Online ISSN: 3104-9338
Print ISSN: 3104-932X

From Benchmark to Bedside: Mapping the Translational Gaps of Artificial Intelligence in Clinical Medicine

Bohan Li,Xinyu Song*

Geely University of China, ChengDu, China, 641423

Received: August 23 2026

Revised: August 30, 2026

Accepted: August 30, 2026

Published online: August 30, 2026

To appear in: *International Journal of Advanced AI Applications*, Vol. 2, No. 9 (September 2026)

* Corresponding Author: Xinyu Song (songxinyu@guc.edu.cn)

Abstract. Artificial intelligence (AI) has reached specialist-level performance across a widening range of clinical tasks, yet celebrated benchmark results have repeatedly failed to translate into patient benefit. This survey reviews 29 core sources (2016–2024) through an original framework of three translational gaps—validation, utility, and governance—that jointly explain the distance between benchmark scores and bedside value; reported figures are quoted only from peer-reviewed studies, pivotal trials, or first-party regulatory documents. Capability measurement has matured across three generations—specialist image classifiers, examination-scale language models exceeding 86% on USMLE-style questions, and workflow-level deployments such as autonomous diabetic-retinopathy screening and ambient clinical documentation. Each gap independently severs the link between performance and benefit: deployed models degrade under dataset shift, an externally audited sepsis model fell far below its reported accuracy, nearly half of the citations in generated medical content can be fabricated, and human–AI collaboration can underperform either party alone. Governance responses—risk-tiered regulation, equity audits, and post-market surveillance—are converging on medicine's established logic: grading AI claims with the same evidence hierarchy that evidence-based medicine applies to therapies.

Keywords: *medical imaging; large language models; external validation; algorithmic bias; regulatory approval*

1. Introduction

1.1. Background and the Central Contradiction

Medicine has become the most visible testing ground for artificial intelligence (AI). Deep learning systems now read retinal photographs, chest radiographs, histopathology slides, and dermatological images; large language models (LLMs)—neural networks trained on web-scale text—answer licensing-examination questions, draft clinical documentation, and summarize patient records. Commentators have described this convergence as the arrival of "high-performance medicine," in which machine perception and recall augment clinical judgment across specialties [1]. Health systems have acted on the promise: AI-enabled products are embedded in radiology workflows, emergency departments, and primary-care clinics. The speed of adoption is unusual even by the standards of recent medical technology; the deeper novelty is that medicine, of all professions, already possesses a mature institutional answer to the question of what counts as evidence—and AI systems are arriving faster than that machinery has learned to grade them.

That institutional answer is worth describing at the outset, because this survey's entire argument is that AI evaluation is converging on it from first principles. Evidence-based medicine ranks forms of evidence by their resistance to self-deception: mechanistic reasoning at the bottom, retrospective observation above it, controlled comparison higher still, and pre-registered randomized trials with independent surveillance at the top—paired with the institutions that operate the hierarchy, from ethics boards to trial registries to pharmacovigilance systems [1]. The hierarchy was not designed for software, but its logic is software-agnostic: it asks of any claim not how impressive it is but what would have revealed it to be wrong. A benchmark score is impressive in exactly the way a mechanistic argument is impressive—and, as the sections that follow document, it fails in exactly the way mechanistic arguments fail: persuasively, retrospectively, and without any procedure that would have caught its errors before patients absorbed them. The survey's governance conclusion can therefore be stated here and defended for the rest of the paper: the question is not whether medicine needs new institutions for AI, but whether AI evaluation will adopt the institutions medicine already built, or repeat the centuries it took to construct them.

The promise, however, has coexisted with a persistent pattern of disappointment, and the pattern is not accidental. In 2017, a Nature study reported that a deep neural network classified skin cancer at dermatologist-level performance across 21 disease categories, an achievement widely covered as a turning point for the field [2]. In the same year, an investigative report

documented that IBM's Watson for Oncology—a system deployed to recommend cancer treatment—had generated "unsafe and incorrect" therapy suggestions in hospital settings, contradicted by the physicians meant to use it [3]. The juxtaposition frames this survey's central question: why do headline performances on clinical benchmarks so often fail to become safe, useful, and equitable bedside systems?

This survey argues that the failure is structural rather than incidental, and that it decomposes into three independent translational gaps: the validation gap (benchmark performance does not survive contact with prospective clinical data), the utility gap (technical accuracy is not patient benefit), and the governance gap (a validated system is not automatically an accountable or equitable one). Each gap has its own failure evidence, its own corrective tradition, and its own open problems; conflating them—which much of the public debate does—produces both exaggerated hope and indiscriminate skepticism.

