the field had under-appreciated: a model that is never independently validated in the environment where it is used is, in effect, an untested intervention deployed at scale. The accompanying editorial in JAMA Internal Medicine made the point bluntly — external validation is not an academic nicety but a precondition for safe use.[13]

4.3 Alert fatigue and the human factor

Even an accurate predictive model fails if its outputs are ignored, mistimed, or buried in a flood of low-value notifications. Clinicians exposed to frequent false alarms learn to dismiss them — a well-documented phenomenon in which the very sensitivity that catches true cases also generates the noise that trains users to stop listening. The lesson is that a clinical decision-support tool is a socio-technical system, not a standalone algorithm. Its threshold, its timing, the channel through which it interrupts, and the action it recommends matter as much as its raw discrimination. A model that fires too often, or at the wrong moment, can make care worse even if its mathematics are sound.

4.4 Doing it well

None of this argues against predictive analytics; it argues for discipline. The most credible deployments share common features: prospective evaluation in the target population; calibration- not merely discrimination- checked against local data; silent-mode running before alerts go live; explicit linkage of each alert to a recommended action; integration into the existing workflow rather than bolted onto it; and ongoing monitoring for performance drift as patient populations and clinical practice evolve. These are the same governance principles that recur in Sections 14 through 17, and they are the difference between a model that helps and a model that merely fires.

5. Generative and Conversational AI in Clinical Practice

5.1 From benchmark scores to bedside roles

The most visible development of the 2022–2025 period was the arrival of large language models capable of expert-level performance on medical examinations. Google's Med-PaLM was the first AI system to surpass the passing threshold on US Medical Licensing Examination-style questions, scoring 67.6%; its successor Med-PaLM 2 reached 86.5%- approaching expert level- and was preferred by physicians over comparison answers across several quality dimensions, with a high proportion of its outputs judged consistent with scientific consensus.[11] Comparable general-purpose models such as GPT-4 demonstrated similar capabilities, including correct inclusion of the right diagnosis in a majority of complex diagnostic vignettes.[24]

These scores generated enormous excitement and an important caveat that the field has internalised slowly: answering examination questions is not the same as practising medicine. As one independent computer scientist observed when Med-PaLM's results were published, there is a wide gulf between answering medical questions and actually diagnosing and treating a living patient whose presentation is messy, incomplete, and atypical. Multiple-choice benchmarks reward recall of textbook associations; real practice rewards the integration of an unreliable history, an ambiguous examination, conflicting data, and the patient's own values. Benchmark performance is necessary but nowhere near sufficient evidence of clinical readiness.

5.2 The clearest win: documentation and ambient scribing

Where generative AI has delivered the most credible near-term value is not diagnosis but documentation- the administrative weight that consumes clinician time and drives burnout. A 2025 scoping review of LLMs in clinical documentation found consistent reductions in documentation time and cognitive load: ambient AI note systems generated a large share of notes in real-world ambulatory clinics, and LLM-drafted discharge and palliative summaries were produced in a fraction of the time taken manually while remaining coherent and readable.[9] In one striking comparison an LLM produced a psychiatric discharge summary in about fifteen minutes against roughly ninety minutes for a clinician, with comparable coherence- though, tellingly, the model's summaries lacked diagnostic nuance.[9] Adapted LLMs have even been shown to outperform medical experts on certain clinical-text summarisation tasks.[10] The economic logic is compelling: documentation is high-volume, low-creativity, and a major driver of clinician dissatisfaction, making it an almost ideal target for automation that augments rather than replaces.

5.3 Persistent limitations and hallucination

The same literature is candid about failure modes. Performance varies sharply across clinical contexts: models that handle a straightforward consultation well struggle with temporally complex patient trajectories, and summaries that capture surface events miss the subtle clinical reasoning a

specialist would record.[9] The deepest concern is **hallucination-** the generation of fluent, confident, but false content, an inherent consequence of the statistical, next-token nature of these models. In a domain where a fabricated dose or invented result can cause direct harm, this is not a quirk to be tolerated but a risk to be engineered against, through retrieval-grounding (constraining the model to cite source documents), human review of all consequential output, constrained output formats, and clear labelling of AI-generated content. The clinician who signs an AI-drafted note remains responsible for its accuracy, and the design of these systems must make that review tractable rather than perfunctory.

5.4 Real-world deployment is still early

Despite the hype, rigorous real-world evidence remains thin. A 2025 systematic review of LLMs in actual clinical workflows identified only a handful of qualifying studies, all using generative pre-trained transformers, covering outpatient communication, mental-health support, inbox-message drafting, and data extraction; the reported benefits in efficiency and satisfaction were real but were tempered by performance variability, limited generalisability, regulatory delay, and- critically- a near-absence of post-deployment monitoring.[24] A clinician-focused review of LLM implementation similarly emphasised the difficulty practitioners face in selecting and validating the right model for a given task.[26] The honest summary is that generative AI in medicine is past proof-of-concept and into early adoption, but well short of the mature evidence base that high-stakes clinical use ultimately demands.

6. Drug Discovery and Development

6.1 The structural-biology revolution

Perhaps the most scientifically profound contribution of AI to medicine lies upstream of the clinic, in the laboratory. The protein-folding problem- predicting a protein's three-dimensional structure from its amino-acid sequence- had resisted solution for half a century. DeepMind's AlphaFold solved it with near-experimental accuracy, work recognised by the 2024 Nobel Prize in Chemistry awarded to Demis Hassabis and John Jumper.[6,7] The 2024 successor AlphaFold 3, developed with Isomorphic Labs, extended prediction beyond single proteins to the interactions of proteins, nucleic acids, ligands, and ions- the very interactions that drug design seeks to manipulate.[7] For structural biology this was transformative: questions that once required years of crystallography can now be approached computationally in hours, opening targets that were previously intractable.

6.2 Compressing the discovery timeline

AI is now embedded across the drug-development pipeline, from target identification to clinical-trial optimisation. The most cited efficiency claims are dramatic: Insilico Medicine reported identifying a novel target for idiopathic pulmonary fibrosis and advancing a candidate to preclinical testing in roughly eighteen months- against the four-to-six years such work typically takes- and Exscientia developed an AI-designed molecule for obsessive-compulsive disorder that became among the first such molecules to enter human trials in under twelve months.[5] Recursion Pharmaceuticals combines automated high-throughput cellular imaging with deep learning to screen and repurpose molecules at scale.[5] A 2025 systematic review following PRISMA methodology confirmed the breadth of adoption across therapeutic areas, with machine learning the dominant technique, followed by molecular modelling and simulation and then deep learning.[5] These claims should be read with appropriate scepticism- they are reported largely by the companies that benefit from them- but the direction is consistent and corroborated across independent reviews.

