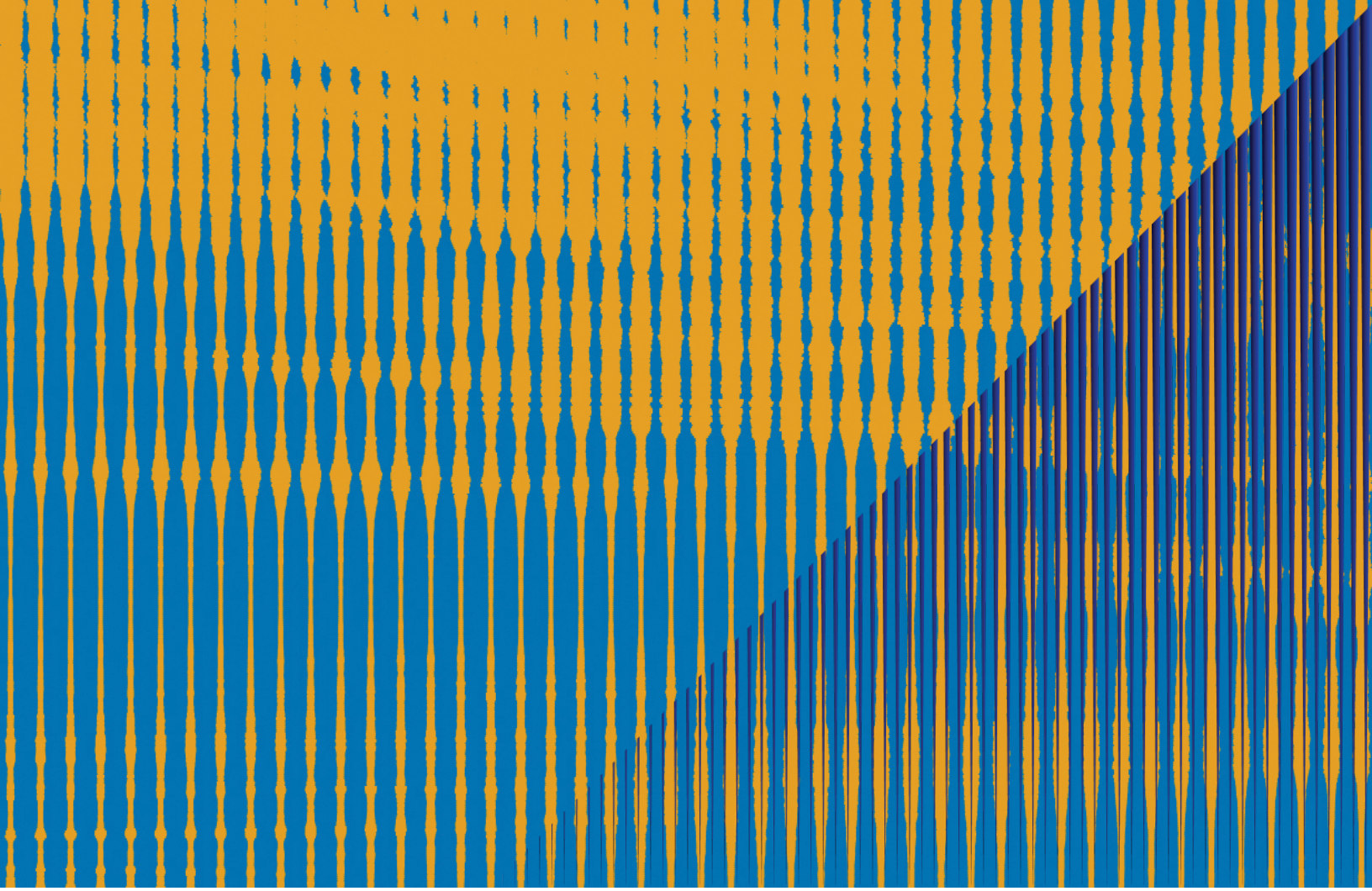


Healthcare AI Industry Report

MAY 2026

ARISE

Table of Contents

Message from ARISE Leadership

Welcome to the inaugural Healthcare AI Industry Report.

Advanced AI systems now perform impressively on benchmark medical tasks, from clinical reasoning and documentation to patient communication and decision support. Recent work from ARISE investigators, including a new publication in *Science* evaluating frontier AI performance on clinical reasoning tasks, underscores how quickly model capabilities are advancing. But this progress also raises a more urgent question: whether these systems are safe, effective, and useful when applied by patients, clinicians, and health systems in the real world.

The 2026 Healthcare AI Industry Report translates the rapidly expanding healthcare AI evidence base into practical guidance for industry leaders. While our State of Clinical AI Report takes a bottom-up approach to synthesize evidence and identify future directions, this report focuses on the issues industry leaders should pay most immediate attention to as healthcare AI moves toward deployment at scale.

We hope this report builds on the momentum around healthcare AI and provides stakeholders across research, clinical practice, and policy with a shared foundation to work from. ARISE collaborates openly with stakeholders across the field, including health systems, regulators, developers, and fellow researchers, to improve healthcare AI. We invite continued dialogue across social and professional networks, and within the broader research community.

Ethan Goh, Adam Rodman, Jonathan H Chen

Executive Summary

A recent McKinsey survey found that more than 80% of surveyed healthcare leaders had deployed their first generative AI use case to end users, with half having done so more than six months ago (Lamb 2026). Frontier models now meet or exceed physician-level performance on a growing set of clinical and administrative tasks (P. G. Brodeur et al. 2026; Goh et al. 2025), and AI-enabled companies captured 54% of the $14.2 billion raised by US digital health startups last year (Zweig et al. 2026).

Early real-world deployments are beginning to show what that may look like at scale. 200+ National Health Service practices deployed triple cardiovascular disease detection with an artificial intelligence-enabled stethoscope (TRICORDER), and showed 2.3x higher detection rates of heart failure and 3.5x for atrial fibrillation (Kelshiker et al. 2026). The MASAI study of nearly 106,000 women showed AI-supported mammography increased cancer detection without raising false positives (Hernström et al. 2025). Public adoption of healthcare AI doubled in the past year (Zweig et al. 2026), with the fastest growth among populations that have historically been the hardest for the health system to reach (Montero et al. 2026).

This report addresses the three questions industry leaders are asking:

1 **Is this technology safe, and how do we prove it?** The field cannot yet answer either question with sufficient confidence. In fact, failures that significantly impact clinical care, including hallucination, omission, automation bias, and silent degradation, are not well represented in current evaluation frameworks. NOHARM found that even the best models produced harmful recommendations in 1 in 11 consultations, and 77% were errors of omission (Wu et al. 2025). Model capability and model safety are distinct, and the field has focused on the former. The next step is evaluation that separates them, and institutions equipped to validate AI against their own clinical use case.

---

2 **Where is AI ready to deploy, and how do we get human–AI collaboration right?** AI alone is outperforming physicians using AI in a growing number of studies, which means there is real value to capture today. Collaborative interfaces and workflows will significantly improve clinician performance, with potential to compound across millions of patient encounters. Prospective studies with patient outcomes as the endpoint are a key missing layer of evidence, and the builders, health systems, and investors who back them will define what "AI that improves care" looks like in practice.

---

3 **What system-level conditions can ensure healthcare AI creates real-world value?** Model capability is only one part of the answer. The harder challenge is building the institutional infrastructure around AI: payment models that reward clinical outcomes, regulatory pathways that can evaluate evolving AI systems, privacy frameworks suited to generative AI, and accountability structures for AI-assisted clinical decisions. Today, that infrastructure does not yet exist. The UK's MHRA AI Airlock shows what regulatory infrastructure for foundation models can look like when built deliberately (Medicines and Healthcare products Regulatory Agency 2026). Health systems should designate clear institutional ownership for AI governance and reduce shadow AI use by providing sanctioned tools. Investors and researchers need to price governance into capital decisions and develop the accountability frameworks that determine liability when AI-assisted decisions cause harm. And because consumer use of healthcare AI chatbots is highest in populations that health systems have traditionally struggled to reach (Montero et al. 2026), institutions that tailor AI for these groups have an opportunity to expand access. The organizations best positioned to lead are those that treat the surrounding infrastructure as the product: evaluation, governance, workflow design, and prospective evidence.

Key Takeaways

Chapter 1: Safety is the missing foundation

- **Technology capability is improving much more rapidly than the peer-review process can keep up with.** Frontier models have saturated benchmarks and crossed performance thresholds across many clinical and administrative tasks, creating a significant capability overhang. The bottleneck is no longer raw performance but the ability to characterize how and why these systems fail in healthcare, especially for generative and agentic AI.
- **Safety and capability must be evaluated as distinct domains.** Medical AI Superintelligence Test (MAST), a joint healthcare benchmarking effort across Stanford, Harvard, and Beth Israel Deaconess Medical Center (BIDMC), shows that stronger overall performance does not translate to safer behavior in clinical environments. Models that excel in knowledge and workflow tasks can be markedly weaker in harm avoidance. Frontier models also perform significantly worse at visual and multimodal clinical reasoning, compared to text reasoning.
- **Organizations need to build internal evaluation capabilities.** External benchmarks provide a directional signal, but no public benchmark matches a 1:1 internal use case. Institutional maturity is increasingly defined by whether a system has the right evaluation expertise and data pipeline in place.

Chapter 2: Deployment and adoption are happening; how do we improve clinical outcomes?

- **AI alone is outperforming physicians using AI in a growing number of studies.** Optimizing human and AI collaboration requires deliberate tool and workflow design, paired with real-world studies that show where models and clinicians each fall short.
- **Human oversight will remain essential, but a human-in-the-loop model can be flawed at scale.** Where access, clinician time, or resources are constrained, requiring human review for every AI-supported task creates the appearance of safety while limiting impact. The open question is which tasks can safely move toward autonomy, under what safeguards, and with what escalation pathways. Answering that requires a taxonomy of clinical and administrative work, paired with measurement of current human performance, risk, cost, and access constraints.
- **The next stage of evidence will be prospective.** Google's Articulate Medical Intelligence Explorer (AMIE) studies illustrate this shift: moving from simulated consultations to a prospective real-world feasibility study at BIDMC, and now toward a nationwide randomized study with Included Health to evaluate conversational AI in virtual care workflows.

Chapter 3: System-level conditions can stand between capability and value

- **Payment models shape the AI that gets built and deployed.** In fee-for-service environments such as academic medical centers, deployment tends to focus on revenue cycle management, coding, and other revenue-growth use cases. In value-based or single-payer systems such as the VA or the UK's National Health Service, the focus is cost-oriented. The same technology lands differently depending on how care is paid for.
- **If regulation does not define accountability, litigation will.** The current US policy environment is oriented toward faster AI deployment and reduced regulatory friction. But without clear federal guardrails, liability will not disappear; it will move downstream to the clinicians, health systems, and vendors closest to deployment. When adverse outcomes occur, courts may assess responsibility across the full chain: from model design and validation to local implementation, monitoring, and whether reasonable safeguards were in place.
- **Infrastructure determines whether capability translates to value.** HIPAA was not designed for foundation models that retain and re-surface information. De-identification is no longer sufficient, Business Associate Agreements (BAAs) do not cleanly extend to chained agentic workflows, and AI tool-use outside of officially sanctioned channels remains the most common day-to-day privacy risk. Closing these gaps requires named institutional ownership of AI governance.

How to Cite This Report

Perez, A., Tusty, M., Morgan, D., Liu, C., Wegner, L., Dutta Gupta, N., Kanjee, Z., Jain, P., Mehta, R., Walton, C., McCoy, L., Nateghi Haredasht, F., Eltahir, A. A., Griot, M., Lopez, I., Lacar, K., Schoeffler, A., Han, B., Zheng, A., Wu, D., Ravi, V., Brodeur, P., Handler, R., Manrai, A., Zwaan, L., Rodman, A., Goh, E., & Chen, J. (2026). The 2026 Healthcare AI Industry Report. ARISE, Stanford, CA.

The authors would also like to thank Abigail Foresman, David J. Wu, John Emmett Worth, Macy Toppan, Marshall Berton, Pavan Shah, Samuel O'Brien, and Zina Jawadi for their contributions.

Education and engagement

Stanford Computational Medicine Colloquia →

Stanford AI in Healthcare Leadership and Strategy: from Innovation to Implementation →

Generative AI and Agentic AI Online Course →

Public Data and Tools

- **Raw data and figures.** Underlying literature and high-resolution versions of all charts presented in this report are available at [Google Drive link].
- **MAST benchmarking tool.** [link to tool]

While human researchers authored and led the writing of this report, researchers used AI tools for finding studies to feature, synthesizing literature, and copy-editing initial drafts. The authors wrote the original copy and utilized AI tools for iterations, with the final review authored by the authors.

Introduction

Frontier models now match or exceed physician performance on a growing range of clinical and administrative tasks, and real-world deployments are starting to show what that capability looks like at scale, from the MASAI study showing AI-supported mammography increased cancer detection to xx.

Adoption has followed, with half of surveyed US healthcare leaders having deployed at least one generative AI use case by the end of 2025 (Lamb 2026). The long-term opportunity to disrupt healthcare across outcomes, costs, and productivity is substantial, with estimates suggesting that the US alone could benefit by $200-$360 billion annually using only technology available today (Sahni et al. 2023). However, many deployments have added complexity instead of reducing it, and reported gains often fall short of what leaders expected. For example, the TRICORDER trial across 200+ NHS practices found no increase in population-level heart failure detection, despite strong algorithmic performance, because most clinicians never used the device and it was not integrated into the EHR. This report aims to explore research on what AI can do and what it can deliver in clinical practice.