The independence of the gaps is the framework's operative claim, and it cuts both ways. Because they are independent, no amount of progress on one closes another: a model validated prospectively at multiple sites has addressed the first gap and none of the others; an interventional outcome study of an unvalidated model measures a moving object; a regulated deployment of a utility-blind system enrolls regulators in harms it does not prevent. Each of the field's celebrated failures can be located on exactly one gap, which is why single-lever explanations of the benchmark-to-bedside distance keep failing—the deployment pessimist points at validation evidence, the adoption optimist at utility trials, and each dismisses the other's gap as an implementation detail. The gaps also differ in who can close them: the validation gap is closed by evaluators and regulators, the utility gap by trialists and health systems, the governance gap by legislators and professional bodies—an allocation of labor the literature rarely makes explicit but that determines which failures should be assigned to which institutions. The three gaps are, in short, the field's problem structure, and the survey organizes its evidence to show that they are real, measured, and separately actionable.

1.2. Survey Methodology

We conducted a targeted literature search across PubMed-indexed general and specialty journals (Nature Medicine, JAMA, NEJM, The Lancet Digital Health, npj Digital Medicine, PLOS Medicine), AI conference proceedings, and first-party regulatory publications (U.S. FDA, European Union, WHO). The search covered 2016–2024 for technical work, 2016 being the year the modern clinical-deep-learning literature opened with large-scale retinal screening. Candidates were retained if they (i) reported empirical clinical or examination performance of

AI systems, (ii) provided independent or adversarial evaluation of deployed systems, or (iii) constituted primary governance instruments (regulations, guidance, official device lists). Purely speculative commentary was excluded. Reported figures are quoted only from peer-reviewed publications, pivotal trials, or first-party documents—never from secondary commentary. This procedure yielded the 29 core sources organized below.

The discipline behind that last sentence bears explicit description, because surveys of medical AI are unusually exposed to citation drift. Every quantitative claim in this survey was traced to the document that produced it: performance figures to the original benchmark or trial publications, audit findings to the auditing studies, regulatory counts to the regulator's own published list, and legislative obligations to the legislative texts themselves. Author lists and bibliographic details of every source were verified against its primary record before inclusion, and sources that could not be verified with confidence were excluded rather than cited from memory. Where this survey draws an inference the source does not state—that a finding constitutes a validation failure, for instance—the inference is presented as interpretation, not attributed to the source. The survey holds itself to the evidentiary standard it argues the field should hold AI systems to: traceable claims, primary provenance, and no number whose origin cannot be named in one step. Readers who find a figure here can follow it to its document; that is the property Sections 3 and 4 will argue deployed clinical AI currently lacks for its own claims.

1.3. Positioning Against Existing Surveys

The field's self-understanding is framed by several recent reviews. Rajpurkar et al. survey AI in health and medicine by modality—image, text, and multimodal systems—and catalogue technical progress and deployment barriers [4]. Thirunavukarasu et al. review LLMs in medicine specifically, organizing the material by clinical task [5]. Kelly et al. articulate the challenges of delivering clinical impact—data quality, workflow integration, and regulation—providing the closest prior framing, though written before the LLM era and without the quantitative failure evidence that has since accumulated [6]. All three organize the field by what AI is (modalities, tasks) or what obstacles exist (challenge lists); none treats the inferential status of performance claims as the organizing problem. Table 1 contrasts their coverage with this survey's, whose distinctive commitment is to grade every performance claim by its translational status—the logic evidence-based medicine (EBM) already applies to therapies.

Table 1. Coverage comparison between recent surveys of AI in medicine and this work.

Dimension	Rajpurkar et al. [4]	Thirunavukarasu et al. [5]	Kelly et al. [6]	This survey
Primary scope	Multimodal techniques	LLMs by clinical task	Barriers to impact	Benchmark-to-bedside evidence
Organizing logic	Modality	Task category	Challenge list	Three translational gaps
Failure evidence	Selected examples	Brief	Conceptual	Quantified: shift, hallucination, external audit
Governance treatment	Brief	Brief	Conceptual	Regulation, equity, post-market surveillance

The remainder follows the gap structure. Section 2 reviews what has actually been measured, across three generations of capability. Section 3 presents the three gaps and the evidence for each. Section 4 analyzes the governance responses closing them, and Section 5 concludes.

2. Measured Capability: Three Generations of Clinical AI

2.1. The Specialist Era: Deep Learning on Medical Images

The modern capability story begins with specialist image classifiers. Gulshan et al. trained a convolutional network on more than 128,000 retinal fundus photographs and detected referable diabetic retinopathy with an area under the receiver operating characteristic curve (AUC) above 0.99 on two independent validation sets—performance matching ophthalmologists [7]. Esteva et al. classified skin cancer across 21 categories at dermatologist level using 129,450 clinical images [2]. De Fauw et al. translated optical coherence tomography scans into referral recommendations matching expert clinicians on sight-threatening retinal disease, and showed the representation transferred across scanning devices [8]. Chest radiography followed: CheXNet detected pneumonia at radiologist level [9], and the CheXpert dataset—224,316 radiographs of 65,240 patients with automated uncertainty-aware labels—industrialized benchmark construction for the modality [10].

Three properties defined the era. Measurement was single-task: one disease, one image type, one decision. Ground truth was retrospective: labels extracted from existing reports rather than from what happened to patients. And the comparator was the specialist: success meant matching clinicians on their home turf, not improving outcomes. These properties made spectacular headline results possible—and, as Section 3 shows, planted the first translation gap. They also set a template that survives into the present: the field's most-cited capability claims are still, at bottom, agreement statistics computed on archives, however sophisticated the architecture

reading the archive has since become.