6.3 From in silico to in human

The field is now crossing the threshold that matters: AI-designed molecules entering clinical trials. Isomorphic Labs, the Alphabet drug-discovery subsidiary built around AlphaFold, raised more than US$600 million in 2025 and announced preparations to dose its first patients, beginning with oncology candidates.[8] This is the test that will ultimately validate or deflate the AI-drug-discovery thesis. Computational design can generate promising candidates quickly and cheaply, but the brutal arithmetic of drug development- in which the overwhelming majority of candidates fail in trials for reasons of efficacy or toxicity that no current model can fully predict- remains. AI may improve the odds and shorten the early phases; it has not repealed the laws of biology, and the expensive, slow, failure-prone clinical phases are precisely where AI has so far had the least proven impact.

6.4 Regulators respond

In 2025 the FDA issued draft guidance on the use of AI to support regulatory decision-making for drugs and biological products, proposing a risk-based, credibility-focused framework that emphasises transparency and validation of the models used in submissions.[4] The emergence of formal regulatory expectations is itself a marker of maturity: AI in drug development has moved from a competitive curiosity to a function that regulators feel compelled to govern, and the central question is shifting from whether AI can generate candidates to whether the evidence behind AI-derived claims can be trusted.

7. Precision Medicine, Genomics, and Oncology

7.1 The data-rich frontier

Oncology is where the threads of this review converge, because cancer care generates exactly the heterogeneous, high-dimensional data that multimodal AI is built to integrate: radiological images, digitised pathology, genomic and transcriptomic profiles, and longitudinal clinical records. The ambition is a system that synthesises all of these into a single, coherent picture of an individual tumour and its likely response to therapy.[46]

7.2 AI in genomics

Genomics is among the most natural homes for machine learning, because the central tasks are pattern-recognition

problems at enormous scale. Deep-learning models now assist in variant calling (distinguishing true genetic variants from sequencing artefacts), in predicting the functional consequences of variants, and in interpreting the clinical significance of the many variants of uncertain significance that confront clinical geneticists. The same protein-structure prediction that revolutionised drug discovery also helps interpret how a given mutation might disrupt a protein's function. The promise is to compress the laborious, expert-intensive work of genome interpretation and to surface clinically actionable findings that manual review would miss; the risk is that models trained predominantly on populations of European ancestry may interpret variants less accurately in under-represented populations, importing a new axis of genomic inequity.

7.3 Multimodal oncology in practice

Progress is tangible. Computational-pathology foundation models can now predict molecular alterations and prognosis from routine slides, as discussed in Section 8.[19,25] Commercial precision-oncology platforms process molecular data to match patients to targeted therapies far faster than manual review allows, and multimodal early-detection efforts combine imaging, blood-based biomarkers, and genomic signatures to catch cancers earlier.[46] A 2025 review in The Lancet Digital Health documented the breadth of AI applications across cancer pathology while emphasising the recurring prerequisites for clinical translation: well-curated validation datasets, regulatory approval, and viable reimbursement.[25]

7.4 The methodological warning

The cautionary note here is methodological. Multimodal models are powerful precisely because they exploit subtle correlations across data types- but subtle correlations are also where confounding and dataset-specific artefacts hide. A model that appears to predict outcomes from a slide may in fact be detecting the signature of a particular scanner, staining protocol, or institution rather than any biological property of the tumour. Such "shortcut learning" can produce spectacular in-sample performance that collapses on external data. Rigorous, multi-site, prospective validation is therefore even more essential in the multimodal setting than in the single-modality one, and the burden of proof rises with the complexity of the model.

8. Digital Pathology and Computational Diagnosis

Pathology is undergoing the same digital-then-AI transition that radiology underwent a decade earlier, and it is doing so directly in the foundation-model era. The digitisation of glass slides into whole-slide images created, almost overnight, a vast new data modality ideally suited to deep learning. In 2024 researchers introduced Virchow, then the largest computational-pathology foundation model, trained on a vast corpus of slides and capable of pan-cancer detection across nine common and seven rare cancers with a specimen-level area under the curve of 0.95.[19] A separate pathology foundation model published in Nature the same year demonstrated cancer diagnosis and prognosis prediction across multiple tissue types.[20]

Most striking is the capacity of these models to infer molecular features directly from morphology. AI classifiers built on a digital-pathology foundation model identified actionable genomic alterations in non-small-cell lung cancer- EGFR, ALK, BRAF, and MET- directly from routine haematoxylin-and-eosin slides, achieving areas under the curve from 0.83 to 0.96 on an independent set of 968 samples.[25] If validated and deployed responsibly, such tools could provide a rapid, inexpensive molecular pre-screen from the same slide a pathologist already examines, accelerating the path to targeted therapy in settings where full molecular testing is slow or unavailable.

Pathology also illustrates the workforce logic of medical AI especially clearly. Many regions face a severe and worsening shortage of pathologists, and the volume of slides per pathologist is rising. AI that triages, pre-screens, quantifies, and flags can extend the reach of a scarce specialty without removing the pathologist from the diagnostic loop- the recurring pattern of augmentation rather than replacement. As in radiology, the limiting factor is not whether the models work in retrospective studies but whether they are validated prospectively, integrated into laboratory workflow, and reimbursed sustainably.

9. Surgery and Robotics

9.1 What robotic surgery is- and is not

Robotic surgery is widely misunderstood. In the dominant systems- led by Intuitive Surgical's da Vinci platform, alongside more than sixty other robot-assisted minimally invasive surgery systems- the robot does not perform surgery autonomously. The surgeon operates at a console, and the surgeon's hand movements are translated into precise, tremor-filtered, scaled motions of the instruments.[37] The value lies in dexterity, visualisation, and ergonomics, not in autonomy. AI is now being integrated into these systems incrementally- for enhanced visualisation, instrument tracking, surgical-step recognition, and real-time guidance- moving the technology gradually toward a more active role in the decision-making loop.[36]

9.2 The frontier of autonomy

The research frontier is genuine autonomy, and it is advancing faster than many surgeons expected. In a widely noted 2025 demonstration, a hierarchical, language-conditioned imitation-learning system learned from videos of gallbladder surgery and then autonomously performed the procedure on porcine specimens, responding to and learning from voice feedback much as a trainee would from a mentor.[38] Reviews in Science Robotics now seriously discuss the technological path toward autonomous surgical robots intended to reduce the wide, surgeon-dependent variation in outcomes that characterises complex procedures.[36] These are early, tightly controlled, non-human demonstrations- a pig gallbladder is not a patient- but they mark a meaningful shift from teleoperation toward systems that can execute defined surgical sub-tasks.

9.3 Why caution is warranted

Surgery raises the stakes of every concern in this review. An autonomous system that misidentifies an anatomical structure, fails to recognise an anomaly, or mishandles unexpected bleeding can cause irreversible harm in seconds, with no opportunity for the deliberate human review that mitigates risk in, say, documentation. The questions of accountability, validation, and explainability are therefore at their most acute here. The realistic near-term trajectory is not the autonomous surgeon but the progressively more capable assistant: AI that improves preoperative planning, enhances intraoperative perception, automates well-defined sub-tasks under supervision, and provides real-time decision support- always within a framework where a human surgeon remains responsible and able to intervene.

