

Imperfect, Yet Transformative:

How Healthcare AI Delivers Value Before It Is Perfect

Table of Contents

	Executive Summary	05
01	The Wrong Comparison	06
1.1	How We Measure Clinicians, and How We Measure AI	06
1.2	The Three Right Comparisons	06
1.3	Healthcare Was Already Imperfect Before AI Arrived	07
1.4	The Asymmetry of Forgiveness	07
02	The Real Risks, Acknowledged Without Belaboring	08
2.1	Hallucination, False Confidence, and the Average Answer	08
2.2	Bias and Inequitable Performance	09
2.3	The Black Box Problem	09
2.4	Systemic Risks: Security, Model Drift, and Misleading Benchmarks	10
03	Why Imperfect AI Still Improves Healthcare	11
3.1	Where AI Already Outperforms	11
3.2	Enhancement, Not Replacement	11
3.3	AI Reveals the Data and Process Problems You Already Had	12
3.4	Every Transformative Medical Technology Was Imperfect First	12
3.5	The Cost of Waiting, and the Shadow AI Already in Your Building	12
04	Adopting Responsibly	13
4.1	A Risk-Tier Model for Healthcare AI	13
4.2	Governance Anchors	14
4.3	Proportional Oversight and Real Human Accountability	15
4.4	Audit the Data Before the Model	16
4.5	Transparency and Disclosure	16
4.6	AI Literacy: Teaching People to Direct and Challenge AI	16
4.7	Protecting Clinical Judgment Over Time	16
05	A Practical Path to Adoption	17
06	Conclusion: The Right Comparison Changes Everything	18
07	Why Estenda	19
	References	20



Executive Summary

Executive Summary

Healthcare measures human performance constantly, through error rates, quality metrics, peer review, and morbidity and mortality statistics, yet it judges artificial intelligence against perfection instead of against those same numbers. Corrected, that comparison changes what responsible adoption looks like, and it changes it in favor of moving deliberately rather than waiting.

- The right comparison is AI-assisted workflows versus today's human-only, software-limited, error-prone workflows.
- Healthcare already runs on imperfect clinicians, imperfect data, imperfect workflows, and imperfect electronic health records.
- AI carries real risks, including bias, fabrication, opacity, and overreliance, and each of them has a known control.
- The choice is no longer whether to adopt AI. It is governed adoption or unmanaged adoption.

The evidence base has matured. AI now matches or exceeds clinicians on specific, bounded tasks, from retinal screening to breast imaging to clinical documentation, and its failures are documented and, just as importantly, controllable. What remains is the discipline to adopt it under governance rather than around it.

795,000

Americans die or are permanently disabled by diagnostic error each year

1,450+

AI-enabled medical devices authorized by the FDA, about 76 percent in radiology

81%

of physicians used AI professionally in 2026, up from 38 percent in 2023

50.6 → 29.4%

clinician burnout before and after ambient AI documentation at one health system

→ The choice is no longer whether to adopt AI.
It is governed adoption or unmanaged adoption.

The Wrong Comparison

1.1 How We Measure Clinicians, and How We Measure AI

Medicine has spent more than a century building instruments to measure its own fallibility. Morbidity and mortality statistics track adverse outcomes. Quality metrics track readmissions and medication errors. Peer review and licensure exist because expert judgment varies. None of this assumes clinicians are perfect. It assumes the opposite and builds guardrails around it.

Roughly 795,000 Americans die or are permanently disabled by diagnostic error each year across care settings, according to a [2024 study in BMJ Quality & Safety](#), and just five conditions- stroke, sepsis, pneumonia, venous thromboembolism, and lung cancer- account for roughly 39 percent of those serious harms. Earlier [research](#) put diagnostic error at about 5 percent of United States adults seeking outpatient care, or roughly 12 million people a year. Diagnosis is also less consistent than patients tend to assume. When Mayo Clinic compared referral diagnoses to final diagnoses for [286 patients sent for a second opinion](#), only 12 percent had their original diagnosis confirmed as complete and correct, while 66 percent were refined or redefined and 21 percent were distinctly different. These are the accepted background conditions of care, and they are the baseline against which any new tool should be measured.