This report evaluates healthcare AI across three core domains: clinical decision support, administrative workflows, and direct patient engagement. These categories were derived through a synthesis of academic and industry literature, informed by the O*NET occupational taxonomy, which categorizes work activities and tasks, and refined with input from domain experts. Together, they represent areas where AI is already being deployed in clinical settings and where its impact can be meaningfully assessed. As the report grows year over year, we hope to deepen the analysis within each domain. We also acknowledge that, due to our clinical care focus, there is much within the scope of AI in healthcare that this report does not cover, such as life sciences (drug discovery, genomics, precision medicine), public/population health, and solutions that indirectly interact with clinicians and patients, including payor operations.

Figure 1: Scope of healthcare AI covered in this report

Clinician Decision Support

Prioritize and escalate care: AI identifies urgency, triages needs, and flags when timely intervention or escalation is required.

Interpret findings and diagnosis: AI helps interpret tests, symptoms, imaging, labs, or other clinical findings to support assessment and diagnosis

Plan and manage treatment: AI supports treatment selection, care planning, and ongoing management decisions based on patient context, evidence, and clinical judgment.

Monitor progress and predict risk: AI tracks patient status over time and estimates future risk to support prevention and early intervention

Communicate and coordinate care: AI supports follow-up, referrals, care transitions, and communication across care teams.

Administrative Workflow

Document clinical workflow: AI supports notes, forms, EHR actions, and other structured workflow tasks that operationalize care.

Manage workforce capacity and resources: AI helps allocate or forecast system level resources, staffing, scheduling or operational capacityl to improve care delivery and efficiency

Coordinate patient access and intake: AI manages appointment scheduling, referral routing and prior authorization to streamline patient access to care services.

Manage revenue cycle and reimbursement: AI supports clinical coding, claims processing, and billing workflows to reduce administrative burden and improve revenue cycle accuracy

Direct Patient Engagement

Enhancing patient understanding and accessibility: AI translates clinical information into plain language, adapting content across languages or literacy levels and supports patient comprehension

Faciltiating patient engagement and support: AI supports patients between clinical encounters through health coaching, behavior change guidance, medication adherence, and chronic disease self-management

Patient provider messaging: AI faciltiates messaging with patients, families, and responsible caretakers.

This report focuses on healthcare AI in three domains: Clinician Decision Support, Administrative Workflow, and Direct Patient Engagement. It does not include: life sciences (drug discovery, genomics, precision medicine), public/population health, and solutions that indirectly interact with clinicians and patients, including payor operations.

This report is inclusive across AI model types, including predictive, generative, and agentic AI. Each type of AI comes with its own set of qualities and failure modes. Predictive AI has a long-standing presence in clinical medicine, with decades of research developing rigorous evaluation metrics. Generative and agentic AI, by contrast, are comparatively new to the clinical setting, and the field is still actively developing the benchmarks, evaluation frameworks, and safety standards needed to assess them with the same rigor. Failure modes also differ meaningfully across categories. Where relevant, these distinctions are flagged throughout the report.

How to Read This Report

The three chapters of this report are sequential, with each building on the previous. Together, they address the leading questions industry leaders have about healthcare AI today.

- **Chapter 1: Is this technology safe, and how do we prove it?** Examines what is known and unknown about how healthcare AI fails, why current benchmarks miss real-world failure modes, and what evaluation can look like.
- **Chapter 2: Where is AI ready to deploy, and how do we get human–AI collaboration right?** Turns to the readiness tiers across AI types and clinical tasks, and what real-world evidence suggests about outcomes.
- **Chapter 3: What system-level conditions can ensure healthcare AI creates real-world value?** Considers the system-level conditions, including incentive alignment, government, and privacy infrastructure, to determine whether technically capable AI translates to real value for providers.

The chapters build on each other: safety foundations enable deployment; deployment at scale exposes the system-level conditions that determine outcomes. The conclusion translates these findings into specific recommendations for health systems, builders, investors, and researchers.

Chapter 1: Safety is the missing foundation

1.1 Current healthcare AI evaluation frameworks prioritize accuracy over safety. Evidence for how and why health care AI fails remains sparse, especially for generative and agentic AI.

A fundamental tenet of clinical medicine is the imperative to “do no harm.” In healthcare AI, this principle means avoiding patient or clinician harm while providing consistent, reliable performance (Goldberg et al. 2024). Since the launch of ChatGPT in 2022 (OpenAI 2022), generative AI in healthcare has advanced quickly, from passing medical knowledge exams to attempting clinical reasoning, handling multi-step workflows, and enabling real-world deployment. However, the focus on capability has not been matched by a comparable focus on characterizing how these systems fail. Moreover, the evidence on failure modes for generative and agentic AI remains especially thin. Deployment in this space is outpacing the field’s ability to characterize the risks. This gap has led to two competing narratives: one highlighting strong performance and positive outcomes, the other warning against premature deployment.

Figure 2: A reference model for AI safety

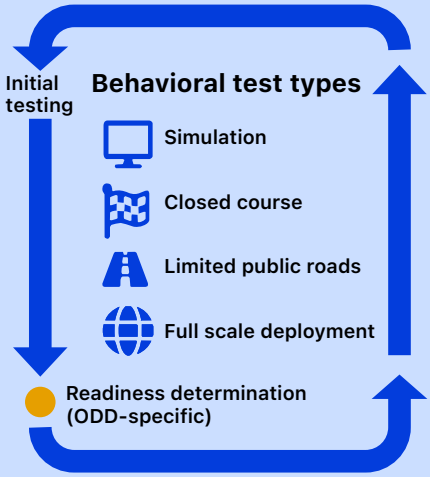
Case example: Waymos undergo a complementary and iterative testing processes

Waymo defines safety as the absence of unreasonable risk perationalized through layered hazards, iterative testing, and continuous monitoring
Adapted from Waymo Safety Methodologies and Safety Readiness Determinations (2020) and Waymo Safety Case Approach (2023)

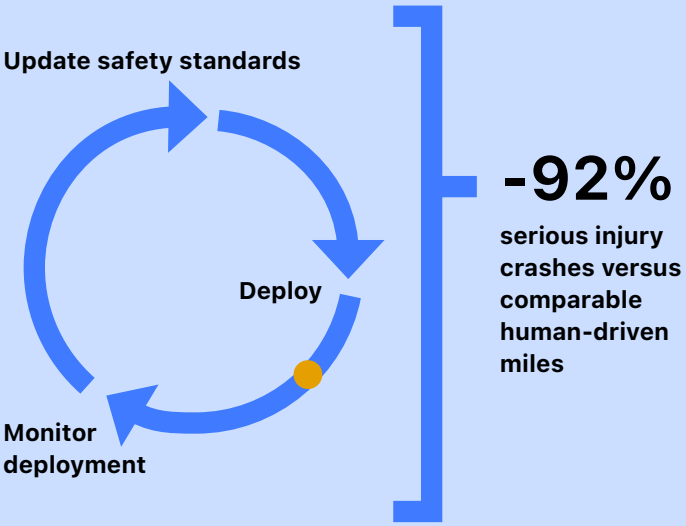
Safety as an emergent development property → **Safety as continuous confidence growth**

Safety as an acceptable prediction and/or observation

- **Architectural:** the physical car components and hardware for backup power
- **Behavioral:** interactions with other vehicles or users of road and abiding by driving rules



- **In service operations:** systems to ensure the tech remains safe like risk management, and field safety programs



The Waymo analogy is illustrative but not directly transferable to medicine. Unlike Waymo, where the destination is fixed upfront, healthcare involves defining the goal itself, a complex, clinician–patient decision. As a result, the exact mechanisms for safety will be different.

AI in healthcare lacks both the framework and the evidence to make a comparable claim to Waymo's 92%

Waymo's safety framework illustrates a layered, lifecycle approach to safety. It is incorporated all along, from architecture and behavioral testing through continuous post-deployment monitoring; which is safety infrastructure that healthcare AI has yet to build.

Other industries deploying autonomous technologies have grappled with a similar gap. Waymo, Alphabet's autonomous vehicle company, offers a useful example of the rigor that safety testing for healthcare AI will require. Rather than defining safety as solely strong performance on driving benchmarks, Waymo grounds its determination in a concept borrowed from decades of system safety engineering: the absence of unreasonable risk (Favaro et al. 2023). The team built a layered framework spanning architecture, behavioral, and operations, in which safety was part of the full lifecycle rather than a one-time readiness check (Webb et al. 2020). The result was 92% reduction in pedestrian-injury crashes compared to human drivers (Waymo, n.d.). At this time, AI in healthcare lacks a comparable definition of "safe enough" and an equivalent infrastructure to demonstrate it.

Building that infrastructure requires first understanding where and how failures occur: in the model itself, in the interaction between humans and AI, and in the broader systems into which these tools are deployed, each of which is explored in the sections that follow.

Defining evaluation for AI in healthcare remains an open problem.

Several frameworks offer useful organizing principles. The WHO framework centers on six principles: protecting human autonomy, promoting well-being and safety, ensuring transparency, fostering responsibility, ensuring inclusivity, and promoting sustainable AI (WHO 2023). FUTURE-AI, developed by 117 experts across 50 countries, organizes safety around fairness, robustness, explainability, universality, traceability, and usability (Lekadir et al. 2025). Stanford's FURM framework (Fair, Useful, and Reliable AI Models), now required for every AI system proposed for deployment at Stanford Health Care, takes a more operational approach: it pairs principle-based ethical review with workflow simulations to estimate achievable benefit, financial sustainability projections, and post-deployment monitoring plans (Callahan et al. 2024). FURM sits within Stanford's broader GUIDE-AI initiative, which develops methods to govern AI across its full lifecycle of use, implementation, development, and evaluation. All of these frameworks reflect genuine expert consensus, but the field is moving faster than the frameworks. Determining that an AI system is safe enough to deploy remains unsolved across institutions and jurisdictions, and no single framework has achieved broad clinical adoption. Until consensus emerges, every institution is responsible for defining and adopting the evaluation methods that best fit its commitments to safety, fairness, and patient care.

1.2 AI in healthcare can fail at three levels: the model, the human-AI interaction, and the system. Often, it can fail at more than one.

Table 1: Unique safety risks by model types in healthcare AI

Where are the distinctive safety risks by model types in healthcare AI?

Categories of AI in medicine from the JAMA Summit Report on Aritificial Intelligence (Angus et al 2025)

	Model behavior	Human-AI interaction	Human-AI interaction
Predictive AI (discriminative AI)	Silent performance degradation	Over-reliance on risk scores	Undetected data drift
Generative AI	Hallucinations, omissions, overconfidence	Sycophany, user-dependent output quality	Identification-based data leaks
Agentic AI	Goal misspecification, reward hacking, multi-step error accumulation, tool-use (e.g., API misspecification, or misuse)	Loss of meaningful oversight	Cascading failures across sytems, unsafe tool interaction
Shared	Bias, poor data, distribution shift	Over-reliance, under-trust, suboptimal or inconsistent usage	Poor integration, lack of monitoring

While this matrix is non-exhaustive, it frames distinct model types and failure modes when deploying different models.

1.2a Model behavior: how models may fail on their own

Safety failures can precede deployment entirely, embedded in how a model was built and on what data. Model-level failure takes many forms, but three are particularly illustrative of the patterns healthcare AI deployments encounter today.

The first is distribution shift. This is the degradation of models trained on data from one clinical context when deployed in another (Han 2025). For example, across external validations of Epic's proprietary clinical decision support tools spanning over 34 sites, real-world performance was consistently more modest in comparison to Epic's own reported performance. Epic's sepsis model achieved an AUROC of 0.65 (vs. Epic-reported 0.76–0.83), the readmission model 0.70 (vs. 0.73), and the end-of-life care index 0.76 (vs. 0.89) (Patel et al. 2026). Similarly, Google's mammography AI required recalibration after prospective deployment revealed a distribution shift, including higher recall rates on newer versus older Hologic mammography systems (Kelly et al. 2026).

Definition:
AUROC (Area Under the Receiver Operating Characteristic Curve) measures how well a classifier distinguishes between two groups, such as patients with vs. without sepsis, where 1.0 is perfect, and 0.5 is no better than chance.

The second is training bias. A widely used care management algorithm systematically underestimated risk for Black patients. The model inherited the structural inequities for how care is spent on Black patients, who receive less care at equal levels of illness (Obermeyer et al. 2019). So because the model optimized for healthcare cost rather than health need, it identified only 17.7% Black patients at the 97th percentile of risk when the bias-corrected figure should have been 46.5%. Bias can also come from inputs, or the patterns that models pick up on, which may be invisible to clinicians. Bias can also enter through inputs invisible to clinicians, with models shown to proxy for race through features like imaging texture or lab values (Gichoya et al. 2022), and is further complicated by the fact that equal accuracy, equal predictive reliability, and equitable outcomes cannot all be satisfied simultaneously when disease prevalence differs across populations (Chouldechova 2017, Kleinberg et al. 2017). Race is among several variables, including gender, socio-economic status, and language, that contribute to bias in clinical care. While some argue that it may be easier for models to correct for these disparities, that depends on whether institutions build the infrastructure to audit and surface these biases.

The third is hallucination and omission in generative models, where systems produce confident but incorrect reasoning or fail to surface critical information, without signaling uncertainty or without clear pathways for escalation to a clinician when needed. Even top-performing models answer confidently when the correct answer is intentionally left out of a multiple-choice, rarely admitting uncertainty and struggling to detect unanswerable questions (Griot et al. 2025). Unlike a clinician who can flag uncertainty, a model that is wrong and confident produces no warning. Recognizing the limits of one's own knowledge is one of the first things physicians are trained to develop, and current models do not reliably do that.