10. Mental and Behavioural Health

10.1 A field of acute need and acute risk

Mental health presents an unusual combination: enormous unmet need, a severe shortage of clinicians, and a therapeutic modality- conversation- that maps directly onto the capabilities of large language models. Hundreds of millions of people worldwide experience mental disorders, and the gap between need and available care is one of the largest in all of medicine. Conversational AI promises scalable, low-cost, always-available support. It also presents distinctive dangers, because the users are by definition vulnerable, the failure modes can be lethal, and the line between a supportive tool and a harmful one can be a single poorly chosen response.

10.2 The first randomised evidence

The most important development in this area is the first randomised controlled trial of a purpose-built generative-AI therapy chatbot. In a 2025 study published in NEJM AI, researchers at Dartmouth tested Therabot- a model fine-tuned on cognitive-behavioural-therapy and psychotherapy principles rather than a generic chatbot- in 210 adults with major depressive disorder, generalized anxiety disorder, or clinically high risk for an eating disorder.[39] Compared with a waitlist control, four weeks of access to Therabot produced statistically significant symptom reductions across all three groups, sustained at the eight-week follow-up, with high engagement (participants sent an average of about 260 messages) and high satisfaction.[39]

Crucially, the same study reported its safety events with candour: expressions of suicidal ideation required staff intervention fifteen times, and inappropriate responses — such as the model offering medical advice it should not have- required correction thirteen times.[39] These numbers are the point, not a footnote: they show both that a fine-tuned model can deliver measurable therapeutic benefit and that human oversight remained necessary to manage real risk. The authors themselves framed the result as feasibility evidence requiring larger and longer studies, not as a licence for unsupervised deployment.

10.3 Privacy as a first-order concern

Mental-health data are among the most sensitive a person possesses, and deploying language models on them squarely engages the privacy techniques introduced in Section 2. Research has begun to combine federated learning with differential privacy to fine-tune mental-health models without centralising the underlying conversations, explicitly targeting compliance with HIPAA and GDPR.[40] The mental-health-chatbot market- valued near US$1 billion in 2022 and projected to grow several-fold- will test whether commercial pressure to scale can be reconciled with the confidentiality, safety, and clinical rigour that vulnerable users deserve.[40]

11. Remote Monitoring, Wearables, and Consumer Health

11.1 Medicine leaves the building

AI has extended the reach of monitoring far beyond the hospital. Consumer wearables now run on-device algorithms that detect arrhythmias, estimate physiological parameters, and flag anomalies continuously and unobtrusively. The defining study remains the Apple Heart Study, conducted with Stanford Medicine, which enrolled 419,297 participants in just eight months- the largest study of its kind.[18]

Its results are instructive in both directions. Only 0.52% of participants received an irregular-pulse notification, allaying fears of mass over-notification; among those who returned an ECG patch, atrial fibrillation was confirmed in 34%, and the positive predictive value of a notification accompanied by a subsequent irregular reading was 0.84.[18] Atrial fibrillation is intermittent, so a confirmation rate of one-third is unsurprising rather than disappointing. The study demonstrated that a consumer device could safely surface a clinically meaningful, often-silent arrhythmia at population scale- a genuine expansion of preventive cardiology, and a template for the virtual, app-based clinical trial.

11.2 The double edge of consumer health AI

The same capabilities raise hard questions. A notification is not a diagnosis; it initiates a cascade of consultations, monitoring, and anxiety, some of which will prove unnecessary. The boundary between consumer wellness device and regulated medical device blurs, and with it the question of who is responsible when an algorithm worn by tens of millions of people is wrong. Newer features- atrial-fibrillation history estimation, sleep-apnoea and hypertension notifications, and the integration of continuous glucose monitoring with AI forecasting- push further into clinical territory.[43] The market is expanding rapidly: smart wearable ECG monitors alone were valued near US$2 billion in 2024 with projected double-digit annual growth.[44] That growth only sharpens the need for clear standards on evidence, equity of access, and the management of false positives at scale- false positives that, multiplied across a population, can overwhelm the very clinicians the technology was meant to support.

11.3 From monitoring to digital biomarkers

The deeper significance of wearables is not any single alert but the continuous, longitudinal data stream they generate. Where a clinic visit captures a single snapshot, a wearable captures a trajectory- heart-rate variability, activity, sleep architecture, gait, and increasingly continuous glucose- from which machine learning can extract digital biomarkers: computable signals that track the onset or progression of disease. Subtle changes in typing dynamics, voice, or movement are being investigated as early signatures of neurological and psychiatric conditions, and continuous physiological monitoring is enabling the decentralised clinical trial, in which participants are followed remotely at a scale and granularity that site-based trials cannot match, as the Apple Heart Study demonstrated. The promise is a shift from episodic, reactive medicine toward continuous, anticipatory care. The peril is the same continuous stream's implications for privacy and for the medicalisation of ordinary life, and the unresolved question of who owns, interprets, and is accountable for a torrent of data that no clinician has time to read.

12. Administrative and Operational AI

The least glamorous applications of healthcare AI may be among the most economically significant. A large share of healthcare spending is administrative rather than clinical, and much of that administrative work — coding, billing, prior authorisation, scheduling, documentation, supply-chain management, and claims processing — consists of high-volume, rule-bound tasks well suited to automation. AI that schedules operating theatres efficiently, predicts no-shows, optimises staffing, automates medical coding, or streamlines the revenue cycle can free resources for care without touching a clinical decision, and with a lower risk profile than diagnostic or therapeutic AI.

But operational AI is not risk-free, and one application has become a cautionary example: the use of algorithms by insurers to make or support coverage decisions. Reporting and litigation in the United States have alleged that automated tools were used to deny care- for example, to override clinician judgement on the length of post-acute care- prompting class-action lawsuits and congressional scrutiny.[45] The episode is a reminder that "administrative" AI can have profoundly clinical consequences when it sits between a patient and the care they need, and that the governance principles developed for clinical AI- transparency, accountability, the right to human review- apply with equal force to the algorithms that decide what will be paid for.

The reader evaluating operational AI should therefore apply a simple test: does the system merely make an existing administrative process faster and cheaper, or does it make or shape a decision that affects a patient's access to care? The former is low-risk efficiency; the latter is a clinical decision in administrative clothing and demands clinical-grade governance.

13. Algorithmic Bias and Health Equity

13.1 How a fair-seeming model becomes unfair

The most important single study in the ethics of medical AI is Obermeyer and colleagues' 2019 analysis in Science of a commercial risk-prediction algorithm applied to roughly 200 million people in the United States.[15] The algorithm was designed to identify patients with complex needs for extra care. It contained no information about race. Yet it was profoundly biased: at any given risk score, Black patients were considerably sicker than white patients. The mechanism is a lesson in how technical choices encode social inequity. The model predicted healthcare *costs* as a proxy for healthcare *needs*. Because less money has historically been spent on Black patients with equivalent illness- a product of unequal access- the model concluded, accurately as to cost but catastrophically as to need, that those patients were healthier.