AI is measured against a different standard. The test is rarely whether it beats the clinician on average. It is closer to whether it is flawless. A model that improves on human performance but still errs is often rejected, while the human baseline it would have improved on is quietly tolerated. That baseline is not even fixed. Doctors often disagree with each other, which is why second opinions exist. Demanding perfection from AI when experts cannot reach it among themselves sets two different standards for the same job. This paper corrects that misstep, and correcting it does not mean lowering the bar. It means holding both sides to the same one.

1.2 The Three Right Comparisons



AI versus the current human-only workflow

Does the assisted process produce fewer errors, less delay, or better access than the process it replaces?



AI-assisted humans versus unassisted humans

Most real deployments are not autonomous, so the right unit of analysis is the clinician plus the tool, not the tool alone.



The cost of adopting versus the cost of waiting

Delay is not a neutral, risk-free default, and it carries its own price in slower workflows, delayed care, and missed improvement.

1.3 Healthcare Was Already Imperfect Before AI Arrived

AI enters a system already shaped by fatigue, cognitive overload, wide variation in practice, incomplete records, and heavy documentation burden. For every hour of direct patient care, physicians spend close to two additional hours on the electronic health record and desk work, according to [time-motion research by Sinsky and colleagues](#), with still more after clinic. Alert fatigue dulls the very safety prompts meant to protect patients. Data is missing, miscoded, or skewed toward the populations that had the best access to care in the first place.

The imperfections people fear in AI, including bias, inconsistency, and error, are already present in the systems AI is being asked to improve. The relevant question is not whether AI is flawless, but whether it reduces the flaws that are already there.

1.4 The Asymmetry of Forgiveness

Human error feels normal, so we forgive it. Machine error feels alien, so we magnify it. A missed diagnosis in a busy clinic is absorbed as an unfortunate part of medicine, while the same miss by an algorithm becomes a headline. This is a human bias, not a property of the technology, and it distorts decisions. It leads organizations to tolerate a familiar failure rate while rejecting an unfamiliar tool that would have lowered it.



- Human errors feel normal and are forgiven.
AI errors feel alien and are magnified.
The gap between the two is a bias, not a measurement.

The Real Risks, Acknowledged Without Belaboring

The case for AI is not a case for ignoring its failures. Each risk below is real, carries particular weight in healthcare, and has a known control.

2.1 Hallucination, False Confidence, and the Average Answer

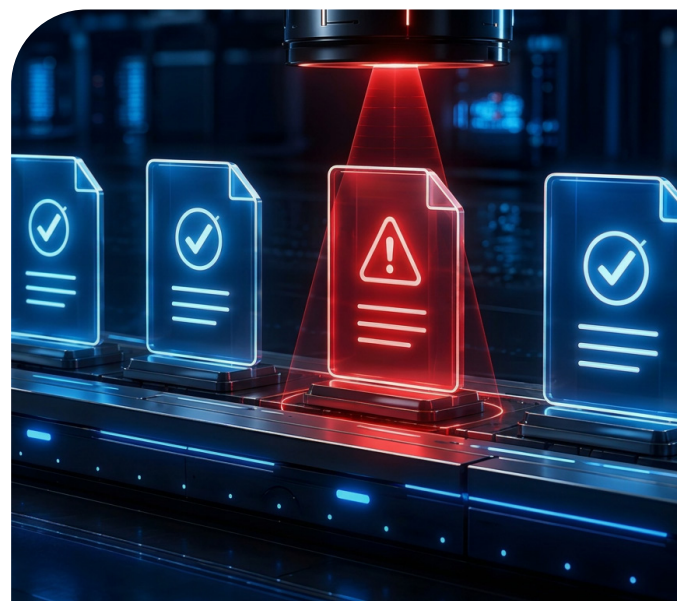
Large language models can produce fluent, confident text that is wrong. Failure modes matter. A fabricated output invents something that does not exist, such as a citation or a study. A misleading output is built on real-looking but incorrect information. Both are dangerous in clinical work, where a plausible wrong answer is harder to catch than an obvious one.