1.2b Human-AI behavior: how the interaction can introduce new challenges

Strong benchmark performance does not guarantee that a model will perform effectively in a clinician hands, and the interaction itself can introduce failure modes that benchmarks don't surface. A retrospective evaluation across 16 primary care clinics in Nairobi, Kenya, tested an LLM-based decision support tool in nearly 1,500 patient visits. The tool's advice was sound most of the time, but in roughly one in thirteen visits, it made a harmful recommendation, and clinical officers were more than twice as likely to act on harmful advice as helpful advice (Agweyu et al. 2026). A companion study provided helpful context: clinicians lost confidence in the tool when its recommendations did not fit local clinical contexts, either because medications were unavailable locally or because guidance strayed from local guidelines. That eroded trust may have contributed to why harmful recommendations were followed more readily (Obong'o et al. 2026). Whether this pattern extends to physician decision-making in better-resourced settings remains less studied, a study found that physician prompting style alone drives 20 percentage points of variability in model outputs, suggesting the dynamic is not unique to lower-resource settings (Lopez et al. 2026).

The same degradation occurs when patients, rather than clinicians, are the human variable. In a randomized controlled study of 1,298 participants, LLMs correctly identified medical conditions in 94.9% of cases when tested in isolation (Bean et al. 2026). But when real users interacted with those same models, accuracy dropped to below 34.5%, or no better than participants using no AI assistance at all. A separate evaluation across 12,000 vignettes and 12 specialties found that diagnostic accuracy fell significantly in interactive patient conversation compared to static vignettes, across every model tested (Johri et al. 2025). The conversational format patients already use is precisely where models perform worst, and most safety benchmarks were never designed to test it. Both of these examples illustrate how human interaction changes model performance, and usually in ways that benchmarks alone do not predict.

What these examples share is that human interaction changes model performance in unpredictable ways, and that direction and magnitude are context dependent. The real-world performance of AI is shaped not only by the clinical task itself, but also by contextual factors such as the user's expertise, the care setting, and the relative strengths of the human and the AI involved. Accordingly, the optimal interaction strategy varies across situations. In pattern recognition tasks where AI has been shown to outperform clinicians, deferring to its output may be appropriate, whereas in more complex clinical reasoning tasks, where human judgment remains stronger, AI outputs warrant closer scrutiny (Zwaan et al. 2026). Qualitative findings from a mixed-methods evaluation further show how workflow integration, trust, and user expertise shape real-world use (Obong'o et al. 2026). More broadly, performance depends on how human–AI interaction is calibrated to task context and relative strengths. As both AI capabilities and human expertise evolve, this calibration becomes a moving target, helping explain gaps between benchmark and real-world performance.

1.2c System-level failures: governance, infrastructure, and silent degradation

Even well-designed models, used appropriately, can fail when the surrounding infrastructure is broken. Before deployment, the governance frameworks needed to guide these decisions are largely absent. Existing proposals like SA-ROC (Sensitivity-Adjusted Receiver Operating Characteristic) attempt to define confidence thresholds for predictive AI, though validation has been limited to narrow tasks like mammography classification and does not extend to generative or agentic systems, where outputs are free-text, and workflows span multiple steps (Kim et al. 2026).

During deployment, AI recommendations enter a clinical environment already overwhelmed by automated guidance. Clinical decision support systems, the alert-based predecessors to today's generative AI tools, are often overridden because alerts fire too frequently, for the wrong patients, with insufficient context to be actionable. In one case, drug-drug interaction (DDI) alerts had override rates above 90% (McGreevey et al. 2020). Generative AI recommendations are now being deployed into this same environment. The result is an environment where clinicians dismiss automated guidance, yet automation bias means they also fail to override incorrect recommendations.

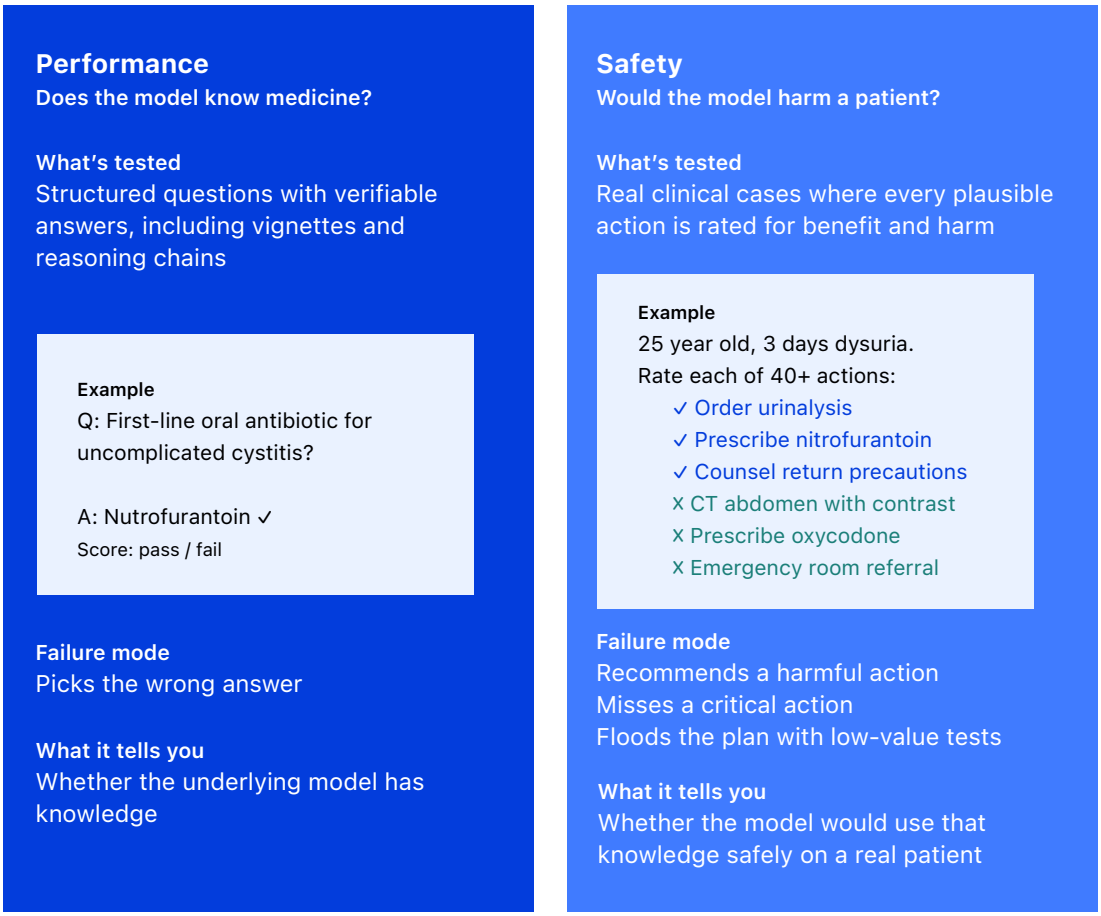
After deployment, models can degrade silently. When models are deployed on static datasets, performance can erode as patient populations and clinical practices change. Proactive monitoring pipelines can detect these shifts before performance drops become clinically consequential, and updating models when drift is detected outperforms both ignoring it and updating on a fixed schedule. For example, during the COVID-19 pandemic, drift-triggered updating improved predictive performance by 0.44 AUROC points over a locked model (Subasri et al. 2025), though systematic monitoring to detect silent model degradation remains rare in practice.

1.3 Benchmarking is still catching up to deployment realities.

A systematic review found that LLM evaluations (n = 519 studies from 2022-2024) mostly focused on medical knowledge, with only 5% using real patient data. Administrative tasks, fairness, bias, and the ability to produce harmful content were understudied (Bedi et al. 2025). With the increasing number of failure modes and real-world deployments, the benchmarking field has begun to respond. MedSafetyBench, MedGuard, and Clinical Safety-Effectiveness Dual-Track Benchmark (CSEDB) each introduced safety dimensions but largely evaluate overall accuracy or adherence to guidelines as primary outcomes rather than the severity or frequency of harm when models make mistakes (Agarwal et al. 2024; Yang et al. 2024; Wang et al. 2025). Most benchmarks also evaluate models in isolation rather than against practicing clinicians, and without a human baseline, AI error rates are difficult to interpret in the context of the care they replace or augment. Beyond what they measure, benchmarks can distort readiness. Goodhart's law predicts that when scores become the target, developers optimize for evaluation rather than underlying capability (Goodhart 1984).

Figure 3: Evaluating performance and safety are related, but should be evaluated separately

Figure 3: Across 31 frontier models on the NOHARM benchmark, safety performance correlated only moderately with existing AI and medical-knowledge benchmarks (Pearson r = 0.61–0.64). Example case adapted from the NOHARM benchmark (Wu et al. 2025b).



NOHARM (Numerous Options Harm Assessment for Risk in Medicine), released in 2025, represents a more direct attempt to measure harm and is one of the few benchmarks to include a head-to-head comparison with practicing physicians. Across 100 real consultations, even the best-performing models produced harmful recommendations in roughly 1 in 11 clinical consultations, rising to 1 in 5 among the worst, with failures of omission driving the majority (77%) of serious errors (Wu et al. 2025c). This finding affirms the need for more research to understand failure modes for AI in healthcare, especially as deployment continues to accelerate in clinical decision-making and towards increasingly autonomous systems.

1.4 Why medicine has not had an “MMLU moment” for evaluation.

In general, AI benchmarks such as MMLU (the Massive Multitask Language Understanding benchmark) helped create a common way to compare models across broad domains of knowledge and reasoning. Healthcare has begun to move in this direction with MedHELM (Bedi, Cui, et al. 2026), which established an important framework for evaluating models across a broad range of tasks. However, the field still lacks a shared standard that frontier developers, AI vendors, and health systems consistently align around.

Evaluations remain scattered, often outdated by the time they are published, and not always comparable across model types. This matters because benchmark design shapes deployment decisions. No benchmark can fully capture real-world clinical readiness, but they can shape which models advance to deployment evaluation in the first place, and benchmark design therefore carries real consequences. Constrained study designs can make systems appear superhuman before they are ready, while overly restrictive evaluations can understate capability and delay tools that would help patients (Kanjee et al. 2023; Rodman et al. 2025). Most existing benchmarks compound the problem by testing a single dimension of clinical performance, whereas clinical readiness is task-specific and cannot be captured by any one evaluation.

What this gap calls for is evaluation that moves at the pace of model development and reflects how clinical work actually unfolds. AgentClinic, which reformats static MedQA cases as sequential clinical simulations with 20+ turns and multi-modal aspects, found accuracy dropped to below a tenth of static-benchmark scores for some models, with the best performer reaching just 62% (Schmidgall et al. 2026). MAST is another attempt to contribute to that effort, curating a high-signal set of benchmarks...MAST is one attempt to contribute to that effort, curating a high-signal set of benchmarks targeting specific cognitive traits: diagnostic and management reasoning, multimodal interpretation, safety and harm avoidance, calibration under uncertainty, and agentic performance in realistic EHR workflows. All frontier models are evaluated rapidly following release, general-purpose and specialist clinical models appear on the same leaderboard, and governance policies are published openly on GitHub. The first full evaluation illustrates why a framework like this matters: it surfaces patterns across safety, reasoning, and modality that no prior benchmark was structured to reveal.

- **Stronger overall capability does not necessarily mean safer clinical behavior.** Some models that performed extremely well on knowledge and workflow tasks were much weaker on measures of harm avoidance. In other words, a model may appear highly capable while still behaving unsafely in clinical contexts. This reinforces a central point of this report: performance and safety must be evaluated as distinct, but necessary domains.
- **Success on traditional reasoning benchmarks does not reliably translate into success on realistic, multi-step clinical tasks inside the EHR.** Several models that performed strongly on static reasoning tasks performed poorly when asked to carry out more agentic workflows, whereas a different set of models proved much more effective in these interactive environments. Clinical care is a sequence of tasks, not a single one, and performance on any one in isolation says little about performance across the chain.
- **Multimodal performance emerged as a distinct strength rather than a simple extension of text capability.** Some middle-of-the-pack models overall performed exceptionally well on image-based tasks such as diagnosis in dermatology and radiology. This supports the view that visual clinical reasoning (which includes not only the above fields but also interpretation of key physical examination findings across all branches of medicine) is a separate competency and should be assessed in its own right, rather than inferred from text performance.

Recommendation: Treat safety evaluation as distinct from capability benchmarking. Ownership maps to the three failure points: builders own model-level failures (1.2a) and share ownership of human-AI failures (1.2b) with health systems, which validate those interactions locally and own system-level failures (1.2c). Researchers and investors act as independent checks, ensuring all three layers are addressed before claims of clinical value are made.