The consequences were not abstract. The researchers estimated that correcting the bias would more than double the proportion of Black patients automatically identified for additional help, from 17.7% to 46.5%.[15] Crucially, the bias was remediable: reformulating the target away from cost and toward direct measures of illness substantially eliminated it. The episode established what is now a foundational principle- that the choice of prediction target, the label the model is trained to predict, can be a decisive and hidden source of bias even when no protected attribute appears anywhere in the data. The problem was not in the algorithm's mathematics but in the question it was asked to answer.

13.2 The many faces of bias

- **Data bias:** training cohorts that under-represent populations- by race, sex, age, geography, ancestry, or device type- yield models that perform worse for the under-represented, a particular danger for imaging and genomics models built at a handful of academic centres.
- **Label bias:** the Obermeyer case- a flawed proxy standing in for the true quantity of interest.
- **Deployment bias:** a model accurate in development drifts as it meets populations, practices, and equipment unlike those it was trained on.
- **Feedback loops:** a biased model influences clinical decisions, which generate new biased data, which retrains the model - entrenching and amplifying the original error over time.

13.3 Equity as a design requirement

The corrective is not to remove demographic data- "fairness through unawareness" failed precisely in the Obermeyer case- but to measure performance explicitly across subgroups, scrutinise the choice of prediction target, demand representative training and validation data, and monitor for disparate impact after deployment. Health equity cannot be retrofitted; it must be a design requirement from the first line of the problem statement. Where AI is built carelessly it will scale existing injustice with mechanical efficiency. Where it is built carefully it can, as the Obermeyer team showed by fixing their own example, actively reduce disparity. The technology is not inherently equitable or inequitable; it

faithfully reflects the values and the rigour of those who build it.

14. Evidence, Validation, and Reporting Standards

A theme has recurred in every preceding section: the distance between what models can do in a paper and what they do in practice. It deserves direct treatment, because it is the single greatest threat to the credibility of medical AI.

14.1 Internal validity is not external validity

A model that achieves excellent metrics on data drawn from the same source as its training data has shown *internal* validity- it has learned that distribution. The clinically relevant question is *external* validity: does it work in a different hospital, a different population, on different equipment, in a different year? The Epic Sepsis Model showed in the starkest terms that the answer can be no, and that the gap may go undetected for years because the model is never independently tested.[12,14] The radiology-approval review reinforced the point at the level of the whole field: prospective and clinical testing remain the exception, not the rule, even among cleared devices.[1]

14.2 Discrimination, calibration, and clinical utility

Three distinct questions are too often collapsed into one. **Discrimination** — can the model rank patients correctly?- is what most papers report. **Calibration** — do its probabilities mean what they claim? — is reported far less often, yet a poorly calibrated model misleads clinicians even when it ranks well. **Clinical utility** — does using the model improve patient outcomes?- is the question that actually matters and is answered least often of all, because it requires prospective study of the model embedded in care, not retrospective analysis of its predictions. A field that publishes discrimination metrics and deploys on that basis is building on the shallowest of the three foundations.

14.3 Performance drift and the maintenance problem

Unlike a drug, a deployed model is not static in its effect: the world around it changes. Patient populations shift, clinical practice evolves, coding conventions change, and new equipment alters the data. A model can therefore degrade silently over time- a phenomenon termed dataset or concept drift. This implies an obligation the field has barely begun to meet: continuous post-deployment monitoring, with mechanisms to detect drift and to recalibrate or retire models that have decayed. Medical AI is not a product to be shipped but a system to be maintained, and the absence of monitoring infrastructure is one of the most serious gaps in current practice.

14.4 Reporting standards: making evidence legible

In response to widespread concerns about poor and inconsistent reporting, the research community has developed a family of AI-specific reporting guidelines that now constitute the de facto standard for credible work. **TRIPOD+AI** (2024) provides a 27-item checklist for studies developing or validating clinical prediction models, superseding the 2015 TRIPOD statement and harmonising guidance across regression and machine-learning methods.[30] **CONSORT-AI** and **SPIRIT-AI** (2020) extend the gold-standard guidelines for randomised-trial reports and protocols to AI interventions.[31,32] **DECIDE-AI** (2022) addresses the early, live clinical evaluation of AI decision-support systems, focusing on usability, human factors, and safety- precisely the dimensions that retrospective studies ignore.[33] **CLAIM** (updated 2024) governs medical-imaging AI, and STARD-AI addresses diagnostic-accuracy studies.[34,35]

Complementing these, software-engineering practices such as **model cards-** short documents that state a model's intended use, training data, performance characteristics, and limitations- and production-readiness checklists are increasingly expected.[45] Taken together, these standards encode the lessons of the field's failures into a shared grammar of trustworthy reporting. A reader assessing any medical-AI claim can reasonably ask which of these guidelines the work followed; their absence is itself informative.

Table 3: The principal reporting guidelines for medical AI. Together they bridge algorithm development (TRIPOD+AI, CLAIM) through early clinical evaluation (DECIDE-AI) to definitive trials (SPIRIT-AI, CONSORT-AI)

Guideline	Scope	What it ensures
TRIPOD+AI (2024)	Prediction-model development & validation	Transparent, complete reporting; supersedes TRIPOD 2015
SPIRIT-AI (2020)	Clinical-trial protocols for AI interventions	Pre-specified design, oversight, safeguards
CONSORT-AI (2020)	Clinical-trial reports for AI interventions	Appraisable trial quality and bias assessment
DECIDE-AI (2022)	Early live evaluation of AI decision support	Usability, human factors, real-world safety
CLAIM (2024) / STARD-AI	Medical-imaging & diagnostic-accuracy AI	Standardised imaging-AI and accuracy reporting

14.5 A typology of failure

It is useful to draw together, in one place, the recurring ways medical-AI systems fail, because the failures are patterned rather than random and because recognising the pattern is the first step to preventing it. The cases examined in this review - the Epic Sepsis Model, the Obermeyer cost-proxy algorithm, hallucinating language models, shortcut-learning pathology classifiers- are not isolated mishaps but instances of a small number of generic failure modes that recur across domains.

Distribution shift is the most common: a model trained on one population, institution, or era meets data drawn from another and silently degrades. **Specification failure-** optimising the wrong objective- produced the Obermeyer result, where predicting cost faithfully meant predicting need badly. **Shortcut learning** occurs when a model latches onto a spurious correlate (a scanner watermark, a staining artefact, the fact that portable chest films are taken on sicker patients) rather than the intended signal. **Automation bias and alert fatigue** are human-side failures: clinicians either over-trust a confident output or learn to ignore a noisy one. And **hallucination** is the signature failure of generative systems-

fluent fabrication presented with unwarranted confidence. Each mode has a corresponding defence, and good practice consists largely in anticipating which modes a given deployment is most exposed to.