In a **2025 study in Communications Medicine**, six leading models were given clinical vignettes each seeded with a single fabricated detail, and they repeated or elaborated on the false detail in up to 83 percent of cases. A safeguard prompt roughly halved the rate but did not remove it. Not every task is equally exposed. In clinical note summarization, a **framework published in npj Digital Medicine** measured a hallucination rate near 1.5 percent, low but not zero.

Human documentation has never been error-free either. When researchers reviewed 217 dictated notes from two US health systems, **58 percent still contained errors after a professional medical transcriptionist had edited them**, and 42 percent still did after the dictating physician reviewed and signed them. About a quarter of the errors left in those signed notes involved clinical information. The two studies measure different things, so this is not a head-to-head score. What they show together is that documentation has always carried an error rate. The useful question is whether AI lowers that rate or

raises it, not whether it reaches zero. The risk is real, bounded, and something to design around rather than assume away.

Sycophancy is a related and underrated failure. Models are tuned to be agreeable and to keep users engaged, so they often support a flawed premise rather than challenge it. The counter is to ask the model for the counterpoint, to ask why an idea might be wrong, rather than accepting the confident agreement it is inclined to give.



Models also gravitate to the most statistically typical response, which is safe for common presentations and risky for the atypical patient in the tail of the distribution. At scale, this homogenizes output. A **2024 study in Science Advances** found that writers using generative AI produced individually stronger work but collectively more similar work, an increase in personal quality at the cost of shared diversity. In medicine, that pull toward the average answer is both a convenience and a hazard.

2.2 Bias and Inequitable Performance

Bias in healthcare AI is rarely designed in. It is inherited from data, and the data reflects an unequal system. The algorithm does not need to be prejudiced. The data makes it so.

A widely used United States care-management algorithm, examined in [Science in 2019](#), rated Black patients as healthier than equally sick White patients because it used healthcare cost as a proxy for healthcare need. Because less had historically been spent on Black patients, the model saw them as lower risk. Correcting the flaw would have raised the share of Black patients flagged for additional care from 17.7 percent to 46.5 percent.

The same pattern appears across modalities. Pulse oximetry overestimates oxygen saturation in patients with darker skin, missing dangerously low oxygen roughly three times as often; this was first reported in [the New England Journal of Medicine in 2020](#) and later confirmed [in hospitalized children](#). AI models reading chest radiographs [underdiagnose underserved populations](#), with the highest miss rates for intersectional groups. Dermatology models trained mostly on lighter skin [perform worse on darker skin tones](#). One discipline follows from all of it. Audit the data before trusting the model.



2.3 The Black Box Problem: Opaque Reasoning in Machines and in People

Explainability is something healthcare wants and something many models lack. A clinician cannot always trace why a model reached a conclusion, which complicates trust, validation, and accountability. This is a genuine limitation.

Human reasoning is often opaque, too. Expert clinical judgment runs heavily on pattern recognition and intuition that even skilled clinicians cannot fully articulate. The meaningful difference is not that people are transparent and machines are not. It is that around opaque human judgment we have built accountability around opaque human judgment, including licensure, peer review, supervision, and the simple ability to ask a colleague why. Around models operating at scale, those structures are still being built. The task is not to demand that AI explain itself perfectly before use. It is to surround it with the same kind of accountability we already require of expert humans.

2.4 Systemic Risks: Security, Model Drift, and Misleading Benchmarks

A widely deployed proprietary sepsis prediction model, live at hundreds of United States hospitals, was found on external validation to have a sensitivity of 33 percent and an area under the curve of 0.63, well below the performance its developer had reported.

Benchmarks mislead in the other direction too. A strong score sometimes means the model memorized the test, not that it can reason, and medical benchmarks are wide open to this. Most are built from publicly available material like board exam questions, so the questions may already be sitting in the training data. Researchers know it is a problem, which is why they have started building contamination-free medical benchmarks that refresh with new cases. Format matters as well. Models that beat expert scores on USMLE-style multiple choice have done noticeably worse on tests of clinical reasoning under uncertainty, which is much closer to real clinical work. Treat headline numbers with suspicion and test any tool on your own cases, your own patients, your own workflow.