The archive structure of the era's landmark studies rewards a second look, because it determined what the studies could and could not show. Gulshan et al.'s retinal classifier was validated on two data sets—one drawn from the same screening network as training data, one from a different network—both curated and retrospective: images already captured, already graded, chosen to measure agreement rather than consequence [7]. Esteva et al.'s dermatology classifier was tested on photographs already verified by biopsy or follow-up, a design that guarantees a clean reference standard and guarantees equally that the test distribution is nothing like a clinic's intake, where prevalence is lower, quality is uncontrolled, and the difficult cases were already referred to specialists [2]. De Fauw et al.'s OCT referral system went furthest of the three toward clinical realism—its cohort was drawn from a real care pathway, and the analysis included simulated deployment scenarios—but its comparison target remained the clinicians' own decisions rather than downstream outcomes [8]. None of these qualifications diminishes the studies; they were measurement milestones executed with unusual care. The point is what their designs share: in every case the patient had already been triaged, imaged, and diagnosed by the ordinary machinery of care before the AI entered the pipeline, and the AI's agreement with that machinery was the entire result. What happens when the AI enters before the machinery—not after it—is the subject of Section 3.

2.2. The Generalist Era: Medical Question Answering

The second generation shifted from images to language. MedQA assembled large-scale question answering from professional medical examinations used in the United States and China, operationalizing clinical knowledge rather than perception [11]. On the MultiMedQA suite built from six such datasets, the instruction-tuned Flan-PaLM reached 67.6% on MedQA—surpassing prior state of the art by more than 17 points—and its instruction-tuned derivative Med-PaLM was the first model to exceed the USMLE passing threshold [12]. Its successor Med-PaLM 2 reached 86.5%, and in blinded comparison physicians preferred its consumer-question answers to those written by physicians themselves on eight of nine evaluated axes [13]. GPT-4, with no medical specialization at all, exceeded the USMLE passing score by over 20 points and was markedly better calibrated than GPT-3.5 [14]; ChatGPT itself performed at or near the passing threshold across all three USMLE steps [15].

Yet the same studies that established these scores qualified them. Human evaluation of long-form answers found Flan-PaLM's outputs repeatedly inferior to clinicians' on factuality and potential harm even where multiple-choice accuracy was state of the art [12]. The measurement

object had quietly migrated: from whether a model answers like a specialist on a curated question, to whether it behaves like a clinician on an open one—and the second claim was not established by the first.

The human-evaluation arm of the Flan-PaLM and Med-PaLM 2 studies is the era's most instructive evidence precisely because it measured the migration directly. Alongside the multiple-choice results, the investigators had clinicians rate long-form answers on nine axes—factuality, potential for harm, comprehension, and related quality dimensions—and the pattern held across both model generations: answers that were state of the art on examination scoring remained repeatedly rated inferior to clinicians' own on the axes that matter to a patient [12]. Med-PaLM 2 closed much of the gap—in blinded comparison its answers were preferred to physician-written ones on eight of nine axes—but the residual axis tells the honest story: the axis on which the model still trailed was potential for harm, the one dimension where an error is not an inconvenience but an injury [13]. An examination cannot register that dimension; a multiple-choice key has no row for it. The generalist era's deepest result is therefore not the 86.5% but the demonstration that examination scores and clinical-answer quality are different measurements that happen to correlate loosely—and that a field reading the first number as evidence about the second was making a categorical error the studies themselves had already refuted in their own appendix tables [12,13]. The era also internationalized evaluation: MedQA drew its questions from professional examinations on both sides of the Pacific, and the cross-country hallucination benchmarks discussed further in Section 3.2 ask whether models trained on predominantly English, Western corpora remain sound when questioned under other jurisdictions' testing standards [11]. Capability, in short, is now measured globally—but mostly with the Anglophone model as protagonist.

2.3. The Workflow Era: From Models to Deployed Systems

The third generation measures systems, not models. IDx-DR became the first FDA-authorized fully autonomous diagnostic AI, referring or clearing diabetic-retinopathy patients in primary-care offices without any physician interpretation, on the strength of a prospective pivotal trial [16]. Ambient AI scribes—LLMs that listen to clinical encounters and draft documentation—represent the fastest LLM deployment to date: one deployment study across a large integrated care system reported 968 physicians with at least 100 encounters each and an estimated 15,000 hours of documentation time saved over 2.5 million uses [17].