Table 4: A typology of recurrent medical-AI failure modes with their characteristic defences. Most real-world incidents are instances of one or more of these patterns rather than novel surprises

Failure mode	Illustrative case	Principal defence
Distribution shift	Epic Sepsis Model across new sites	External, prospective validation; drift monitoring
Specification failure	Cost used as a proxy for need (Obermeyer)	Scrutinise the prediction target; subgroup audits
Shortcut learning	Scanner / staining artefacts in imaging	Multi-site data; ablation and stress testing
Automation bias / alert fatigue	Ignored or over-trusted CDS alerts	Human-factors design; actionable, well-timed alerts
Hallucination	Fabricated values in LLM-drafted notes	Retrieval-grounding; mandatory human review

15. Explainability, Trust, and Human Factors

Many of the most accurate medical models are also the most opaque: deep networks with hundreds of millions or billions of parameters whose reasoning cannot be inspected directly. This creates a tension between performance and interpretability that the field has not resolved, and that becomes more acute as foundation models grow larger.

Why explainability matters. Clinicians are accountable for their decisions and reasonably reluctant to act on a recommendation they cannot interrogate. Explanation supports appropriate trust- helping users recognise when a model is likely right and, equally important, when it is likely wrong. It also supports error detection, regulatory scrutiny, medico-legal defensibility, and the patient's interest in understanding decisions that affect their care.

The limits of current methods. Post-hoc explanation techniques- saliency maps that highlight influential image regions, or feature attributions that rank influential inputs- are useful but can be unstable and misleading, sometimes offering a plausible story that does not reflect the model's actual computation. There is an active debate over whether the better path is to explain complex models after the fact or to prefer inherently interpretable models where the stakes are highest. For the foundation models and LLMs now entering practice, faithful explanation remains an unsolved research problem, and a confident-sounding natural-language "explanation" from an LLM may itself be a hallucination.

Calibrated reliance, not blind trust or blanket distrust. The goal is neither automation that sidelines the clinician nor a veto that wastes the model's value, but *calibrated reliance*: a working relationship in which the human knows the model's domain of competence and its failure modes, and the system is designed to make those boundaries visible. The recurring evidence that LLMs and predictive models fail on atypical or out-of-distribution cases is precisely why the human must remain meaningfully in the loop for the foreseeable future- not as a formality, but as the component that handles the cases the model cannot. Human-factors research, embodied in guidelines such as DECIDE-AI, is therefore not peripheral to medical AI but central to whether it helps or harms.[33]

16. Privacy, Security, and Data Governance

AI is data-hungry, and health data are among the most sensitive any person holds. Three distinct risks must be managed together. The first is privacy: the aggregation of records for training creates opportunities for re-identification, and models can in some circumstances memorise and inadvertently regurgitate training data. The second is security: AI systems are software, and software can be attacked- through data poisoning that corrupts a model during training, through adversarial inputs crafted to fool a deployed model, or through theft of the valuable models and datasets themselves. The third is governance: the question of who may use which data, for what purpose, with what consent, and under what accountability.

The technical responses introduced in Section 2- federated learning and differential privacy- directly address the privacy dimension by allowing models to learn from distributed data without centralising it and by bounding what can be inferred about any individual.[40] These are complemented by legal frameworks: HIPAA in the United States and the GDPR in Europe set the baseline obligations for health data, and the EU AI Act layers additional data-governance requirements onto high-risk medical AI, as discussed in Section 17. But technology and law are necessary, not sufficient. The deeper requirement is trust: patients are more willing to allow their data to advance medical knowledge when they believe it will be handled responsibly and used for their benefit, and every high-profile breach or misuse erodes the social licence on which the entire enterprise depends.

A particular tension deserves emphasis. The same pooling of data that improves model fairness and generalisability- by ensuring that under-represented populations are included- also increases privacy risk. Federated and privacy-preserving methods are valuable precisely because they offer a way to pursue the first benefit without fully incurring the second cost. The governance challenge is to build the institutional and technical infrastructure that lets medicine learn continuously from data while honouring the confidentiality that the doctor–patient relationship has always promised.

17. Regulation, Governance, and Ethics

17.1 The United States: a pathway under strain

The FDA has authorised the great majority of AI devices through the 510(k) pathway, which establishes substantial equivalence to an existing predicate device.[1] The pathway is efficient but poorly suited to systems that learn and change: a model is not equivalent to a scalpel. The agency has been developing frameworks for adaptive algorithms- notably the concept of **predetermined change-control plans** that allow a model to be updated within agreed bounds without a new submission for every change- and in early 2025 issued draft guidance for AI-enabled devices addressing transparency,

bias, and a total-product-lifecycle approach to oversight.[36] The FDA has also signalled it will explore tagging devices that incorporate foundation models and LLMs, an acknowledgement that the existing categories strain to contain the new technology.[1]

17.2 The European Union: comprehensive and risk-based

The EU Artificial Intelligence Act (Regulation 2024/1689), which entered into force in August 2024, is the first comprehensive legal framework for AI and is explicitly risk-based.[22,41] AI used for medical purposes generally qualifies as "high-risk," triggering obligations on risk management, data governance, transparency, and human oversight that layer on top of the existing Medical Device and In Vitro Diagnostic Regulations.[41,48] Obligations phase in over several years- prohibitions and AI-literacy duties from early 2025, general-purpose-AI rules from August 2025, core high-risk obligations from August 2026, with an extended transition to August 2027 for AI already regulated as a medical device.[42] Its extraterritorial reach means non-EU developers are bound if their systems affect people in the EU.

17.3 Global governance and the WHO

In January 2024 the World Health Organization issued ethics-and-governance guidance specifically for large multimodal models in health, building on its 2021 framework and offering more than forty recommendations spanning governments, developers, and providers- a recognition that the most general models pose distinctive risks and require coordinated, anticipatory governance rather than reactive regulation.[21]

17.4 The ethical core

Beneath the regulatory architecture lie enduring ethical questions. **Accountability:** when an AI contributes to a harmful decision, responsibility is distributed among developer, deploying institution, and clinician in ways the law is still working out. **Privacy:** AI is data-hungry, and the aggregation of sensitive health data for training creates risks that consent frameworks struggle to address. **Autonomy:** patients have an interest in knowing when AI is involved in their care and in not being reduced to the output of a model. These are not obstacles to be cleared on the way to deployment; they are the terms on which deployment earns legitimacy.