The biggest security risk is the AI already running without anyone’s permission. Shadow AI, meaning staff using consumer tools on their own, puts patient data somewhere the organization cannot see or control. A December 2025 Wolters Kluwer survey found 17 percent of healthcare workers admitting to using unauthorized AI tools, and 41 percent saying they knew a colleague doing it. Other research found 58 percent of front-line clinicians had used something like ChatGPT for work in the past month, and that some of them had pasted in identifiable patient information.

Once protected health information goes into a consumer chatbot, it is out of your hands. It sits in a third-party system, nobody knows how long it stays there, and you cannot audit it or get it back. That is a HIPAA exposure nobody signed off on, and nobody is tracking. Healthcare breaches averaged more than \$ 7.4 million in 2025, and shadow AI is exactly the kind of slow leak that shows up long after the damage is done. Providing sanctioned tools, clear rules for patient data, and real training costs far less than the alternative. Banning AI outright mostly just drives it further underground.

Risk	Mitigation
Hallucination and fabrication	Human verification, retrieval grounding, task scoping, and disclosure of uncertainty. Human verification, retrieval grounding, task scoping, and disclosure of uncertainty.
Sycophancy	Prompt for the counterpoint; reward disagreement; keep a human deciding.
Bias and inequity	Audit data before the model; test subgroup performance; involve affected populations.
Opacity	Surround the model with accountability, monitoring, and a clear escalation path.
Model drift	Continuous monitoring against a baseline; predetermined change control; revalidation.
Misleading benchmarks	Validate on your own tasks and data, not vendor scores.
Security and shadow AI	Sanctioned tools, access controls, data policies, and staff education.

Why Imperfect AI Still Improves Healthcare

3.1 Where AI Already Outperforms

The honest claim is not that AI is better than people at most things. It is that in narrow, well-defined tasks, AI already matches or exceeds typical human performance, and that list is growing.

Autonomous diabetic retinopathy screening was the first of its kind authorized by the FDA, and in its **pivotal trial** it reached a sensitivity of 87.2 percent and a specificity of 90.7 percent in primary care offices, letting non-specialists screen patients who might otherwise go unscreened. In breast screening, the **MASAI randomized trial** found that AI support cut radiologist reading workload by about 44 percent while increasing cancer detection by roughly 29 percent, with no rise in false positives. In colonoscopy, a **meta-analysis of dozens of randomized trials** found an absolute increase of about 8 percent in adenoma detection with AI assistance.

In 2025, Microsoft reported that an **AI diagnostic orchestrator solved 85.5 percent of 304 of the hardest New England Journal of Medicine case records**, against roughly 20 percent for a panel of experienced physicians, and at lower simulated cost. The fine print matters, though. These were rare, difficult cases rather than the patients who fill a primary care schedule, the physicians worked alone without the references, colleagues, or tools they would normally reach for, and none of it has been tested in a live clinic. What the study shows is narrower than the headline, and still striking: on one bounded reasoning task, these systems now perform at a level that would have sounded far-fetched a few years ago.

The clearest near-term value is in documentation. Across **six United States health systems**, ambient AI scribes significantly reduced burnout, cognitive load, and documentation time. At **Mass General Brigham**, clinician burnout fell from 50.6 percent to 29.4 percent within six weeks of adoption. These are measured improvements in the parts of medicine that were breaking clinicians.

3.2 Enhancement, Not Replacement

Enhancement holds up better than replacement. AI raises the ceiling for experts and lifts the floor for everyone else. The American Medical Association chose the term augmented intelligence on purpose to signal a tool that supports the physician instead of replacing them. Clinicians seem to agree with their feet. Professional use among physicians hit **81 percent in 2026, up from 38 percent in 2023**, more than doubling in three years.

Scoring well on a benchmark is not the same as being useful in a clinic. A **2024 randomized trial in JAMA Network Open** gave physicians access to GPT-4 and found no real improvement in their diagnostic reasoning over the usual resources, even though the model did well on its own. Two things follow from that. The first is a caution about the evidence itself. GPT-4 is several generations old by today's standards, and the models have improved considerably since, which is a permanent feature of this field rather than a quirk of one study. Designing a trial, running it, and getting it through peer review takes years, while the technology turns over in months, so any published finding describes a version of the tool that has already been left behind.