The two workflow deployments are also, read carefully, a controlled lesson in how to make translation succeed—because they chose the two easiest cells of the task grid and still illustrate

every remaining gap. Autonomous screening succeeded at IDx-DR only after a pivotal trial designed to regulatory standard, a task with a binary refer/clear decision, a modality with a stable imaging protocol, and a care setting where the consequence of error is another appointment rather than a missed diagnosis [16]. Ambient scribes succeeded first because their output is draft: a clinician reviews and signs every note, so the human-in-the-loop is structurally built in and errors are absorbed by the professional rather than transmitted to the patient [17]. Neither deployment required solving the hard problems of Section 3—they were selected, deliberately or not, so that validation is well-defined and error is recoverable. The lesson generalizes in an uncomfortable direction: the workflow tasks now measured are exactly the ones where the translational gaps are narrowest, while the tasks where AI could most change outcomes—autonomous management of undifferentiated patients, high-stakes treatment selection—remain outside both the benchmark literature and the deployment literature. The workflow era measures where translation is easy; patients benefit where translation is hard [16,17]. Table 2 summarizes the three generations. Figure 1 situates them within the translational-gap framework of this survey.

Table 2. Three generations of clinical AI capability measurement.

Generation	Representative systems	Measurement object	End point
Specialist vision (2016–2019)	Retinal DR [7], skin cancer [2], OCT referral [8], chest X-ray [9,10]	Single-task image classifier	Agreement with specialists (AUC)
Generalist QA (2021–2023)	MedQA [11], Med-PaLM [12], Med-PaLM 2 [13], GPT-4 [14]	LLM on examination questions	Accuracy; physician preference
Workflow systems (2018–)	Autonomous screening [16], ambient scribes [17]	Deployed system in clinical process	Task completion; time saved

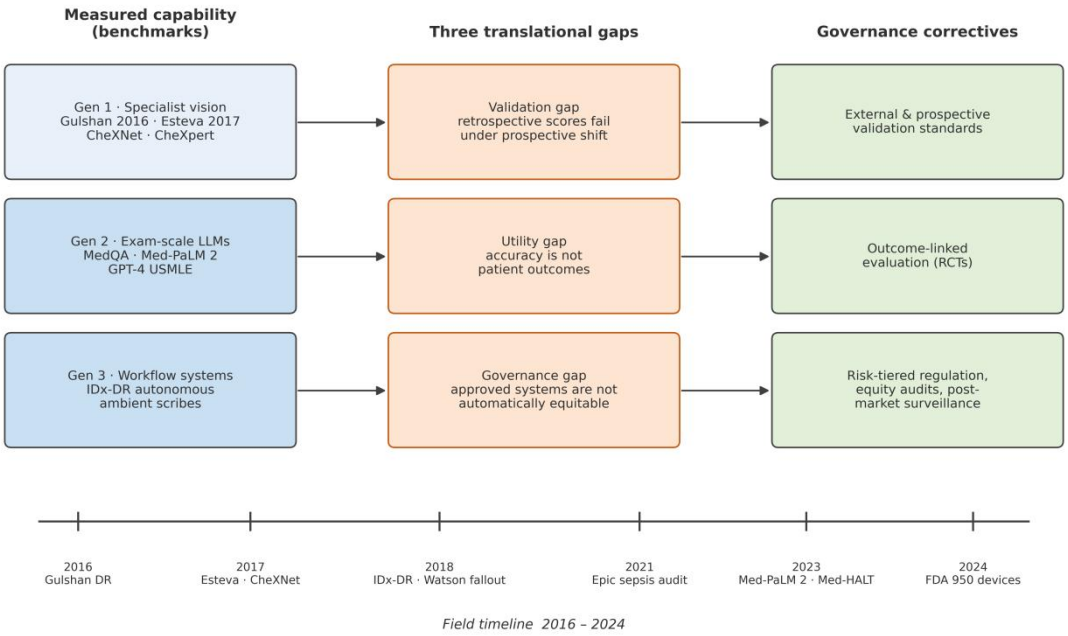

Figure 1. The three translational gaps of clinical AI. Left: three generations of measured capability. Middle: the independent gaps separating benchmark performance from bedside value. Right: the governance correctives each gap demands. Bottom: the field's timeline, 2016–2024.

2.4. The Capability Boundary

Aggregate capability, however, is bounded in ways aggregate scores conceal. Examination performance measures closed-book knowledge under standardized prompts—the setting examinations use precisely because it is not practice [12,14]. Long-form evaluation repeatedly exposed gaps in reasoning, currency of knowledge, and calibrated uncertainty that multiple-choice accuracy does not register [12,14]. Whole regions of practice remain thinly measured at all: dosage and drug-interaction reasoning under real-world polypharmacy, rare-disease presentation, multimodal integration of imaging with narrative records, and the epistemics of not knowing—a clinical skill examinations reward only in passing. And deployment evidence remains concentrated where measurement is easiest: documentation is now measured at industrial scale [17], while diagnostic autonomy beyond retinal screening remains rare [16]. There is an asymmetry worth naming: the tasks AI has penetrated fastest are those whose outputs are most easily reviewed—text a clinician edits, images a radiologist confirms—while the tasks patients most need help with are precisely those whose outputs are hardest to check. What AI systems can do, in short, has been measured well; what deployed systems do to patients has not. That distinction is the subject of the next section.