17.5 Liability and the accountability gap

Among these questions, liability is the one most likely to shape adoption in practice, because it determines who bears the cost when an AI-influenced decision goes wrong. The current allocation is unsettled and arguably incoherent. A clinician who follows a flawed AI recommendation may be liable for not overriding it; a clinician who overrides a correct one may be liable for ignoring available information- a bind that discourages exactly the calibrated judgement the technology requires. Vendors limit their exposure through the framing of their products as decision support rather than decision makers, shifting final responsibility to the clinician even as the tool shapes the decision. Meanwhile the insurance-coverage litigation noted in Section 12 shows that when algorithms make adverse determinations, accountability can become diffuse to the point of disappearing. Resolving this accountability gap- through clearer standards of care, transparent allocation of responsibility across the developer–deployer–user chain, and mechanisms for redress- is not a peripheral legal detail but a precondition for the trust on which clinical adoption ultimately rests.

Table 4: The principal regulatory and governance frameworks shaping health AI as of 2026. They differ sharply in comprehensiveness but converge on risk-proportionality, transparency, and human oversight.

Framework	Jurisdiction / Body	Defining feature for health AI
510(k) / De Novo pathways	US FDA	Market entry via substantial equivalence; ~97% of AI devices use 510(k)
AI Act (2024/1689)	European Union	Risk-based; medical AI generally high-risk with phased duties to 2027
MDR / IVDR	European Union	Underlying device regulation; third-party conformity assessment
LMM ethics guidance (2024)	World Health Organization	40+ recommendations specific to large multimodal models
AI in drug development (draft, 2025)	US FDA	Risk-based credibility framework for models in submissions

18. 18. Economics, Workforce, and Health Systems

18.1 The investment case

The economic logic of healthcare AI is rooted in two pressures every health system faces: rising demand from ageing populations and chronic disease, and a workforce that cannot scale to meet it. AI is attractive precisely because it promises to extend scarce expertise- a single screening algorithm can read images at a volume no individual could match, and an ambient scribe can return hours of clinician time to patient care. Market forecasts reflect this logic, projecting the healthcare-AI market toward roughly US$190 billion by 2030 at compound growth near 37–39%, with diagnostics, decision support, and robot-assisted surgery among the leading segments.[3,43]

These projections should be treated as directional rather than precise; market-sizing reports vary widely in methodology and incentive, and a healthy scepticism toward any single forecast is warranted. What is robust is the direction and the underlying driver: the value of AI in health is greatest where it relieves a binding human-capacity constraint, and weakest where it merely adds another layer of technology to an already-functioning process. Return on investment in healthcare AI has, in practice, proven harder to realise than vendors promise, because value depends on workflow redesign, change management, and sustained maintenance, not on the algorithm alone.

18.2 Will AI replace clinicians?

The recurring public question is whether AI will replace doctors. The evidence of this review points to a more accurate framing. AI excels at narrow, well-bounded tasks with abundant data and clear ground truth; it remains weak at the integrative, contextual, relational, and ethical dimensions of care- the messy, atypical, and human work that constitutes much of medicine. The realistic trajectory is augmentation and task-shifting: AI absorbs the repetitive and the routine, freeing clinicians for judgement, communication, and the cases that defy pattern-matching. The often-quoted formulation- that AI will not replace clinicians, but clinicians who use AI may replace those who do not- captures the near-term reality. The clinician most exposed is not the one replaced by AI but the one who declines to use it while a colleague does.

18.3 Access, equity, and the global picture

AI could narrow or widen global health inequities depending on how it is deployed. Autonomous screening of the IDx-DR type could bring specialist-level diagnosis to settings with no specialists- a profound potential gain for low- and middle-income regions.[16] Yet the same regions often lack the digital infrastructure, local validation data, and regulatory capacity that safe deployment requires, and models trained on wealthy-country data may transfer poorly. Realising the equity upside requires deliberate investment in local data, infrastructure, and governance- it will not happen by default, and the path of least resistance leads toward concentration of benefit in the systems that least need it.

18.4 Why the returns so often disappoint

A persistent gap separates the value AI promises in a pilot from the value it delivers at scale, and the reasons are instructive. Many deployments stall not because the model underperforms but because the surrounding work- integration with the electronic health record, clinician training, workflow redesign, ongoing monitoring, and maintenance- is underestimated and underfunded. A model is a small component of a large socio-technical system, and the unglamorous integration work typically costs far more than the algorithm itself. Pilots succeed in the controlled conditions of motivated early adopters and dedicated support; scale-up exposes the brittleness of tools that were never engineered for the messiness of routine practice. The lesson, repeated across health systems, is that procuring a model is the beginning of the cost, not the end of it, and that organisations which budget only for the technology and not for its assimilation are systematically disappointed.

18.5 AI literacy and the education gap

The safe use of medical AI depends on a workforce that understands it well enough to use it critically- to know when to trust an output, when to question it, and how its failure modes manifest. Yet medical and nursing curricula have only begun to incorporate the statistical and computational literacy this requires, and most practising clinicians received no formal training in it at all. The risk is a workforce that either over-trusts opaque tools through unfamiliarity or rejects valuable ones through misplaced suspicion. Building genuine AI literacy- not the ability to build models, but the ability to appraise and supervise them- is as much a precondition for safe deployment as any technical safeguard, and notably the EU AI Act now makes AI literacy a formal obligation for organisations deploying these systems. The clinician of the coming decade will need to be a competent supervisor of machines, and the education system has not yet caught up to that reality.

18.6 The environmental and infrastructure cost

Finally, the computational substrate of modern AI carries a real and often-ignored cost. Training and running large foundation models consumes substantial energy and water and depends on specialised hardware concentrated in a few data centres and supply chains. For a health system weighing the adoption of large-model tools, this implies both a direct operating cost and an environmental footprint that sits awkwardly with healthcare's own sustainability commitments. It also implies a dependency: institutions that build critical clinical workflows on externally hosted models inherit the reliability, pricing, and continuity risks of their providers. None of this argues against the technology, but it belongs in an honest accounting of what AI in medicine costs- not only in dollars per licence, but in energy, infrastructure, and strategic dependence.

19. Synthesis and the Path Forward

The evidence assembled here supports the thesis stated at the outset: the limiting factor in medical AI is no longer capability but accountability. The models can read the scan, draft the note, fold the protein, pass the examination, and even reduce depressive symptoms in a randomised trial. What remains underbuilt is the surrounding system- the validation, governance, equity safeguards, and human–machine collaboration- that turns capability into reliable benefit.

Two patterns run through every domain examined. First, the gap between benchmark and bedside is real and recurrent: examination scores, retrospective metrics, and vendor claims systematically overstate real-world clinical value, and only prospective, external, outcome-focused evaluation closes the gap. Second, AI in healthcare is socio-technical: the Epic Sepsis Model failed not because its mathematics were uniquely poor but because it was deployed without validation into a human workflow it did not fit, and the Obermeyer algorithm harmed not through malice but through an unexamined choice of prediction target. The failures are organisational and conceptual as much as computational, which is encouraging- because organisational and conceptual problems are ones we know how to solve.