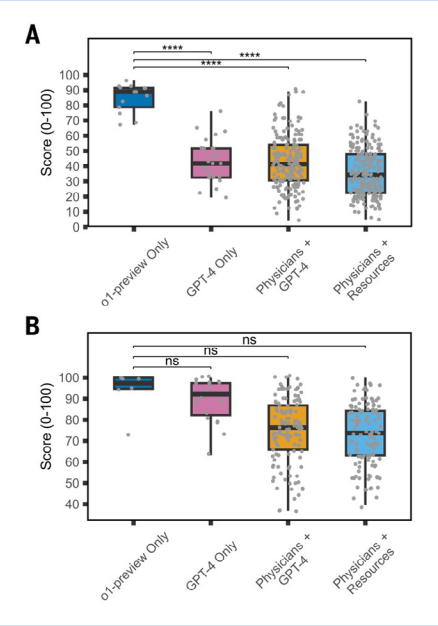
- **Builders:** Report failure modes alongside performance. Build evaluation into the product so customers can monitor performance against their own data.
- **Health systems:** Designate a single institutional owner for AI evaluation. Validate tools on local data before deployment. Ensure continued evaluation on a defined cadence, well past deployment, and including the following model updates.
- **Investors:** Make evaluation infrastructure part of due diligence, either build an in-house toolkit or partner with trusted evaluators. Inform how portfolio companies and investments measure harm, with a continued focus on safety post-investment.
- **Researchers:** Prioritize benchmarks that capture harm severity, omission, and calibration under uncertainty over aggregate accuracy.

Chapter 2: Deployment and adoption are happening; how can we ensure improvement of clinical outcomes?

2.1 For many tasks in simulated environments, AI is already at or above physician-level performance.

Frontier models are accelerating in benchmark performance faster than expected, with the majority of models exceeding or equaling physicians (P. G. Brodeur et al. 2026). Imaging remains the dominant use case with high-impact studies in pathology, neurology, nephrology, oncology, pulmonology, and cardiology (P. Brodeur et al. 2026). Such real-world applications are quickly growing and improving patient care worldwide. For example, TRICORDER, an evaluation across 200+ National Health Service (NHS) practices caring for 1.5 million patients, demonstrated 2.3x higher detection rates for heart failure, 3.5x for atrial fibrillation, and 1.9x for valvular heart disease compared with standard practice over 12 months (Kelshiker et al. 2026). Similarly, the MASAI study involving 105,934 women showed that AI-supported mammography screening achieved higher cancer detection rates with no increase in false positives and fewer aggressive cancers missed (Hernström et al. 2025).

A new study in Science reports that o1-preview outperformed physicians across six clinical reasoning experiments, including NEJM Clinicopathologic Conference cases, NEJM Healer Diagnostic cases, Grey Matters management cases, landmark diagnostic cases, probabilistic reasoning cases, and a blinded evaluation of real emergency department cases at BIDMC. Notably, blinded physician raters could not reliably distinguish AI-generated differentials from human ones, responding "can't tell" in 83.6% and 94.4% of cases (P. G. Brodeur et al. 2026). That said, headlines claiming "AI beats doctors in the ER" overstate the finding. The ED subset compared internal medicine attendings (not ER physicians) on already-admitted patients and scored final diagnostic accuracy, a metric that doesn't reflect what emergency medicine actually optimizes for (Walker 2026). These results support continued study of LLMs as second-opinion and reasoning-support tools, with prospective trials still needed to understand how they can help in real workflows.



That, combined with the growing body of research on multimodal and multi-agent models, has provided sufficient evidence that, from a performance standpoint, the technology operates at exceedingly high rates. For example, MARCUS, a vision-language model trained on ECG, echocardiography, and MRI imaging data, performed 34-45% better than GPT-5 and Gemini 2.5 Pro across modalities and, in multimodal cases, outperformed frontier models by nearly 3x (O'Sullivan et al. 2026). However, especially with multimodal models, the illusion of readiness may be overstated by multimodal benchmarks. Stress tests reveal that LLMs in health can rely on shortcuts instead of robust multimodal reasoning. One study found that when diagnostic images were removed from questions clinicians had verified as visually dependent, most models scored well above random chance, suggesting they were drawing on memorized text patterns rather than visual reasoning (Gu et al. 2025). This raises questions about the validity of benchmarks and the reliability of these models in practice.

Why closed models dominate in healthcare

Open models still lag behind frontier models in raw performance. Epoch AI found open-weight models lag behind state-of-the-art proprietary models by roughly 3 months on average (Emberson 2025). In clinical settings, long-context tasks like chart review and multilingual patient interactions require the latest reasoning capabilities. Additionally, much of the time the decisions on adoption rely heavily on existing infrastructure, which means that models that integrate best into the EHR can often win. However, health systems also weigh PHI exposure, regulatory version control, or fine-tuning on local patient populations, where an open model they own may be reasonable. Open models may also shift who gets to build. As Yun Liu, Senior Staff Research Scientist, Google Research, put it, open models may give clinicians a path to say "I'm an ED doctor and this is the problem I'm trying to solve" rather than waiting on enterprise procurement. See Table 2 for further details.

Table 2: Considerations for closed and open models

Dimension	Closed Models (e.g., GPT, Claude, Gemini)	Open Models (e.g., Llama, MedGemma, DeepSeek)
Raw performance	Frontier, leading in most benchmarks	Larger gaps on long-context, multilingual
Data privacy / PHI	Data leaves the institution	Full data isolation
Enterprise integration	Embedded in Epic / Oracle EHR contracts	Requires in-house development work
Customization	Limited and vendor-controlled	Tune according to needs
Version control	Models change frequently, increasing FDA or regulatory clearance difficulty	Reproducible for regulation
Costs at scale	Higher costs on average	Lower costs on average

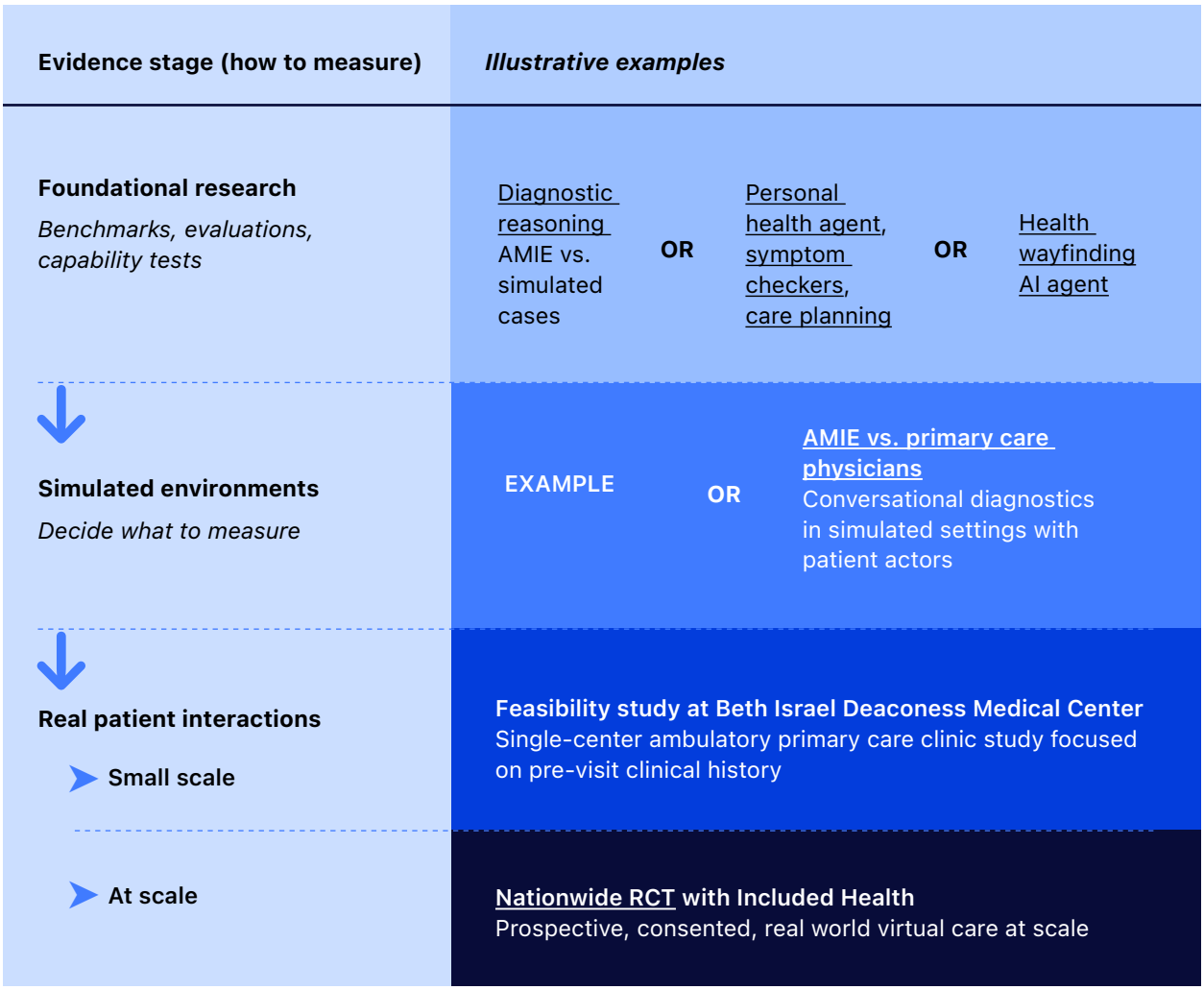
2.2 More real-world evidence is needed to understand when AI improves care and when it adds system complexity.

Early successes can be partly attributed to the relative ease and low risk of the tools deployed, along with a nascent evaluation landscape. AI scribes and administrative tasks could be seen as simpler and lower-stakes than autonomous or semi-autonomous clinical decision-making tools, and even workflow and admin-related tools raise implementation challenges, including potential legal risks (Daily Journal 2026). As risk and complexity rise in AI implementation, the field needs more rigorous and unbiased evaluation methods to ensure deployed tools have a real chance of improving clinical outcomes in live workflows.

Real-world validation of patient-facing AI remains mostly limited to retrospective datasets or controlled pilots, and whether AI-assisted care translates into better patient outcomes remains largely unanswered. The studies that do exist in 2026 illustrate both the promise and the persistent methodological gaps. Most predictive AI RCTs power on accuracy or performance rather than hard clinical outcomes (Soleymanjahi et al. 2024; Hernström et al. 2025). Trials powered on outcomes have produced mixed results. An AI-ECG mortality alert in Taiwan reduced 90-day all-cause mortality from 4.3% to 3.6% (Lin et al. 2024), while an AI-based deterioration monitoring at UVA did not show any improvement in patient outcomes (Keim-Malpass et al. 2026).

For generative AI, the evidence base is thinner still, even as adoption races ahead. Penda Health's collaboration with Open AI is one of the exceptions. OpenAI partnered with Penda Health in Kenya across 39,849 patient visits, reporting 16% fewer diagnostic errors and 13% fewer treatment errors in clinics using AI Consult compared to those without it (Korom et al. 2025). While one of the first real-world studies, it also had its limitations. The study was cluster-assigned rather than randomized, was funded and co-analyzed by OpenAI, and measured errors via independent physician review of 5,666 randomly selected visits rather than patient outcomes directly. Patient-reported outcomes showed no statistically significant differences between arms during the study period.

Figure 4: Google's staged evidence strategy for AMIE shows evidence before scale



Where more rigorous scaffolding exists, it has been built incrementally. Google DeepMind's AMIE has followed a logical progression, moving from foundational research to testing in simulated environments to real patient interaction in small cohorts and at scale. See Figure 4. The logic mirrors how medical students are trained coupled with clinical trial multi-phase structures. "Humans go through one pathway, medical devices go through another," as the AMIE team describes it. "When there is a device with human-AI capabilities, it needs a blended approach." As AI moves from controlled tasks into clinical environments, performance increasingly depends not on model capability alone, but on how clinicians interact with these systems in practice.

2.3 As AI capabilities advance, a central question is how to allocate tasks between physicians and AI systems in ways that improve patient outcomes, reduce physician cognitive load, and strengthen health system efficiency.

Health systems have already begun to automate and augment administrative workflows, with clinical documentation as the clearest example. Across five academic medical centers, scribe adoption was associated with modest reductions in documentation time and an increase of 0.49 visits per week, and physicians in other studies have consistently reported reduced burnout and more time with patients (Olson et al. 2025). This study, however, also noted that xx.

Documentation is only the entry point. Broad estimates project a 5-10% reduction in U.S. healthcare spending from AI use cases already in practice, including administrative automation, fraud and billing error reduction, improved risk prediction, and clinical support (Sahni et al. 2023; Singhal et al. 2025). Early evidence suggests that AI scribes can raise, rather than lower, per-visit reimbursement costs by surfacing previously underdocumented elements that drive higher-intensity coding. For example, Riverside Health reported an 11% rise in physician work relative value units (wRVUs) and a 14% increase in documented Hierarchical Condition Category (HCC) diagnoses per encounter following the adoption of an ambient scribe (Dai et al. 2025). The next step may be to manage administrative workflows as an integrated system spanning billing, revenue cycle management, scheduling, pre-authorization, and claims, rather than piecemeal deployments. One entry for this could be agents, however, early benchmarks suggest there is still more research needed to optimize agent interaction in workflows. On HealthAdminBench, the best agent achieved only 36.3% end-to-end task success with portal guidance and 19.3% without it, even as the strongest system reached 82.8% on individual subtasks (Bedi, Welch, et al. 2026). The open question remains: how to build tools that meaningfully improve clinical outcomes rather than magnify the broken workflows of the existing health system.

2.4 Human-AI teaming is the frontier

The most consequential question is how clinical AI and clinicians perform together. In theory, clinicians and AI could achieve complementarity by flagging each other's errors. In practice, this is not yet happening at scale. Physicians with AI often outperform those with no resources but fail to outperform AI alone (Goh et al. 2025), indicating opportunities for improvement in human-AI interaction. Since humans and AI will continue to collaborate in the long term, the question of how to design their collaboration deserves serious investment. Designing AI that collaborates with clinicians looks very different from designing AI that replaces them. The evidence now points clearly to where the gap lies, why it persists, and how to begin closing it.