The boundary analysis also predicts where the field's measurement effort will concentrate next, and why that prediction is itself uncomfortable. The pattern across all three generations is

that capability follows measurability: imaging benchmarks preceded imaging deployment, examination benchmarks preceded documentation deployment, and the tasks currently unmeasured—polypharmacy reasoning, undifferentiated presentations, epistemic humility about not knowing—are precisely the tasks without a data collection tradition [12,14]. If the pattern holds, the next capability frontier will be set by whatever new measurement infrastructures are built rather than by modeling breakthroughs, which inverts the field's usual narrative of algorithm-driven progress. It also creates a quiet selection pressure worth naming: tasks that resist measurement will continue to resist deployment, not because they are harder for AI but because nobody can claim anything about them—and the aggregate effect is a clinical AI portfolio optimized for the measurable rather than the important. Recognizing that pressure is the boundary section's practical yield: what the field chooses to build benchmarks for is, increasingly, a public-health decision, and it is currently being made by default rather than by design [12,14].

3. The Three Translational Gaps

3.1. The Validation Gap: Retrospective Scores Under Prospective Conditions

The first gap is metrological. A retrospective benchmark measures performance on a static, curated dataset; a deployed model meets a moving clinical world. Zech et al. provided the canonical demonstration: chest-radiograph classifiers exploited hospital-specific signatures—platform artifacts and disease-prevalence patterns of their home institutions—and performance consistently degraded when tested at other hospitals, because the model had partially learned where an image came from rather than what it showed [18]. The critique extends to deployed systems at scale. An independent external audit of Epic's widely implemented sepsis-prediction model—deployed across hundreds of U.S. hospitals with a reported AUC of 0.76–0.83—measured an AUC of 0.63, and found the model missed two-thirds of sepsis cases while alerting on many patients who never developed it [19]. Dataset shift itself has been given a clinical taxonomy—changes in practice, population, measurement, and label conventions—along with the argument that safe deployment requires informed clinician-users plus technical governance, because shift cannot be eliminated, only monitored [20].

In EBM terms, a retrospective benchmark carries roughly the weight of mechanistic plausibility: necessary, encouraging, and not licensure-grade. The corrective is external and prospective validation on moved data—a standard the field has named far more often than it

has funded. The persistence of the gap has an incentive explanation: public benchmarks are cheap, fast, and competitive, and they reward the institutions that build them; external validation is slow, expensive, unglamorous, and its findings—whoever pays for them—are publishable precisely when they are damaging. No amount of methodological exhortation overcomes that asymmetry; only funding structures and regulatory requirements can, which is why the validation gap is ultimately a governance problem even though its failure mode is statistical.

Zech et al.'s demonstration repays mechanistic unpacking, because it shows the failure is not poor engineering but rational learning of spurious structure. Their classifiers, trained on chest radiographs from multiple institutions, could partly predict which hospital an image came from—an astonishing feat that is entirely predictable once one notices that an institution's radiology plates, patient positioning, acuity mix, and disease prevalence leave signatures in pixels [18]. A model trained on one institution's data has no incentive to distinguish those signatures from pathology: within the training distribution, they correlate perfectly with disease, and the loss function rewards exploiting them. Performance only collapses when the correlation breaks—at a new institution, where the signature no longer tracks the disease—and the collapse is therefore guaranteed by the training setup, not by negligence [18]. The Epic sepsis audit supplies the deployed-system counterpart at commercial scale: the model's vendor-reported AUC of 0.76–0.83 became 0.63 under independent evaluation, with two-thirds of sepsis cases missed and a warning burden falling heavily on patients who never developed sepsis—a pattern the audit attributed partly to the model's reliance on features correlated with when sepsis was suspected rather than whether it was present [19]. Between them, the two studies define the validation gap's signature: a model that is excellent where it was made and unreliable where it is used, with no internal signal distinguishing the two regimes [18,19].

3.2. The Reliability Gap: Hallucination in Generated Medical Content

The second gap is generative. When an LLM fabricates a medical citation, the error is not cosmetic: it is a false representation of the clinical evidence base, with the burden of detection falling on time-pressed clinicians. The measured rates are sobering. Auditing ChatGPT-generated medical content, Bhattacharyya et al. found that 47% of the 115 generated references did not exist and a further 46% existed but were inaccurate—only about 7% were authentic and accurate [21]. The Med-HALT benchmark extended the question internationally, testing memory and reasoning hallucination across examination questions from multiple countries and showing that models fabricate plausible-but-false clinical knowledge, not merely citations [22]. Hallucination improves with model generation but does not vanish: in bibliographic citation

tasks, fabrication fell from roughly 55% for GPT-3.5 to about 18% for GPT-4 [23]. Retrieval-augmented generation (RAG), which grounds generation in retrieved documents [24], narrows the fabrication channel but does not close it—the model can still mischaracterize what a retrieved source says, and grounding introduces its own failure surface in retrieval quality. The residual risk is insidious precisely because grounding works: a grounded system's errors arrive attached to real sources, so its fabrications wear the appearance of scholarship. Detection, moreover, is professionally asymmetric—checking whether a citation exists is mechanical, but checking whether a cited study supports the claim made of it requires exactly the expertise that automation was supposed to economize.