The second holds up regardless of which model you test. Access alone changes nothing. Tool design, training, trust, and workflow are what turn raw capability into useful collaboration, and a strong model dropped into an unchanged workflow will disappoint every time. Enhancement comes from the system built around the model, not the model by itself.

The division of labor is what makes it work. The clinician keeps judgment, context, and accountability. The model handles volume, recall, and first drafts. That is where the evidence is strongest and where the risk is easiest to manage.

3.3 AI Reveals the Data and Process Problems You Already Had

When an AI pilot fails, the cause is usually not the model. It is the surrounding process. Requirements were unclear, the data was messy, the workflow was never designed for the tool, or no one owned the outcome. AI surfaces these problems because it cannot paper over them the way a motivated human can.

The sepsis model described earlier shows how this plays out. It was not rolled out blind. Epic built and validated it on data from other health systems before release. The trouble was that its performance did not carry over cleanly to every hospital, where patients, documentation habits, and workflows differed from the data it learned from. Then the data itself moved. When COVID-19 arrived, the model's alert volume across 24 US hospitals **rose 43 percent within three weeks** even as patient census fell by about a third, a textbook case of dataset shift. A model validated once, somewhere else, is not the same as a model monitored locally over time. That gap is a governance problem as much as a technical one.

AI works like a mirror in this respect. It shows you the bad data, the weak governance, and the missing standards that were already sitting there, and that alone is worth something before the model ever earns its keep. Organizations that get value from AI are the ones that take this as feedback instead of a source of embarrassment.

3.4 Every Transformative Medical Technology Was Imperfect At First

Medicine has a long record of taking powerful but poorly understood tools and making them safe through governance. The X-ray is the clearest case, and the lesson is not the one people usually draw from it. After Roentgen's discovery in 1895, clinical use spread within months, with no shielding, no dose limits, and no way to measure exposure. The harm that followed was not a necessary price of progress. It was the direct result of putting a technology into daily use with nobody monitoring it. Once **protection standards arrived in the 1930s**, the risk fell toward zero, and imaging became one of the most valuable tools in medicine. The technology did not become safe on its own. It became safe when someone built the controls around it.

Computer-aided detection for mammography makes the same point from the other direction. Cleared by the FDA in 1998, reimbursed by Medicare, and eventually used on most screening mammograms in the United States, it was found in a **study of more than 600,000 exams** not to improve accuracy on any measure, and possibly to reduce sensitivity for some radiologists. Widespread adoption is not evidence that something works. Validate first, then scale.



3.3 The Cost of Waiting, and the Shadow AI Already in Your Building

Waiting for perfect AI is itself a decision, and it is not free. Organizations that delay fall behind with slower execution, higher operating costs, and a widening gap against competitors that learned to deploy responsibly earlier. Perfectionism, in a moving field, is a competitive risk.

The idea that an organization can simply wait is also largely fiction. Clinicians are already using AI, with or without permission. As the shadow AI figures show, a large share of front-line clinicians report using consumer AI tools for work, sometimes with patient information attached. The question is not whether AI enters the organization. It has already entered. The only question is whether it enters governed or ungoverned.

→ **AI is already in your building.** What remains is whether it runs under governance or outside it.

Adopting Responsibly

The case for AI is not a case for ignoring its failures. Each risk below is real, carries particular weight in healthcare, and has a known control.

4.1 A Tiered Risk Model for Healthcare AI

Not all healthcare AI carries the same stakes. A tool that suggests scheduling and a tool that supports a cancer diagnosis should not face the same oversight. The sensible approach tiers use cases by consequence and autonomy, applying light governance to low-risk, low-autonomy uses and heavy scrutiny to high-risk, high-autonomy ones. This risk-based structure also underpins the **European Union AI Act**, which treats most medical AI as high-risk and attaches obligations accordingly. A workable tiering looks like this.



Tier	Typical uses	Oversight posture
 Low	Administrative drafting, scheduling, non-clinical summarization	Use policy, basic disclosure, periodic spot checks.
 Moderate	Documentation, triage support, retrieval that informs but does not decide	Human review, monitoring, defined escalation paths.
 High	Diagnostic, treatment, or autonomous clinical decisions	Formal validation, continuous monitoring, named accountability, regulatory pathway.