19.1 Recommendations

- **Validate prospectively and locally.** No high-stakes model should be deployed on the strength of internal metrics alone; external, prospective validation in the target population is a precondition, not an enhancement.
- **Report calibration and utility, not only discrimination.** Journals, regulators, and procurers should require evidence that probabilities are trustworthy and that using

the model improves outcomes — and should expect adherence to TRIPOD+AI, CONSORT-AI, and related standards.
- **Treat equity as a design requirement.** Subgroup performance, prediction-target scrutiny, and representative data must be built in from the start and monitored after launch.
- **Monitor continuously.** Deployed models must be surveilled for drift and maintained, recalibrated, or retired- governed as living systems, not shipped products.
- **Keep the human meaningfully in the loop.** Design for calibrated reliance, making each model's competence and failure modes visible to its users.
- **Match regulation to the technology.** Adaptive, learning systems and foundation models need governance built for change, such as predetermined change-control plans, not frameworks retrofitted from static devices.
- **Protect privacy by design.** Where data must be pooled to achieve fairness and generalisability, federated and privacy-preserving methods should be the default, not an afterthought.

19.2 Limitations of this review and directions ahead

This review is a narrative synthesis, not a systematic one; it is selective by design, privileging landmark studies and illustrative cases over exhaustive enumeration, and it reflects the state of evidence available in early 2026 in a field that moves monthly. Several developments are likely to define the next phase and deserve close watching. Multimodal foundation models that reason jointly over imaging, text, genomics, and waveforms will move from research to clinical pilots, raising the validation bar described in Section 7. Agentic systems- models that take sequences of actions rather than producing a single output- will extend AI from advising clinicians to executing multi-step workflows, sharpening every question of accountability and oversight. Regulation will mature, with the EU AI Act's high-risk obligations and the FDA's lifecycle approach moving from text to enforced practice. And the evidence base itself should strengthen as prospective trials, registries, and post-deployment monitoring- long the field's weakest link- finally begin to accumulate. The central uncertainty is not whether the models will become more capable; they will. It is whether the systems of evidence and governance around them will mature quickly enough to keep pace.

19.3 A closing perspective

Artificial intelligence is reshaping modern healthcare- the question posed by this article's title is settled in the affirmative. But the shape of that transformation is not yet determined. The technology is sufficiently powerful to do great good and, deployed carelessly, sufficiently powerful to scale harm and inequity with the same efficiency. The decisive variables are human: the rigour of our evidence, the wisdom of our governance, and the seriousness with which we treat fairness and safety not as compliance exercises but as the substance of good medicine. AI in healthcare will ultimately be judged not by how intelligent its models are, but by how accountable its makers and users choose to be. On present evidence, that choice is still ours to make.

Declarations

Conflicts of interest. The author declares no financial conflicts of interest relevant to this manuscript. **Funding.** This work received no external funding. **Data availability.** All sources are publicly available and cited in the reference list.

Appendix A. A Timeline of Medical AI

The following timeline situates the developments discussed in this article. Dates refer to the landmark event (publication, authorisation, or release) most commonly associated with each entry.

Table A1: A condensed timeline of the milestones discussed in this review. Entries are ordered to foreground the acceleration of the field after 2016 and again after 2022.

Year	Development
1972	MYCIN rule-based expert system for antimicrobial therapy demonstrates early clinical AI
1998	First FDA authorisation of an AI imaging tool (mammography computer-aided detection)
2017	Transformer architecture introduced, enabling the foundation-model era
2016	Deep-learning algorithm matches ophthalmologists for diabetic-retinopathy detection (AUC > 0.99)
2018	IDx-DR becomes the first autonomous diagnostic AI authorised in any field; Viz.ai clears stroke triage; Apple Heart Study launches
2019	Obermeyer et al. expose racial bias in a population-health algorithm affecting ~200 million people
2020	CONSORT-AI and SPIRIT-AI reporting guidelines published
2021	AlphaFold solves protein structure prediction; Epic Sepsis Model external validation reveals poor real-world performance
2022	DECIDE-AI guideline published; ChatGPT popularises large language models
2023	Med-PaLM 2 reaches ~86.5% on USMLE-style questions, approaching expert level
2024	AlphaFold 3 extends prediction to molecular interactions (Nobel Prize in Chemistry); Virchow pathology foundation model; TRIPOD+AI and updated CLAIM published; EU AI Act enters into force; WHO issues LMM guidance
2025	First RCT of a generative-AI therapy chatbot (Therabot); autonomous robotic gallbladder surgery on porcine models; FDA draft guidance on AI in drug development and AI-enabled devices; AI radiology authorisations surpass 1,000

Appendix B. Glossary of Key Terms

Table B1: Definitions of the key technical terms used in this review, intended for the non-specialist reader

Term	Definition
Machine learning (ML)	Building models whose parameters are tuned by data so they perform a task without explicit programming for every case.
Deep learning	ML using multi-layer neural networks that learn hierarchical representations directly from raw data.
Convolutional neural network (CNN)	A neural-network architecture specialised for images, central to medical imaging AI.
Transformer	A neural architecture using attention; the basis of modern foundation models and LLMs.

Foundation model	A large, general-purpose model pre-trained on broad data and adaptable to many downstream tasks.
Large language model (LLM)	A foundation model that processes and generates natural-language text.
Hallucination	Fluent but false content generated by a model; a core safety risk of LLMs in medicine.
Discrimination	A model's ability to rank patients correctly (e.g., area under the ROC curve).
Calibration	Whether a model's predicted probabilities match observed frequencies.
External validation	Testing a model on data from a different setting than it was trained on.
Dataset / concept drift	Degradation of a deployed model as the data or world it operates in change over time.
Federated learning	Training across institutions by sharing model parameters rather than raw patient data.
Differential privacy	A formal technique that bounds what can be learned about any individual from shared data.
510(k) pathway	US FDA route to market via demonstrating substantial equivalence to an existing device.

References

[1] Sivakumar R, Lue B, Kundu S. FDA approval of artificial intelligence and machine learning devices in radiology: a systematic review. *JAMA Network Open. 2025;8(11). DOI:10.1001/jamanetworkopen.2025.42338.*

[2] The Imaging Wire. FDA AI approvals surge past 1,000 for radiology (authorisations through September 2025). *theimagingwire.com, December 2025.*

[3] The Research Insights / Precedence Research. AI in healthcare market: size and forecast to 2030 (US$26.6B in 2024 to ~US$187.7B by 2030, CAGR ~38%). *Market reports, 2025–2026.*

[4] Corona A. How AI is transforming drug discovery; FDA 2025 draft guidance on AI in drug and biological-product development. *Drug Discovery News. 2025.*

[5] From lab to clinic: how artificial intelligence is reshaping drug-discovery timelines (PRISMA systematic review, 2015–2025). *PMC12298131. 2025.*

[6] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature. 2021;596(7873):583–589. DOI:10.1038/s41586-021-03819-2.*