The structure of the Bhattacharyya audit shows the failure mode with unusual clarity, because the authors graded every generated reference rather than sampling. Of 115 references produced across medical prompts, the 47% that did not exist at all were the detectable failure—a database query eliminates them—and the audit's deeper finding is the surrounding 46%: references that were real publications but misdescribed, cited for conclusions they did not reach, wrong in author, journal, or year [21]. This middle band is the industrially important one, because it passes the first check a busy clinician would run—the article exists, the title sounds right—and fails only the second, which requires reading the article. Med-HALT's cross-country design extends the same structure from citations to knowledge: using examination questions from multiple countries' medical curricula, it distinguishes memory hallucination (fabricating facts within the model's claimed competence) from reasoning hallucination (arriving at unsupported conclusions from given information), and finds both present across systems, with performance degrading on questions from regions whose curricula are underrepresented in training data [22]. The composite picture is a failure surface with a gradient: fabrication is densest where verification is hardest—nonexistent sources, misdescribed sources, underrepresented curricula—and thinnest where a checklist already exists. That gradient, not any single number, is the reliability problem: the residual errors are concentrated precisely where the professional's time is scarcest [21,22].

Legal medicine has an instructive parallel here: courts have sanctioned attorneys for filing fabricated AI-generated citations, and the profession's response—verification duties assigned to the licensed professional—is precisely the liability structure medicine is now improvising. The difference is that a fabricated medical citation can mislead treatment decisions for patients who never consented to be downstream of a language model.

3.3. The Utility Gap: From Accuracy to Outcomes

The third gap is the one patients actually experience: technical accuracy is not clinical benefit. Tschandl et al. demonstrated the asymmetry with an interventional study in skin-cancer recognition: AI assistance improved less experienced clinicians but degraded the diagnostic accuracy of experts, who did worse with the tool than either they alone or the AI alone [25].

The study's design deserves its reputation, because it measured collaboration rather than claiming it. Clinicians of graded experience diagnosed dermoscopic cases first unaided, then with AI suggestions presented alongside the image; the crossover structure let the analysis separate the AI's accuracy from the clinician's response to the AI—a distinction that aggregate "human-plus-AI accuracy" numbers, the era's dominant reporting format, structurally destroys [25]. The findings that survived that separation are the utility gap in miniature: experienced clinicians—whose unaided accuracy exceeded the AI's—were pulled toward the AI's errors when these appeared, a deference gradient invisible in any accuracy table; less experienced clinicians improved, meaning the same tool simultaneously raised the floor and lowered the ceiling of clinical performance [25]. The policy consequence is stark: the same deployment is beneficial or harmful depending on the user's expertise, so "does the tool work?" is not a property of the tool but of the tool–user–task triple, and no pre-deployment evaluation that fixes one user population can certify another. The collaboration surface has since acquired its own literature and its own vocabulary—automation bias, algorithmic aversion—but the 2020 study's core result remains the field's cleanest demonstration that the unit of evaluation for medical AI is not the model [25]. Human–AI collaboration is therefore not automatically superior to its components; it has its own performance surface that must be measured. More broadly, the clinical-impact literature has long identified the same deficiency: the overwhelming majority of reported AI performance comes from retrospective studies with no patient-level outcome, and evidence of improved outcomes remains the exception rather than the rule [6]. An AUC is a property of a dataset; mortality, time-to-treatment, and equity of access are properties of a health system. No benchmark score entails the second set. The corrective is outcome-linked evaluation—interventional studies in which the unit of analysis is the patient pathway—but such studies remain scarce relative to the deployment pace documented in Section 2.3. The scarcity has a structural cause as well as a financial one: outcome trials impose time constants—months to years of follow-up—that are alien to software release cycles, so deployment and evidence arrive on different clocks, and the system ships first. Medicine solved this misalignment for drugs by making the evidence clock the licensing clock; whether AI

deployment can be made to wait for its evidence is, at bottom, the regulatory question of Section 4.

Taken together, the three gaps explain the field's founding contradiction without paradox: the 2017 benchmark milestone [2] and the 2017 deployment failure [3] were measurements of different things, taken at different points of the translational pipeline, read as if they were the same.

4. Closing the Gaps: Governance as Clinical-Grade Evidence

The governance responses to the three gaps share a single logic: converting statistical performance into clinical-grade claims requires the same apparatus medicine built for therapies—tiered evaluation, independent audit, and post-market surveillance.

4.1. Regulation and Post-Market Surveillance

Regulatory practice is the most concrete instance. The U.S. FDA's list of AI-enabled medical devices reached 950 authorized products by August 2024, of which roughly 76% are radiological and the overwhelming majority were cleared through the 510(k) substantial-equivalence pathway rather than de-novo trials—a proportion that itself illustrates the validation gap: most authorized devices were never required to demonstrate performance on new, prospective data [26]. The EU Artificial Intelligence Act addresses the lifecycle rather than the entry event: AI systems that are safety components of medical devices fall into the high-risk tier, triggering risk management, clinical evaluation, human oversight, and—critically—post-market monitoring obligations, making deployed-model drift an auditable compliance object rather than a surprise [27].