Table 3: Key selected studies for AI and human performance in clinical tasks in order of publication

Inspired by Eric Topol's "When Doctors With A.I. Are Outperformed by A.I. Alone" posted on Feb 2, 2025

Title	Date and author	AI Type	Findings
Performance of a large language model on the reasoning tasks of a physician	(Brodeur et al., 2026)	Generative AI	Open AI's o1 series outeprformed physicians and older models across five experiments and one real-world study comparing primary care physicians and AI second opinions
Automation bias in large language model–assisted diagnostic reasoning among physicians trained in AI literacy — A randomized clinical trial	(Qazi et al., 2026)	Generative AI	Physicians exposed to deliberately introduced errors had a lower diagnostic accuracy (73.3%) compared to the control group that was not exposed to errors (84.9%)
Physicians and Large Language Model-Generated Discharge Summaries	(Williams et al., 2025)	Generative AI	LLM drafted hospital discharge summary narratives matched physician quality but had more errors; harm potential was low for both
Towards conversational diagnostic artificial intelligence	(Tu et al., 2025)	Generative AI	AMIE outperformed 20 primary care physicians on 30 out of 32 axes for diagnostic accuracy and performance and on 25 out of 26 axes according to patient-actors across 159 case scenarios
Towards accurate differential diagnosis with large language models	(McDuff et al., 2025)	Generative AI	AMIE outperformed unassisted clinicians (59.1% compared to 33.6%). AMIE also improved clinician performance (51.7%), but AMIE-assisted performance did not outperform AMIE alone.
Comparison of initial AI and final physician recommendations in AI-assisted virtual urgent care visits	(Zeltzer et al., 2025)	Generative AI	Physician reviewers rated AI recommendations as more clinically optimal than physicians' decisions (77% vs. 67%). In cases where physicians outperformed AI, researchers attributed it to physicians' adaptability to new clinical information during live consult.
Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading: the MASAI trial	(Kristiansen et al. 2025)	Predictive AI	AI-supported mammography screening matched standard double reading on interval cancer rates while detecting more cancers overall (80.5% vs. 73.8% sensitivity) and reducing radiologist workload.
Artificial intelligence links CT images to pathologic features and survival outcomes of renal masses	(Xiong et al., 2025)	Predictive AI	A deep learning model outperformed six of seven radiologists at predicting renal mass malignancy on CT scans (AUC 0.871), and improved radiologists' accuracy when used as an assistive tool.
GPT-4 assistance fo rimporvemnet of physician performance on patient care tasks: a randomized controlled trial	(Goh et al., 2025)	Generative AI	Physicians with GPT-4 scored 6.5% higher on management reasoning than using conventional resources and there was no significant difference between LLM-augmented physicians and LLM alone.
Large language model influence on diagnostic reasoning: a randomized clinical trial	(Goh et al., 2024)	Generative AI	ChatGPT alone outeprformed physicians in diagnostic reasoning, but physicians who had access to ChatGPT did not perform any better without it.
Pathologist-AI collaboration framework for enhancing diagnostic accuracies and efficiencies	(Huang et al., 2024)	Predictive AI	AI improved pathologists' plasma cell detection rate (38% to 79%), reduced unnecessary follow-up tests by 63% and cut diagnosis time by 62%. In a separte lymph node study, AI improved detection of hard-to-spot cancer spread.
Use of a large language model to assess clinical acuity of adults in the emergency department	(Y.K. Williams et al. 2024)	Generative AI	In a 500 subsample, LLM accuracy (0.88) was comparable to resident physician (0.86), with both using only clinical history and without access to conventional resources.
Evaluation and mitigation of the limitations of large language models in clinical decision-making	(Hager et al., 2024)	Generative AI	Physicians significantly outperformed open-access, downloadable models in real patient cases with both single-turn and multi-turn scenarios.

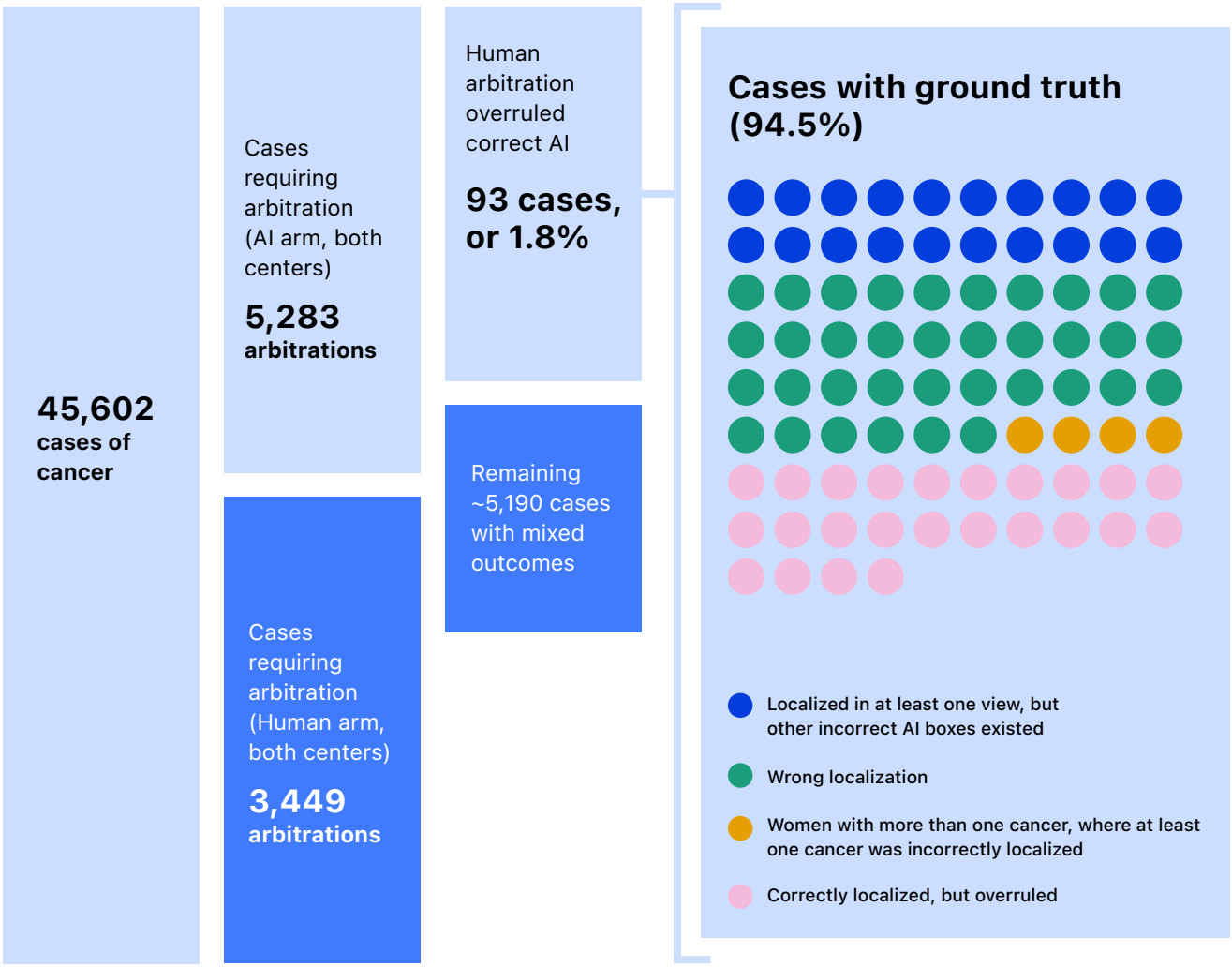
Studies were identified through PubMed searches covering Jan 2024 through May 2026, restricting to peer-reviewed publications and high impact journals. Eligible studies compared AI performance directly against physicians ro licensed clinicians, with both generative AI and predictive AI qualifying. Preprints were excluded, therefore some of the most recent work is not captured.

Complementarity, where human and AI strengths combine to outperform either alone, is uncommon. A 2024 meta-analysis of 106 experiments across healthcare and other domains found that on average, human-AI combinations performed significantly worse than the best of humans or AI alone, with performance losses concentrated in decision-making tasks. A 2025 healthcare-specific reliability analysis pooled 87 conditions from 52 studies, and found that most human-AI teams (95%) outperformed the doctor alone, but only beat the AI alone in 44% of conditions (Liu et al. 2025). In plainer terms, adding AI to a doctor almost always made the doctor better, but adding a doctor to AI usually did not improve AI. Notably, these studies took place in controlled environments, heavily skewed towards imaging and mostly drew from literature before 2023. A re-analysis of the Vaccaro study suggested that experiment designs that incorporated human learning, such as providing feedback, improved results (Berger et al. 2025). The implication then becomes that complementarity requires deliberate engineering, which can include deciding which tasks belong to clinician alone, AI alone, or both together and for shared tasks determining how disagreements get surfaced and resolved. A related open question is how do we identify tasks where AI performs well autonomously but degrades under human oversight, and should those be tested for safe deployment?

Two trust-calibration failures can begin to explain the gap. The first is overreliance, which compounds into deskilling. A multicenter study in Poland indicated that adenoma detection rates decreased 6 percentage points after routine AI exposure (Budzyń et al. 2025). Under time pressure, pathologists abandoned seven percent of initially correct judgments when shown erroneous AI advice. In a multireader mammography study, radiologists' error rates rose on cases where the AI's regional suspicion score was incorrect, even as overall performance improved (Gommers et al., 2025). The second failure mode shows the inverse problem. In a retrospective Nature Cancer imaging study evaluating AI as a replacement for a second human reader across 50,000 women at two NHS screening centers, AI detected more interval and next-round cancers than human readers, but that advantage largely disappeared after arbitration. Human arbitrators overruled 93 cases where AI was correct. Why? In over half of those cases, the AI made localization errors or was inconsistent in its localization, which likely reduced clinicians' trust in otherwise accurate outputs (Warren et al. 2026). Over-trust and under-trust operate at once, often within the same workflow.

Figure 5: Even when the AI was right, clinicians sometimes overruled it largely due to inconsistent or unclear outputs, highlighting that trust hinges as much on reliability as accuracy

Adapted from Warren et al., 2026



The results underscore the importance of consistency and interpretability in AI outputs. Enhancing localization reliability and transparency, while systematically characterizing failure modes, is part of the equation to building clinician trust and enabling robust human–AI complementarity

The gap is closeable, and the design principles are emerging. A 2026 randomized trial of 70 US-licensed clinicians evaluated a purpose-built collaborative interface in which clinicians and AI generate independent assessments, followed by an AI-generated synthesis that surfaces agreements and disagreements. Diagnostic accuracy reached 85% and 82% in this cohort, compared with 75% accuracy in clinicians with conventional resources (Everett et al. 2026). AI alone scored the highest (87%). However, the collaborative approach reliably improved clinician performance, particularly reducing low-scoring cases.

A perspective on medical education and AI; how to avoid deskilling and "never-skilling."

Jason Hom, Chief of Stanford University's Division of Hospital Medicine, noted, "AI is here to stay and will only grow in importance, so we should expose trainees early. But the devil is in the details. We don't want early learners to skip developing foundational medical skills." A recent study highlights concern about false mastery, or "never-skilling," in which students using generative AI completed tasks 48% more successfully but performed 17% worse when it was removed (OECD 2026; Bastani et al. 2025). Some skill loss is expected with any new technology, but that makes it more important to deliberately choose which clinical skills are safe to atrophy and which must remain foundational in training.

As a result, medical schools are starting to adapt in different ways. Stanford Medicine's fall 2025 AI curriculum teaches all MD and PA students to evaluate AI outputs and apply them in patient care, while prohibiting AI on closed-book exams and requiring demonstrated clinical reasoning without it.UCSF has deployed Curate, an AI coach for third-year medical students that provides feedback on illness scripts and clinical notes using vetted institutional content rather than open web sources. NYU Grossman's Institute for Innovations in Medical Education is developing Communication Compass, a Macy Foundation–funded system that analyzes ambient audio from patient-physician interactions to give residents personalized feedback on communication skills. The shared challenge across all of these efforts is integrating AI without eroding the foundational skills on which clinical training is built.

2.5 Public adoption is growing across the patient journey, although trust remains low as AI moves towards clinical decision support

In the last year, public use of AI chatbots for health has doubled (Zweig et al. 2026). One third of US adults use AI chatbot tools for physical and mental health, with larger shares of younger and uninsured adults using them (KFF, 2026). Notably, Black and Latino adults were also more likely to use AI mental health chatbots than white adults. Most who used AI tools did not follow up with a health professional. However, healthcare advice from an AI tool alone is not the preference, with 88% of respondents preferring a healthcare provider over a chatbot when given the choice (Nair et al. 2026).

This growing use of chatbots contrasts with public trust in chatbots. Only half of respondents trusted AI chatbots, and 43% were uncomfortable with healthcare providers using AI in medical care (Nair et al., 2026). Additionally, non-AI users trusted health chatbots 4x less and social media 3x less than AI-users, indicating a divide in consumer trust between those who are using AI and those who are not (Zweig et al. 2026). Physicians also keep their AI use limited, with the majority of AI use for medical research (39%) and chart documentation (28%), with assistive diagnosis as the least used case (17%) (American Medical Association 2026).