[7] Fang Z, Ran H, Zhang Y, et al. AlphaFold 3: an unprecedented opportunity for fundamental research and drug development. *Precision Clinical Medicine. 2025;8(3): pbaf015. DOI:10.1093/pcmedi/pbaf015.*

[8] Isomorphic Labs prepares to launch trials for AI-designed drugs; US$600M financing round (2025). *Clinical Trials Arena / Fortune, 2025.*

[9] Scoping review: the use of large language models in clinical documentation. *International Journal of Nursing Studies / ScienceDirect S0020748925003323. 2025.*

[10] Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine. 2024; 30: 1134–1142.*

[11] Singhal K, et al. Large language models encode clinical knowledge (Med-PaLM); Med-PaLM 2 USMLE-style performance 86.5%. *Nature. 2023; and Google Research, Med-PaLM.*

[12] Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine. 2021;181(8):1065–1070. DOI:10.1001/jamainternmed.2021.2626.*

[13] Habib AR, Lin AL, Grant RW. The Epic Sepsis Model falls short — the importance of external validation (incl. sepsis epidemiology). *JAMA Internal Medicine. 2021;181(8):1040–1041. DOI:10.1001/jamainternmed.2021.3333.*

[14] Ostermayer DG, Braunheim B, Mehta AM, et al. External validation of the Epic sepsis predictive model in two county emergency departments. *JAMIA Open. 2024. DOI:10.1093/jamiaopen/ooae133.*

[15] 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):447–453. DOI:10.1126/science.aax2342.*

[16] Abràmoff MD, Lavin PT, Birch M, et al. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary-care offices. *npj Digital Medicine. 2018; 1: 39.*

[17] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep-learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA. 2016;316(22):2402–2410.*

[18] Perez MV, Mahaffey KW, Hedlin H, et al. Large-scale assessment of a smartwatch to identify atrial fibrillation (Apple Heart Study). *New England Journal of Medicine. 2019; 381: 1909–1917. DOI:10.1056/NEJMoa1901183.*

[19] Vorontsov E, et al. A foundation model for clinical-grade computational pathology and rare-cancer detection (Virchow). *Nature Medicine. 2024; 30: 2924–2935. DOI:10.1038/s41591-024-03141-0.*

[20] Wang X, Zhao J, Marostica E, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature. 2024; 634: 970–978.*

[21] World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. *WHO, January 2024.*

[22] Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act). *Official Journal of the European Union, 2024.*

[23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS). 2017:5998–6008.*

[24] Large language models in real-world clinical workflows: a systematic review of applications and implementation. *Frontiers in Digital Health. 2025. DOI:10.3389/fdgth.2025.1659134.*

[25] Application of AI and digital tools in cancer pathology (review); CanvOI NSCLC biomarker validation (968 samples). *The Lancet Digital Health, 2025; npj Precision Oncology, 2026 (DOI:10.1038/s41698-025-01267-z).*

[26] Li H, Fu J-F, Python A. Implementing large language models in health care: clinician-focused review with interactive guideline. *Journal of Medical Internet Research. 2025;e71916.*

[27] Autonomous artificial intelligence in diabetic-retinopathy testing- lessons on successful health-system

adoption (LumineticsCore, EyeArt, AEYE). *Ophthalmology Science / ScienceDirect, 2025; PMC12553049.*

[28] Autonomous AI in diabetic retinopathy: from algorithm to clinical application (development cost and validation timeline). *PMC8120059, 2021.*

[29] AI in radiology: 2025 trends, FDA approvals, and adoption (Viz.ai deployment across 1,600+ hospitals). *IntuitionLabs, November 2025.*

[30] Collins GS, Moons KGM, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models. *BMJ. 2024. (27-item checklist; supersedes TRIPOD 2015.)*

[31] Cruz Rivera S, Liu X, Denniston AK, et al. Guidelines for clinical-trial protocols for AI interventions: the SPIRIT-AI extension. *Nature Medicine. 2020; 26: 1351–1363.*

[32] Liu X, Cruz Rivera S, Moher D, et al. Reporting guidelines for clinical-trial reports for AI interventions: the CONSORT-AI extension. *Nature Medicine. 2020; 26: 1364–1374.*

[33] Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for early-stage clinical evaluation of AI decision-support systems: DECIDE-AI. *BMJ. 2022;377: e070904.*

[34] Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. *Radiology: Artificial Intelligence. 2024;6: e240300.*

[35] Reporting guidelines for AI studies in healthcare (conventional and LLMs): what's new in 2024 (overview). *Korean Journal of Radiology. 2024.*

[36] Will your next surgeon be a robot? Autonomy and AI in robotic surgery (da Vinci; FDA total-product-lifecycle and predetermined change-control plans). *Science Robotics. 2025. DOI:10.1126/scirobotics.adt0187.*

[37] Banks J. Will AI open the door to fully autonomous robotic surgery? (da Vinci teleoperation; 60+ RAMIS systems). *IEEE Pulse. 2025.*

[38] SRT-H: a hierarchical framework for autonomous surgery via language-conditioned imitation learning (autonomous gallbladder surgery, porcine model). *Science Robotics / Johns Hopkins. 2025.*

[39] Heinz MV, Mackin DM, Trudeau BM, et al. Randomized trial of a generative AI chatbot for mental-health treatment (Therabot). *NEJM AI. 2025;2(4). DOI:10.1056/AIoa2400802. (ClinicalTrials.gov NCT06013137.)*

[40] FedMentalCare / FedMentor: privacy-preserving, federated fine-tuning of LLMs for mental health (differential privacy; HIPAA/GDPR). *arXiv:2503.05786; arXiv:2509.14275, 2025.*

[41] The EU Artificial Intelligence Act (2024): implications for healthcare (high-risk classification of medical AI). *Health Policy / ScienceDirect S0168851024001623. 2024.*

[42] EU AI Act explained: phased obligations and transition periods for medical-device AI to 2027. *Tandem Health, 2026.*

[43] AI in medical devices market expands to ~US$886B by 2034 (Apple AFib history; AI-assisted surgical robotics). *Towards Healthcare / GlobeNewswire, November 2025.*

[44] Smart wearable ECG monitors market (~US$2.0B in 2024; projected ~US$3.5B by 2030). *Grand View Research, 2025.*

[45] AImedReport; model cards and ML Test Score for AI model transparency and production readiness. *PMC11975813, 2025; and original model-card literature.*

[46] Flora DB. The top precision-oncology AI stories of 2024 (Tempus, Sophia Genetics; multimodal early detection). *AI in Precision Oncology. 2025.*

[47] AI surgical planning in 2025: technologies and market trends. *iData Research, 2025.*

[48] The EU AI Act and medical devices: navigating high-risk compliance. *Reed Smith Viewpoints, June 2025.*

[49] Using AI tools in health-insurance coverage decisions: reporting, class-action litigation, and oversight. *Stanford Health Policy / FSI, 2024.*