The FDA list's composition tells the validation-gap story in administrative data. That roughly three-quarters of the 950 authorized AI-enabled devices are radiological is not a statement about where AI works best but about where evaluation is easiest: imaging has standardized inputs, established reference standards, and a regulatory pathway (510(k)) that certifies similarity to a predicate device rather than clinical performance on new data [26]. The pathway's logic—substantial equivalence—is inherited from devices whose behavior does not depend on a training distribution, and the growing share of adaptive, learning-enabled systems inside it is what pushed the regulator toward predetermined change-control protocols: documents that specify, before deployment, what model updates are permissible and what re-evaluation they trigger [26]. The DeepMind-era concern that motivated such thinking—models whose behavior is a function of data no one has enumerated—are, in regulatory terms, devices that keep

changing after the inspection; the change-control plan is the attempt to make continued learning compatible with continued accountability by converting model drift from an event into a documented process [26,27]. The machinery is young and its enforcement untested, but the direction is unambiguous: the permission-to-deploy conversation is moving from "was it validated?" to "who watches it, on what schedule, and what happens when it drifts?"—which is the validation gap being answered with governance instruments because the technical instruments did not arrive [26,27]. The WHO's guidance on ethics and governance of AI for health supplies the normative frame: transparency, inclusiveness, and accountability as design requirements, not aspirations [28]. LLMs stress this machinery: general-purpose chatbots do not fit device categories built for single-purpose instruments, and their deployment is running ahead of any settled regulatory pathway. A second tension is life-cycle: models that keep learning after authorization break the assumption, inherited from device regulation, that the validated artifact is frozen—hence the regulatory interest in predefined change-control plans that make continued learning compatible with continued accountability [26,27].

4.2. Equity as a Governance Obligation

The governance gap also has a distributional dimension. Obermeyer et al. dissected a widely used population-health algorithm and showed it systematically underreferred Black patients—not through any discriminatory variable, but because it used healthcare cost as a proxy for healthcare need, and Black patients historically received less spending for the same illness [29]. The bias was invisible in the model's accuracy metrics and emerged only when the algorithm's decisions were audited against outcomes.

Two features of the episode generalize beyond its case, and both sharpen the governance gap's definition. First, the mechanism was proxy selection, not data preprocessing: the algorithm predicted healthcare costs as a stand-in for healthcare needs—a choice that is individually defensible (costs are well-measured, complete, and predictive) and jointly discriminatory (equal spending does not mean equal illness across populations with unequal access) [29]. No fairness test applied at the feature level would have flagged it, because the proxy is not a "sensitive attribute"; the disparity lived in the validity of the substitution, which is invisible until the model's outputs are compared against an outcome the proxy was supposed to capture. Second, the episode's remediation shows what governance-grade correction looks like: the authors did not merely document the bias but rebuilt the proxy—reformulating the prediction target on illness measures collected directly from patients—which reduced the predicted disparity dramatically at similar accuracy [29]. Detect, re-specify, re-validate: that

loop is ordinary engineering once the audit exists, which locates the binding constraint exactly where this survey has placed it—not in the difficulty of fixing biased systems but in the institutional fact that nothing in the benchmark–clearance pipeline would ever have required the audit that revealed this one [29]. The lesson generalizes: fairness failures are not exotic failure modes but predictable consequences of proxy choices, undetectable at benchmark level and detectable only through governance-style audit. Equity testing before deployment, and surveillance after it, are therefore not additive virtues but load-bearing parts of the corrective for the utility gap.

4.3. Grading AI Claims with the EBM Hierarchy

The framework's synthesis is that the three gaps map onto medicine's existing evidence hierarchy. A retrospective benchmark is mechanistic-plausibility evidence; external and prospective validation is the analogue of a clinical trial; post-market surveillance is the analogue of pharmacovigilance; and structured reporting and audit are the analogue of trial registration and independent review [20,26,28]. Under this reading, the field's current turbulence is not a crisis of AI but a familiar stage of evidence maturation—the same stage at which therapies were once adopted on mechanistic enthusiasm before randomized evidence tamed adoption curves. The practical implication is that no single intervention closes all gaps: validation standards without outcome studies certify the wrong thing; regulation without equity audit entrenches disparity at scale; and none of it substitutes for clinicians who understand what the tools measure and what they do not [20].

The mapping deserves one more turn, because it converts a metaphor into a checklist. For each tier of the drug-evidence hierarchy, the AI analogue already exists in at least prototype form: trial registration maps to pre-deployment publication of evaluation protocols and data provenance, so that the evaluated configuration is the deployed one; randomized comparison maps to the interventional outcome studies of Section 3.3, including the collaboration designs that measure the tool–user–task triple; pharmacovigilance maps to the post-market monitoring obligations now entering force, whose artificial-intelligence-specific form is drift detection against declared performance; and independent review maps to the external audits—the sepsis evaluation [19], the bias dissection [29]—that demonstrated the model–record gap in both directions [20,26,28]. What the analogy also supplies is the expectation setting: nobody asks whether a drug "works"; one asks what evidence grade supports which indication at what dose. The AI field's interminable "does AI work in medicine?" debate dissolves under the same grammar—the answer is a ledger of claims, each with its evidence grade, its population, and its

surveillance plan [20,26,28]. The surveys reviewed in Section 1.3 catalogued the field's artifacts; the EBM mapping catalogs its evidence, and the difference between the two is the difference between a parts list and a pharmacopoeia.