It's important to note that these public adoption numbers reflect a small percent of the global population. Microsoft's AI Economy Institute found that only 16.3% of the world had used generative AI tools by the second half of 2025, leaving roughly 83% who have never interacted with a free AI chatbot, let alone paid for one (Microsoft AI Economy Institute 2026). If the tools are being built and deployed largely for and by a small slice of the global population, then public education and AI-health literacy become prerequisites for realizing any of the access the industry is promising. Equally critical is designing these models and tools for all populations across languages, age, and reading levels, cultural contexts, and degrees of digital fluency. The populations most cited as potential beneficiaries of AI-expanded healthcare access may simultaneously be least positioned to use it safely. A 2025 RCT in Australia found that even a brief educational animation was sufficient to produce appropriate skepticism in non-university-educated adults using ChatGPT for health questions (Ayre et al. 2025). This study shows how the gap between informed and uninformed AI health use is not fixed, and with focused efforts, it is possible to close. Equally critical is designing these models and tools for all populations across languages, age, and reading levels, cultural contexts, and degrees of digital fluency.

Recommendation: Design for human-AI collaboration, and build prospective evidence on what works.

- **Health systems:** Pilot tools in clinical workflows with patient outcomes as the primary endpoint, not benchmark accuracy. For any tool that influences clinical decisions, start with human-in-the-loop oversight to surface errors and harm early.
- **Builders:** Build interfaces with physician input from the start that surface AI confidence, flag disagreements, and capture real-world usage data for continuous improvement.
- **Investors:** Prioritize prospective studies that go beyond model performance to demonstrate that clinicians used the AI in ways that improved outcomes relative to the standard of care. For VCs and PE, this means backing companies committed to that evidence bar; for philanthropic and public funders, it means directly resourcing the studies themselves.
- **Researchers:** Run prospective studies that compare AI alone, clinician alone, and AI plus clinician across defined tasks with measurable patient care outcomes. Test multiple interface designs per setting, with physicians involved in the evaluation, and partner with vendors and health systems to align incentives so that evidence-building is undisruptive.

Chapter 3: System-level conditions can stand between capability and value

By the end of 2025, half of the surveyed US healthcare leaders reported that their organizations had implemented at least one generative AI use case, up from 25% in late 2023, and 82% expect positive ROI from their AI investments (Lamb 2026). In another survey report, 85% of executives shared that AI is increasing revenue, while 80% said it's reducing costs (NVIDIA 2026). Despite the expected and shown value from AI, just 4% of organizations have scaled AI implementation (Qventus 2026). This chapter focuses on two parts of the system that shape how healthcare AI is deployed: the privacy and governance architecture under which it operates, and the financial and structural incentives that determine which tools are built and who they reach.

3.1 The regulatory question facing every health system is what constitutes appropriate governance for healthcare AI.

Rules written too tightly risk locking in evaluation criteria that are obsolete within a release cycle and pushing deployment into ungoverned channels; rules written too loosely leave health systems and clinicians exposed to liability and patients exposed to harm. The US and UK illustrate two different responses to that tradeoff.

The UK has concluded that regulators do not yet know enough to write durable rules for foundation models, ambient scribes, or agentic clinical tools, and has built infrastructure to learn before legislating. The MHRA's AI Airlock, launched in 2024 and now in its second phase with multi-year funding through 2029, is a regulatory sandbox that tests AI as a Medical Device in real NHS settings under controlled oversight, with findings published to inform future regulation (Medicines and Healthcare products Regulatory Agency 2026). Running in parallel, the MHRA-established National Commission into the Regulation of AI in Healthcare is due to publish recommendations for a new regulatory framework in 2026. Whether this approach scales is an open question, but it offers a reference model for how public institutions can structure the evaluation function while private actors build within it.

The US is approaching the same questions from a different starting point. The current administration is signaling fast deployment and the removal of regulatory roadblocks, with no comprehensive federal framework for healthcare AI yet in place. The FDA regulates AI that meets the medical device definition, and the Office of the National Coordinator (ONC) for Health Information Technology's Health Data, Technology, and Interoperability-1(HTI-1) rule requires transparency for predictive decision support in certified EHRs, but no single framework addresses healthcare AI as a category. In that vacuum, private actors are setting de facto standards through deployment, with liability and eventual regulation as backstops rather than guardrails.

In either market, a more specific gap persists. Whether governance arrives through a sandbox, certification rules, or sectoral guidance, no framework currently requires a record of what happens when an AI model acts on a patient's data (e.g., what data it received, what it recommended, or what the clinician did with the output).

3.2 For health systems deploying clinical AI, strict privacy standards are a critical part of what sustains patient trust, clinician trust, and institutional credibility.

Nearly nine in ten physicians name data privacy as the primary facilitator of AI adoption in their practice (American Medical Association 2026). Concerns about data privacy remain one of the highest barriers to public trust, with only 23% of AI users open to sharing health data with health tech companies and 15% willing to share with consumer tech companies (Zweig et al. 2026). These concerns are parallel to rising adoption, widening the gap between adoption and the infrastructure built to support privacy.

Regardless of the regulatory approach a health system operates under, a more fundamental problem exists. Privacy frameworks designed for an earlier era of health technology are not built for foundation models. HIPAA, the foundation of US clinical privacy, illustrates the gap, and similar gaps exist under UK GDPR, the EU AI Act, and other governance. HIPAA's Privacy, Security, and Breach Notification Rules have long governed PHI across human and software. In January 2025, the HHS Office for Civil Rights proposed the first significant update to the Security Rule in over twenty years, explicitly requiring that AI tools be included in risk analysis and management (HIPAA Security Rule To Strengthen the Cybersecurity of Electronic Protected Health Information 2025). State laws in Texas and California have added consent and transparency obligations. But HIPAA was not designed for a world in which the model itself stores information and can surface it later. Foundation models powering clinical documentation, decision support, and patient-facing tools introduce a category of PHI retention risk that HIPAA's architecture does not cleanly address. Three gaps in particular are emerging:

- **De-identification alone is no longer sufficient.** Large EHR foundation models can memorize and expose private information despite de-identification, with the highest risk for patients with rare conditions or fewer people that look like them in the data (Tonekaboni et al. 2025). A separate 2025 study showed a fine-tuned LLM can infer additional clinical attributes from partial, unordered inputs (Rosenblatt et al., 2025). Anonymization of training data is necessary, but it is no longer sufficient to ensure data privacy.
- **Business Associate Agreements (BAAs) with vendors are being stretched beyond their design.** Platform-level HIPAA eligibility for services like Azure OpenAI or Vertex does not automatically extend to every feature, log, or downstream agent call (Maguregui and Hennessy 2025). The BAA framework also assumes a one-to-one health system–vendor relationship that breaks down when agentic AI chains actions across multiple vendors in a single workflow.
- **Shadow AI remains the biggest day-to-day risk.** Nearly half of healthcare organizations permitting generative AI have no formal approval process, and even fewer monitor its use (Censinet/CHIME 2025). So, even when AI use is sanctioned, HIPAA-compliant, and covered by appropriate vendor agreements, no record captures what the tool processed or what it recommended in a given encounter. Without that, drift, bias, and harm in deployed AI are difficult for institutions to detect locally. Yet, this scenario doesn't cover what is likely the most common usecase; staff pasting PHI into consumer chatbots to summarize notes or draft messages. This makes inconsistent logging one of the least-tracked privacy risks in deployed healthcare AI. To solve this, we needa named institutional owner for AI governance, consolidation of the fragmented vendor landscape, and sanctioned tools that are genuinely easier to use than the unsanctioned ones staff reach for today.

3.3 Incentives shape what gets built and who benefits.

Healthcare AI is built and bought inside markets, and those markets shape what gets prioritized and who benefits. The financial and structural incentives that determine where deployment happens are as consequential as the privacy and governance constraints examined above.

Payment models determine which AI gets built. In US fee-for-service environments, deployment gravitates toward revenue cycle management, coding, scheduling, and prior authorization, the use cases with the clearest path to revenue. Investment capital follows the same pattern. AI-enabled companies captured 54% of the $14.2 billion raised by US digital health startups in 2025, up from 37% in 2024, with workflow tools alone capturing 42% of sector funding (Zweig et al. 2026). Hospital adoption mirrors this: between 2023 and 2024, billing automation grew 25 percentage points, and scheduling grew 16 points, while AI for treatment recommendations grew just 2 (Chang et al. 2025). Ambient scribes, often pitched as a means of reducing burnout, also raise per-visit reimbursement by surfacing previously underdocumented elements (Dai et al. 2025).

The financial gain on the provider side has produced a predictable countermove. Medicare Advantage and commercial payers are deploying their own AI to ingest claims at scale, downcode, and automate denials, thereby pushing providers to redirect spend toward defensive RCM AI (Dai et al. 2025). The result is a zero-sum coding arms race that consumes operating budgets and that could fund workflow innovation.

Value-based and single-payer systems reorient AI toward prevention, but aligned incentives are not sufficient on their own. The TRICORDER trial across 205 NHS practices showed detection multipliers of 2.3x for heart failure, 3.5x for atrial fibrillation, and 1.9x for valvular heart disease per use across over 6000 cases, yet missed its primary endpoint of significantly increasing overall heart failure detection because real-world utilization fell short of protocol assumptions (Kelshiker et al. 2026). The same technology lands differently depending on how care is paid for, and even aligned payment does not guarantee population-level outcomes without thoughtful integration.

Capital and capability gaps determine who benefits. AI deployment concentrates where the capacity to evaluate, customize, and govern it already exists. In 2024, 81% of urban US hospitals used EHR-integrated predictive AI, compared with 56% of rural hospitals and 50% of critical access hospitals (Chang et al. 2025). A 2026 Nature Health analysis of 3,560 US hospitals found significant geographic clustering in implementation, with hospitals serving high-need regions significantly less likely to have deployed predictive AI; interoperability and IT infrastructure were the strongest consistent predictors of adoption (Hwang et al. 2026). Smaller margins, thinner technical staff, and weaker vendor leverage compound the gap.

For builders and investors, the deployment barriers are orthogonal to the equity ones. Change management, workflow integration, and cost relative to alternatives are often harder and more expensive than expected. The competitive landscape is also tightening. EHR incumbents are bundling AI directly into their platforms. Epic has folded clinician support, revenue cycle automation, and patient-facing AI into its EHR, and Oracle has followed with an AI-enabled patient portal and revenue cycle suite (Zweig et al. 2026). For health systems, this means AI increasingly arrives by default through the EHR rather than through standalone vendors, compressing the addressable market for point solutions. On the patient side, the pattern inverts. AI chatbot use is higher among uninsured and minority adults precisely because clinical access is constrained. The people most likely to use AI interact with consumer tools outside the clinical system, while the institutions serving these patients have the least infrastructure to deploy AI well inside it.

The implication is that AI is most likely to be deployed responsibly in institutions that already have the strongest evaluation and governance infrastructure, and least likely to be deployed responsibly where patients most need its benefits. With 41% of rural hospitals operating in the red and projected Medicaid reductions under the One Big Beautiful Bill Act estimated to cost these facilities $140 billion over multiple years, digital transformation budgets are being further cut (Topchik et al. 2026). Without deliberate intervention through shared evaluation infrastructure, vendor terms designed for smaller institutions, and public funding tied to deployment in underserved settings, the fastest-growing technology in US healthcare will widen rather than narrow existing disparities in care.

Recommendation: Build the institutional infrastructure that turns capability into value

- **Health systems:** Name an institutional owner for AI governance with a budget and a reporting line into the C-suite. Close the shadow AI gap with sanctioned, easy-to-use tools and clear policies on PHI in foundation models. Ensure BAAs cover agentic and chained workflows before deployment.
- **Builders:** Design for privacy beyond de-identification. Make BAAs explicit about agentic and chained workflows, build audit trails into the product, and stress test pricing for safety-net deployment. Compete on integration depth and change management support, not on model performance alone, since the binding constraint for most buyers is workflow fit, and EHR incumbents are bundling AI into platforms faster than standalone vendors are differentiating against them.
- **Investors:** Stress test thesis against payment-model risk and EHR-bundling risk. A use case that returns capital under fee-for-service may not under accountable care; a use case that the EHR vendor can fold into a platform within 18 months is not a durable moat. Account for integration, change management, and governance costs when underwriting investment.
- **Researchers:** Develop accountability frameworks that delineate liability across developers, institutions, and clinicians when AI-assisted decisions cause harm.

Conclusion

Across clinical decision support, administrative workflow, and patient engagement, frontier models have demonstrated performance that meets or exceeds physician benchmarks in a growing range of tasks. The more difficult question, and the focus of this report, is whether the surrounding infrastructure of evaluation, governance, clinical workflow, and financial incentives is adequate to translate capability into reliable patient benefit at scale.

The evidence suggests not yet, and the gaps are increasingly well-defined. Safety evaluation remains the most critical unmet need. Benchmarks have focused on what models get right rather than how they fail, and the failure modes that matter most in clinical settings, including harms of omission, automation bias, and silent model degradation, are precisely those least well captured by existing frameworks. Human–AI teaming also remains underdeveloped. In theory, clinicians and AI systems should outperform either alone; in practice, multiple studies show that collaboration often fails to deliver an additive benefit. The problem is not only model performance, but the design of the collaboration: how AI is introduced into workflows, how clinicians are trained to use it, when AI should defer or escalate, and how disagreements between human and AI judgment should be resolved.

At the system level, three forces will shape whether the next phase of adoption delivers value or diffuses it. The first is payment structure: what gets built and where depends heavily on whether a health system's incentives point toward growing revenue or reducing costs. The second is regulation: in the absence of a clear federal framework, liability will be adjudicated in court. The third is infrastructure: privacy architecture, governance, and internal evaluation expertise are not optional additions to an AI deployment strategy. They are prerequisites.