5. Conclusion

This survey organized clinical AI along three translational gaps and reached three findings. First, measured capability has matured across three generations—specialist image classifiers, examination-scale language models exceeding 86% on licensing-style questions, and workflow-level deployments—with each generation shifting the measurement object from model to system. Second, three independent gaps—validation, utility, and governance—separate those measurements from patient benefit: retrospective scores degrade under dataset shift [18,19], generated content carries fabrication rates that decline but persist [21,23], and human–AI collaboration can underperform either party alone [25]. Third, the governance responses—risk-tiered regulation, post-market surveillance, and equity audit [26-29]—are converging on a single strategy with deep precedent: grading AI claims with the same evidence hierarchy evidence-based medicine applies to therapies.

The framework travels. Any domain combining statistical systems, licensed professionals, and high stakes—law and finance being the nearest—faces the same sequence: benchmarks arrive first, deployment exposes translational gaps, and evidence hierarchies plus surveillance close them. For medicine specifically, the highest-value open problems follow the gaps: funded infrastructure for external, prospective validation; outcome-linked studies of human–AI collaboration, whose asymmetric effects remain underexplored; standardized reporting of hallucination and grounding failure in clinical content; and a settled regulatory pathway for general-purpose medical LLMs. The distance from benchmark to bedside is real—but it is now mapped, and mapping it is the first step to closing it.

The mapping also explains what the next decade of the field will be about, because the three gaps define three distinct research programs rather than one. The validation program's frontier is shift-aware evaluation: benchmarks that sample deployment conditions rather than archives, and drift monitors that declare when measured performance has lapsed [18,20]. The utility program's frontier is the collaboration and outcome literature: study designs in which the unit of analysis is the tool–user–task triple [25], and endpoints are patient pathways rather than agreement statistics—a literature that will have to grow at the pace of clinical follow-up, not release cycles. The governance program's frontier is administrative: adapting evidence hierarchies built for molecules to systems that learn [26,27], and building the equity-audit

machinery that catches proxy harms before enrollment [29]. The survey's closing observation is that none of these programs is speculative—each has its founding results, cited above, and its institutional carriers already in motion. What remains is the unglamorous work of scale and enforcement, which is precisely the work medicine has done before, for every therapy now taken for granted. The benchmark-to-bedside gap is not a scandal to be lamented but a pipeline to be completed—and its completion is the field's actual task.

References

[1] Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

[2] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

[3] Ross, C., & Swetlitz, I. (2017, September). IBM's Watson supercomputer recommended "unsafe and incorrect" treatments for cancer patients. STAT News.

[4] Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31-38.

[5] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930-1940.

[6] Kelly, C. J., Karthikesalingam, A., Suleyman, M., & Corrado, G. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), 195.

[7] Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA, 316(22), 2402-2410.

[8] De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., … Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine, 24, 1342-1350.

[9] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.

[10] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 590-597).

[11] Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14), 6421.

[12] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker,

A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620, 172-180.
[13] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Aguera y Arcas, B., … Natarajan, V. (2023). Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617.
[14] Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. arXiv preprint arXiv:2303.13375.
[15] Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health, 2(2), e0000198.
[16] Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Webster, J. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digital Medicine, 1, 39.
[17] Tierney, A. A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M. A., Kipnis, P., Liu, V., & Lee, K. (2024). Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catalyst Innovations in Care Delivery, 5(3), CAT.23.0404.
[18] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine, 15(11), e1002683.
[19] Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., Ghous, M., & Singh, K. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine, 181(8), 1065-1070.
[20] Finlayson, S. G., Subbaswamy, A., Singh, K., Fornage, J. F., Chodroff, L. C., Syrowatka, A., Kohane, I. S., & Saria, S. (2021). The clinician and dataset shift in artificial intelligence. New England Journal of Medicine, 385(3), 283-286.
[21] Bhattacharyya, M., Miller, V. M., Bhattacharyya, D., & Miller, L. E. (2023). High rates of fabricated and inaccurate references in ChatGPT-generated medical content. Cureus, 15(5), e39238.
[22] Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2023). Med-HALT: Medical domain hallucination test for large language models. arXiv preprint arXiv:2307.15343.
[23] Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. Scientific Reports, 13, 14045.
[24] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems 33.
[25] Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H. P., Zalaudek, I., & Kittler, H. (2020). Human–computer collaboration for skin cancer recognition. Nature Medicine, 26(8), 1229-1234.
[26] U.S. Food and Drug Administration. (2024). Artificial intelligence-enabled medical devices. U.S. FDA. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices
[27] European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

[28] World Health Organization. (2021). Ethics and governance of artificial intelligence for health. WHO.
[29] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.