The organizations best positioned to lead are not those with access to the most capable models, which are increasingly commoditized. They are those who invest in understanding how AI performs in their specific clinical context, govern its use with institutional rigor, and design human-AI workflows with the same care they would apply to any high-stakes clinical process. The technology is advancing rapidly; the greater challenge is building the conditions under which it can be trusted.

Appendix

Table 1: Non-exhaustive alphabetical appendix of tools in healthcare AI

Task Type	Tool	Current Capabilities
Clinical note generation	Abridge	Real-time transcription of patient-doctor conversations into structured notes
Clinical note generation	Ambience	Ambient AI scribe with integrated coding and clinical documentation integrity
Clinical note generation	DeepScribe AI	AI scribe for specialty visits (notably integrated with Flatiron's OncoEMR)
Clinical note generation	Heidi Health	Ambient scribe used in UK primary care with NHS trial showing 51% reduction in documentation time (Source)
Clinical note generation	Suki AI	Voice-enabled AI assistant integrated with EHR
Clinical note generation	Dragon Ambient eXperience (DAX)	Ambient listening AI with RCT finding DAX saved providers ~2.5 hours/week of documentation (Source)
Clinical decision support	UpToDate Expert AI	Generative AI chat built on UpToDate evidence
Clinical decision support	OpenEvidence	AI-driven medical search/chat platform, expanding now to Medical Coding and Billing (Source)
Clinical decision support	Pathway / DoxGPT	Evidence-based AI reference; acquired by Doximity in 2025
Clinical decision support	Isabel DDx Companion	
Consumer health AI	K Health AI	Patient-facing AI triage chatbot
Consumer health AI	Ada symptom checker	Symptom assessment app
Medical research (literature)	Elicit	AI research assistant (semantic search and summarization)
Developer tools	MedGemma 1.5	
	Hippocratic AI	
	OpenAI for Healthcare	

Table 2: Selected safety benchmarks for clinical AI

Title	Authors, Institution	Purpose	Measuring	Methodology	Findings
NOHARM	Wu et al. 2025 (Stanford, Harvard)	LLM safety (errors of commission/ omission)	% cases with "severe harm"; harm rate by case; completeness	100 real primary care for specialist consultation cases across 10 specialties. 12,747 expert annotations by 29 board-certified physicians rating 4,249 clinical management options for harm severity and completeness. Models were prompted with real-world clinical notes; outputs were scored per option.	Severe harm in up to 22.2% of cases (95% CI 21.6–22.8%); 76.6% of errors were harms of omission. Safety performance only moderately correlated with existing benchmarks (r = 0.61–0.64). Best models outperformed generalist physicians by +9.7%; diverse multi-agent approach improved safety by +8.0% vs. solo models.
MedGuard	Yang et al. 2025 (Google)	LLM safety (Truthfulness, bias, etc.)	Pass/fail on 10 aspects (hallucination, sycophany, bias, privacy)	1,000 expert-verified questions (100 per aspect) across 5 safety principles: Truthfulness, Resilience, Fairness, Robustness, Privacy. Mix of closed- and open-ended formats. Evaluated 11 LLMs; includes human physician comparison. Automated + human scoring; public leaderboard available.	All 11 LLMs performed poorly across most aspects regardless of safety alignment mechanisms. Significant gap vs. human physicians. LLMs consistently underperformed on safety dimensions, even where they matched physicians on medical knowledge tasks. Study underscores need for human oversight and AI guardrails.
MedSafety Bench	Kumar et al. 2025 (Harvard)	LLM safety (refusal of harmful medical requests)	Refusal and safety rate	1,800 harmful medical requests grounded in AMA Principles of Medical Ethics, split into evaluation set (900) and fine-tuning set (900 with safe responses). Harmful prompts generated via GPT-4 and Llama-2-7b. Scored by GPT-4 judge ona 1–5 harmfulness scale (1 = full refusal, 5 = full compliance). Tests general knowledge and medical-tuned LLMs.	Medical LLMs comply with harmful requests more readily than general safety-aligned models (avg. harm score 1.78–3.78 vs. lower for general models). Medical jargon in prompts increased compliance. Fine-tuning on MedSafetyBench improved safety while preserving medical performance. Key finding: Medical fine-tuning degrades safety alignment.
CSEDB	Wang et al. 2025 (Medlinker, Peking Union Medical College Hospital)	Clinical LLM safety & effectiveness	Macro-F1 on safety/ effectiveness criteria; gap analysis	2,069 open-ended Q&A items across 26 clinical departments, developed and validated by 32 specialist physicians. 30 weighted metrics covering critical illness recognition, guideline adherence, and medication safety. Hybrid evaluation: automated DeepSeek-R1 judge + manual concordance validation. Tested 6 LLMs (including GPT-4o, Claude-3.7-Sonnet, DeepSeek-R1).	Average total score 57.2% ± 24.5%; safety score (54.7%) consistently lower than effectiveness (62.3%). Significant 13.3% performance drop in high-risk scenarios ($p < 0.0001$). Domain-specific medical LLMs (MedGPT) outperformed general-purpose models by 15.3% overall and 19.8% on safety. DeepSeek-R1 and GPT-o3 led among general-purpose models.
ClinSafe	Buch et al. in development (Harvard)	Broad clinical AI safety	Proposed: accuracy, bias, calibration, and hallucination metrics	Counterfactual demographic swapping at scale: identical clinical presentations are tested with swapped patient attributes (race, gender, insurance status) to measure variation in LLM outputs. Open-source pipeline; will be hosted on GitHub and HuggingFace. Focuses on deployment safety rather than knowledge performance.	Preliminary findings show LLMs can score 95%+ on medical QA benchmarks while still recommending mental health referrals at 6x higher rates for Black patients on identical presentations. Accuracy benchmarks would not catch this. Formal published findings pending.
ClinSafe	Buch et al. in development (Harvard)	Broad clinical AI safety	Proposed: accuracy, bias, calibration, and hallucination metrics	Counterfactual demographic swapping at scale: identical clinical presentations are tested with swapped patient attributes (race, gender, insurance status) to measure variation in LLM outputs. Open-source pipeline; will be hosted on GitHub and HuggingFace. Focuses on deployment safety rather than knowledge performance.	Preliminary findings show LLMs can score 95%+ on medical QA benchmarks while still recommending mental health referrals at 6x higher rates for Black patients on identical presentations. Accuracy benchmarks would not catch this. Formal published findings pending.

Agarwal, Chirag, Tessa Han, Aounon Kumar, and Himabindu Lakkaraju. 2024. "MedSafetyBench: Evaluating and Improving the Medical Safety of Large Language Models." *Advances in Neural Information Processing Systems 37*, 33423–54. https://doi.org/10.52202/079017-1054.

Agweyu, Ambrose, Paul Mwaniki, Wilkister Musau, et al. 2026. "Safety of a Large Language Model-Based Clinical Decision Support System in African Primary Healthcare." *Nature Health*, ahead of print, March 10. https://doi.org/10.1038/s44360-026-00082-5.

American Medical Association. 2026. *2026 Physician Survey on Augmented Intelligence*. https://www.ama-assn.org/practice-management/digital-health/physician-survey-augmented-intelligence.

Ayre, Julie, Melody Taba, Brooke Nickel, et al. 2025. "Can Co-Designed Educational Interventions Help Consumers Think Critically about Asking ChatGPT Health Questions? Results from a Randomised-Controlled Trial." *Npj Digital Medicine* 8 (1): 672. https://doi.org/10.1038/s41746-025-02056-5.

Bastani, Hamsa, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakcı, and Rei Mariman. 2025. "Generative AI without Guardrails Can Harm Learning: Evidence from High School Mathematics." *Proceedings of the National Academy of Sciences* 122 (26): e2422633122. https://doi.org/10.1073/pnas.2422633122.

Bean, Andrew M., Rebecca Elizabeth Payne, Guy Parsons, et al. 2026. "Reliability of LLMs as Medical Assistants for the General Public: A Randomized Preregistered Study." *Nature Medicine* 32 (2): 609–15. https://doi.org/10.1038/s41591-025-04074-y.

Bedi, Suhana, Hejie Cui, Miguel Fuentes, et al. 2026. "Holistic Evaluation of Large Language Models for Medical Tasks with MedHELM." *Nature Medicine* 32 (3): 943–51. https://doi.org/10.1038/s41591-025-04151-2.

Bedi, Suhana, Yutong Liu, Lucy Orr-Ewing, et al. 2025. "Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review." *JAMA* 333 (4): 319. https://doi.org/10.1001/jama.2024.21700.

Bedi, Suhana, Ryan Welch, Ethan Steinberg, et al. 2026. "HealthAdminBench: Evaluating Computer-Use Agents on Healthcare Administration Tasks." Version 1. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2604.09937.

Berger, Julian, Jason W. Burton, Ralph Hertwig, et al. 2025. "Fostering Human Learning Is Crucial for Boosting Human-AI Synergy." Version 1. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2512.13253.

Brodeur, Peter G., Thomas A. Buckley, Zahir Kanjee, et al. 2026. "Performance of a Large Language Model on the Reasoning Tasks of a Physician." *Science* (New York, N.Y.) 392 (6797): 524–27. https://doi.org/10.1126/science.adz4433.

Brodeur, Peter, Ethan Goh, Emily Tat, et al. 2026. *State of Clinical AI Report*. ARISE Network. https://docs.google.com/presentation/d/1A-TcHQb5Hg3-0MoiUFV199FIDBgmGvYhh9DyEU9C_Po/edit?slide=id.g3b69d2bb95c_10_0#slide=id.g3b69d2bb95c_10_0.

Budzyń, Krzysztof, Marcin Romańczyk, Diana Kitala, et al. 2025. "Endoscopist Deskilling Risk after Exposure to Artificial Intelligence in Colonoscopy: A Multicentre, Observational Study." *The Lancet Gastroenterology & Hepatology* 10 (10): 896–903. https://doi.org/10.1016/S2468-1253(25)00133-5.

Callahan, Alison, Duncan McElfresh, Juan M. Banda, et al. 2024. "Standing on FURM Ground: A Framework for Evaluating Fair, Useful, and Reliable AI Models in Health Care Systems." *NEJM Catalyst* 5 (10). https://doi.org/10.1056/CAT.24.0131.

Chang, Wei, Priscilla Owusu-Mensah, Jordan Everson, and Chelsea Richwine. 2025. *Hospital Trends in the Use, Evaluation, and Governance of Predictive AI, 2023-2024*. ASTP Data Brief No. 80. Assistant Secretary for Technology Policy.

Dai, Tinglong, Joseph C. Kvedar, and Daniel Polsky. 2025. "Policy Brief: Ambient AI Scribes and the Coding Arms Race." *Npj Digital Medicine* 8 (1): 780. https://doi.org/10.1038/s41746-025-02272-z.

Daily Journal. 2026. "Patients Sue Sutter Health, MemorialCare over Alleged AI Recordings." *Daily Journal*, April 9. https://www.dailyjournal.com/article/390722-patients-sue-sutter-health-memorialcare-over-alleged-ai-recordings.

Emberson, Luke. 2025. "Open-Weight Models Lag State-of-the-Art by around 3 Months on Average." *Epoch AI's Data Insights*, October 30. https://epoch.ai/data-insights/open-weights-vs-closed-weights-models.

Everett, Selin S., Bryan J. Bunning, Priyank Jain, et al. 2026. "From Tool to Teammate in a Randomized Controlled Trial of Clinician-AI Collaborative Workflows for Diagnosis." *Npj Digital Medicine*, ahead of print, March 18. https://doi.org/10.1038/s41746-026-02545-1.

Favaro, F., L. Fraade-Banar, S. Schnelle, et al. 2023. *Building a Credible Case for Safety: Waymo's Approach for the Determination of Absence of Unreasonable Risk*. https://waymo.com/safety/waymo-safety-case-approach.

Gichoya, Judy Wawira, Imon Banerjee, Ananth Reddy Bhimireddy, et al. 2022. "AI Recognition of Patient Race in Medical Imaging: A Modelling Study." *The Lancet Digital Health* 4 (6): e406–14. https://doi.org/10.1016/S2589-7500(22)00063-2.

Goh, Ethan, Robert J. Gallo, Eric Strong, et al. 2025. "GPT-4 Assistance for Improvement of Physician Performance on Patient Care Tasks: A Randomized Controlled Trial." *Nature Medicine* 31 (4): 1233–38. https://doi.org/10.1038/s41591-024-03456-y.

Goldberg, Carey Beth, Laura Adams, David Blumenthal, et al. 2024. "To Do No Harm — and the Most Good — with AI in Health Care." *Nature Medicine* 30 (3): 623–27. https://doi.org/10.1038/s41591-024-02853-7.

Gommers, Jessie J. J., Sarah D. Verboom, Katya M. Duvivier, et al. 2025. "Influence of AI Decision Support on Radiologists' Performance and Visual Search in Screening Mammography." *Radiology* 316 (1): e243688. https://doi.org/10.1148/radiol.243688.

Goodhart, C. A. E. 1984. "Problems of Monetary Management: The UK Experience." In *Monetary Theory and Practice*, by C. A. E. Goodhart. Macmillan Education UK. https://doi.org/10.1007/978-1-349-17295-5_4.

Griot, Maxime, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2025. "Large Language Models Lack Essential Metacognition for Reliable Medical Reasoning." *Nature Communications* 16 (1): 642. https://doi.org/10.1038/s41467-024-55628-6.

Gu, Yu, Jingjing Fu, Xiaodong Liu, et al. 2025. "The Illusion of Readiness in Health AI." Version 3. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2509.18234.

Han, Larry. 2025. "Addressing Distribution Shift for Robust and Trustworthy Prediction and Causal Inference in Clinical AI Settings." *JAMA Network Open* 8 (6): e2513705. https://doi.org/10.1001/jamanetworkopen.2025.13705.

Hernström, Veronica, Viktoria Josefsson, Hanna Sartor, et al. 2025. "Screening Performance and Characteristics of Breast Cancer Detected in the Mammography Screening with Artificial Intelligence Trial (MASAI): A Randomised, Controlled, Parallel-Group, Non-Inferiority, Single-Blinded, Screening Accuracy Study." *The Lancet Digital Health* 7 (3): e175–83. https://doi.org/10.1016/S2589-7500(24)00267-X.

HIPAA Security Rule To Strengthen the Cybersecurity of Electronic Protected Health Information, Federal Register § Proposed Rule (2025). https://www.federalregister.gov/documents/2025/01/06/2024-30983/hipaa-security-rule-to-strengthen-the-cybersecurity-of-electronic-protected-health-information.

Hwang, Yeon-Mi, Madelena Y. Ng, Malvika Pillai, Michelle P. Sahai, and Tina Hernandez-Boussard. 2026. "The Landscape of AI Implementation in US Hospitals." *Nature Health* 1 (1): 99–112. https://doi.org/10.1038/s44360-025-00016-7.

Johri, Shreya, Jaehwan Jeong, Benjamin A. Tran, et al. 2025. "An Evaluation Framework for Clinical Use of Large Language Models in Patient Interaction Tasks." *Nature Medicine* 31 (1): 77–86. https://doi.org/10.1038/s41591-024-03328-5.

Kanjee, Zahir, Byron Crowe, and Adam Rodman. 2023. "Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge." *JAMA* 330 (1): 78. https://doi.org/10.1001/jama.2023.8288.

Keim-Malpass, Jessica, Sarah J. Ratcliffe, Matthew T. Clark, et al. 2026. "A Randomized Controlled Trial of Artificial Intelligence-Based Analytics for Clinical Deterioration." *Scientific Reports* 16 (1): 7345. https://doi.org/10.1038/s41598-026-39051-z.

Kelly, Christopher J., Marc Wilson, Lucy M. Warren, et al. 2026. "Diagnostic Accuracy, Fairness and Clinical Implementation of AI for Breast Cancer Screening: Results of Multicenter Retrospective and Prospective Technical Feasibility Studies." *Nature Cancer* 7 (3): 494–506. https://doi.org/10.1038/s43018-026-01127-0.

Kelshiker, Mihir A., Patrik Bächtiger, Camille F. Petri, et al. 2026. "Triple Cardiovascular Disease Detection with an Artificial Intelligence-Enabled Stethoscope (TRICORDER) in the UK: A Cluster-Randomised Controlled Implementation Trial." *The Lancet* 407 (10529): 704–15. https://doi.org/10.1016/S0140-6736(25)02156-7.

Kim, Young-Tak, Hyunji Kim, Manisha Bahl, et al. 2026. "Defining Operational Safety in Clinical Artificial Intelligence Systems." *Npj Digital Medicine* 9 (1): 281. https://doi.org/10.1038/s41746-026-02450-7.

Korom, Robert, Sarah Kiptinness, Najib Adan, et al. 2025. "AI-Based Clinical Decision Support for Primary Care: A Real-World Study." Version 1. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2507.16947.

Lamb, Jessica. 2026. *Generative AI in Healthcare: Adoption Matures as Agentic AI Emerges*. McKinsey & Company. https://www.mckinsey.com/industries/healthcare/our-insights/generative-ai-in-healthcare-current-trends-and-future-outlook.

Lekadir, Karim, Alejandro F. Frangi, Antonio R. Porras, et al. 2025. "FUTURE-AI: International Consensus Guideline for Trustworthy and Deployable Artificial Intelligence in Healthcare." *BMJ* 388 (February): e081554. https://doi.org/10.1136/bmj-2024-081554.

Lin, Chin-Sheng, Wei-Ting Liu, Dung-Jang Tsai, et al. 2024. "AI-Enabled Electrocardiography Alert Intervention and All-Cause Mortality: A Pragmatic Randomized Clinical Trial." *Nature Medicine* 30 (5): 1461–70. https://doi.org/10.1038/s41591-024-02961-4.

Liu, Peng, Jiaxin Zhang, Shuaiqi Chen, and Shanguang Chen. 2025. "Human-AI Teaming in Healthcare: 1 + 1 > 2?" *Npj Artificial Intelligence* 1 (1): 47. https://doi.org/10.1038/s44387-025-00052-4.

Lopez, Ivan, Selin S. Everett, Bryan J. Bunning, et al. 2026. "Clinician Input Steers Frontier AI Models toward Both Accurate and Harmful Decisions." Version 1. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2603.14158.

Maguregui, Aaron, and Jennifer Hennessy. 2025. "HIPAA Compliance for AI in Digital Health: What Privacy Officers Need to Know." *Health Care Law Today*, May 8. https://www.foley.com/insights/publications/2025/05/hipaa-compliance-ai-digital-health-privacy-officers-need-know/.

McGreevey, John D., Colleen P. Mallozzi, Randa M. Perkins, Eric Shelov, and Richard Schreiber. 2020. "Reducing Alert Burden in Electronic Health Records: State of the Art Recommendations from Four Health Systems." *Applied Clinical Informatics* 11 (01): 001–012. https://doi.org/10.1055/s-0039-3402715.

Medicines and Healthcare products Regulatory Agency. 2026. *AI Airlock: The Regulatory Sandbox for AIaMD*. April 10. https://www.gov.uk/government/collections/ai-airlock-the-regulatory-sandbox-for-aiamd.

Microsoft AI Economy Institute. 2026. *Global AI Adoption in 2025—A Widening Digital Divide*. https://www.microsoft.com/en-us/corporate-responsibility/topics/AI-Economy-Institute/reports/Global-AI-Adoption-2025/.

Montero, Alex, Julian Montalvo III, Audrey Kearney, Isabelle Valdes, Ashley Kirzinger, and Liz Hamel. 2026. *KFF Tracking Poll on Health Information and Trust: Use of AI For Health Information and Advice*. March 25. https://www.kff.org/public-opinion/kff-tracking-poll-on-health-information-and-trust-use-of-ai-for-health-information-and-advice/.

Nair, Vishnu, Rebecca Handler, and Andre Kumar. 2026. *Public Trust Remains The Limiting Factor For AI Adoption In U.S. Healthcare*. January 18. https://arise-ai.org/blog/public-trust-remains-the-limiting-factor-for-ai-adoption-in-us-healthcare.

NVIDIA. 2026. *State of AI in Healthcare and Life Sciences: 2026 Trends*. https://www.nvidia.com/en-us/industries/healthcare-life-sciences/

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–53. https://doi.org/10.1126/science.aax2342.

Obong'o, Christopher, Grace Njenga, Dickson Otiangala, et al. 2026. "Mixed-Methods Evaluation of Clinician Experiences and Adoption Patterns of an EHR-Integrated Generative AI-Based Clinical Decision Support Uptake by Clinicians in Kenya." *BMJ Digital Health & AI* 2 (1): e000207. https://doi.org/10.1136/bmjdhai-2025-000207.

OECD. 2026. *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*. OECD Digital Education Outlook. OECD Publishing. https://doi.org/10.1787/062a7394-en.

Olson, Kristine D., Daniella Meeker, Matt Troup, et al. 2025. "Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout." *JAMA Network Open* 8 (10): e2534976. https://doi.org/10.1001/jamanetworkopen.2025.34976.

OpenAI. 2022. *Introducing ChatGPT*. November 30. https://openai.com/index/chatgpt/.

O'Sullivan, Jack W., Mohammad Asadi, Lennart Elbe, et al. 2026. "MARCUS: An Agentic, Multimodal Vision-Language Model for Cardiac Diagnosis and Management." Version 1. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2603.22179.

Patel, Hardik, Salvatore Crusco, Derek Hansen, et al. 2026. "A Systematic Review and Meta-Analysis of Externally Validated Epic Clinical Decision Support Tools." *Journal of General Internal Medicine*, ahead of print, March 31. https://doi.org/10.1007/s11606-026-10381-y.

Qventus. 2026. *Beyond the Pilot: How CIOs Are Operationalizing AI Across Health Systems in 2026*. https://www.qventus.com/resources/resource-library/cio-research-report-2026/.

Rodman, Adam, Laura Zwaan, Andrew Olson, and Arjun K. Manrai. 2025. "When It Comes to Benchmarks, Humans Are the Only Way." *NEJM AI* 2 (4). https://doi.org/10.1056/AIe2500143.

Sahni, Nikhil, George Stein, Rodney Zemmel, and David Cutler. 2023. *The Potential Impact of Artificial Intelligence on Healthcare Spending*. No. W30857. National Bureau of Economic Research. https://doi.org/10.3386/w30857.

Schmidgall, Samuel, Rojin Ziaei, Carl Harris, et al. 2026. "AgentClinic: A Multimodal Benchmark for Tool-Using Clinical AI Agents." *Npj Digital Medicine*, ahead of print, April 27. https://doi.org/10.1038/s41746-026-02674-7.

Singhal, Shubham, Drew Ungerman, and Jason Azzoparde. 2025. *Future of US Healthcare: Gathering Storm 2.0 or a Golden Age?* https://www.mckinsey.com/industries/healthcare/our-insights/future-of-us-healthcare-gathering-storm-2-point-0-or-a-golden-age.

Soleymanjahi, Saeed, Jack Huebner, Lina Elmansy, et al. 2024. "Artificial Intelligence–Assisted Colonoscopy for Polyp Detection: A Systematic Review and Meta-Analysis." *Annals of Internal Medicine* 177 (12): 1652–63. https://doi.org/10.7326/ANNALS-24-00981.

Subasri, Vallijah, Amrit Krishnan, Ali Kore, et al. 2025. "Detecting and Remediating Harmful Data Shifts for the Responsible Deployment of Clinical AI Models." *JAMA Network Open* 8 (6): e2513685. https://doi.org/10.1001/jamanetworkopen.2025.13685.

Tonekaboni, Sana, Lena Stempfle, Adibvafa Fallahpour, Walter Gerych, and Marzyeh Ghassemi. 2025. "An Investigation of Memorization Risk in Healthcare Foundation Models." Version 1. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2510.12950.

Topchik, Michael, Troy Brown, Melanie Pinette, Billy Balfour, Ana Wiesse, and Renee Burnham. 2026. *2026 Rural Health State of the State*. Chartis. https://www.chartis.com/insights/2026-rural-health-state-state.

Walker, Graham. 2026. *The ER Has Three Prime Directives. Diagnosis Isn't One of Them.* May 2. https://www.linkedin.com/pulse/er-has-three-prime-directives-diagnosis-isnt-one-them-walker-md-31hcc/?trackingId=Z%2FSOKL%2F5R1W4X%2BBaOKFmLw%3D%3D.

Wang, Shirui, Zhihui Tang, Huaxia Yang, et al. 2025. "A Novel Evaluation Benchmark for Medical LLMs Illuminating Safety and Effectiveness in Clinical Domains." *Npj Digital Medicine* 9 (1): 91. https://doi.org/10.1038/s41746-025-02277-8.

Warren, Lucy M., Jenny Venton, Kenneth C. Young, et al. 2026. "Impact of Using Artificial Intelligence as a Second Reader in Breast Screening Including Arbitration." *Nature Cancer* 7 (3): 507–21. https://doi.org/10.1038/s43018-026-01128-z.

Waymo. n.d. *Waymo Safety Impact Dashboard*. https://waymo.com/safety/impact/.

Webb, Nick, Dan Smith, Christopher Ludwick, et al. 2020. "Waymo's Safety Methodologies and Safety Readiness Determinations." Version 1. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2011.00054.

WHO. 2023. *WHO Calls for Safe and Ethical AI for Health*. May 16. https://www.who.int/news/item/16-05-2023-who-calls-for-safe-and-ethical-ai-for-health.

Wu, David, Fateme Nateghi Haredasht, Saloni Kumar Maharaj, et al. 2025a. "First, Do NOHARM: Towards Clinically Safe Large Language Models." Version 2. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2512.01241.

Wu, David, Fateme Nateghi Haredasht, Saloni Kumar Maharaj, et al. 2025b. "First, Do NOHARM: Towards Clinically Safe Large Language Models." Version 2. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2512.01241.

Wu, David, Fateme Nateghi Haredasht, Saloni Kumar Maharaj, et al. 2025c. "First, Do NOHARM: Towards Clinically Safe Large Language Models." Version 2. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2512.01241.

Yang, Yifan, Qiao Jin, Robert Leaman, et al. 2024. "Ensuring Safety and Trust: Analyzing the Risks of Large Language Models in Medicine." Version 1. Preprint, arXiv. https://doi.org/10.48550/ARXIV.2411.14487.

Zwaan, Laura, Adam Rodman, and Taro Shimizu. 2026. "Trust, Scrutiny, or Collaboration? A Performance-Based Framework for Human–AI Interaction in Medicine." *NEJM AI* 3 (5). https://doi.org/10.1056/AIe2600354.

Zweig, Megan, Jackie Kimmel, and Madelyn Knowles. 2026. *2025 Year-End Digital Health Funding Overview: A Tale of Two Markets*. Rock Health. https://rockhealth.com/insights/2025-year-end-digital-health-funding-overview-a-tale-of-two-markets/.