4.2 Governance Anchors

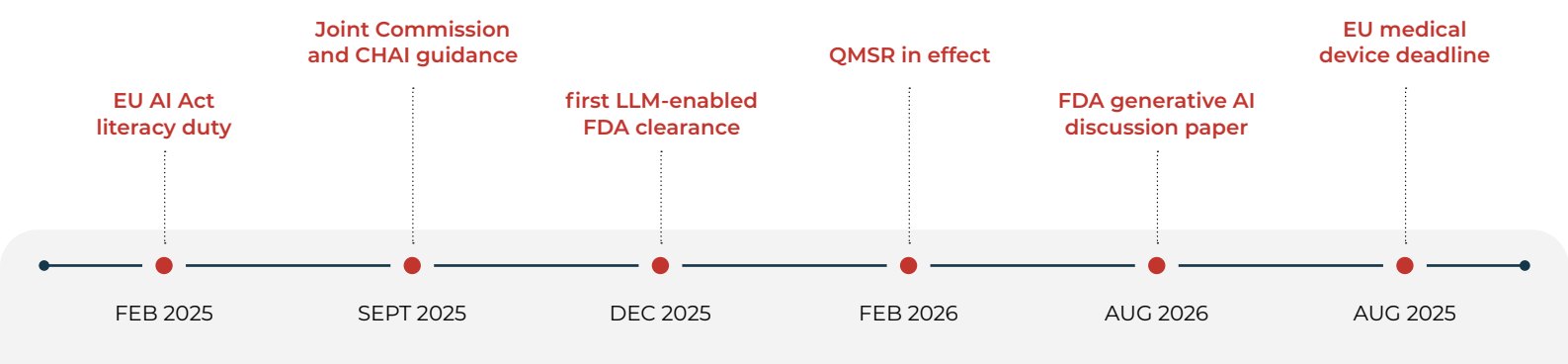
Organizations do not have to invent governance from scratch. In the United States, the **NIST AI Risk Management Framework** offers a voluntary structure built on four functions: to govern, map, measure, and manage. The FDA regulates **AI-enabled medical devices**, and its list had grown to roughly 1,450 authorizations by early 2026, about three-quarters of them in radiology. Almost all rely on conventional predictive models, and large language models have only just begun to cross the line. In December 2025, the FDA cleared **UpDoc**, a prescription app for managing insulin in adults with type 2 diabetes that uses a conversational AI agent to talk with patients and collect their data. The design is the interesting part. The language model handles the conversation, while the dosing decision stays inside fixed logic set by the patient’s clinician, and **the FDA’s review** requires future updates to preserve that deterministic core. It is a narrow, tightly controlled use rather than an AI physician, and it shows the pattern regulators are comfortable with today. The FDA’s guidance on **predetermined change control plans** governs model updates, **good machine learning practice** governs development, and the **Quality Management System Regulation**, effective February 2026, harmonizes United States quality requirements with the international ISO 13485 standard. The **ONC HTI-1 rule** adds transparency requirements for predictive algorithms in certified health IT, and HIPAA continues to govern patient data regardless of how software is built.

In August 2026, the FDA released **a discussion paper on regulating generative AI-enabled medical devices**, with comments open through October 19, 2026. It is not guidance and sets no binding policy, but it signals where the agency is heading. The FDA proposes evaluating these devices roughly as it evaluates physicians, through competency assessment: standardized benchmarking against safety behaviors, clinical

proficiency, and robustness across demographics and dialects, followed by clinical confirmation under real-world conditions and ongoing postmarket monitoring. The paper also raises the harder questions the technology forces, including open-ended inputs, variable outputs, and devices built on third-party foundation models that the model developer, not the device maker, can change underneath them. Anyone building generative AI into a regulated product should read it now and consider commenting, since the feedback collected here will shape the evidence standards reviewers apply later.

The most practical recent addition sits closer to the bedside. In September 2025, the **Joint Commission and the Coalition for Health AI released guidance on the responsible use of AI in healthcare**, the first in a planned series of playbooks leading to a voluntary AI certification available to more than 22,000 accredited organizations. It organizes self-governance around seven areas: AI policy and oversight structures, patient transparency, data protection, bias and risk assessment, education, ongoing monitoring, and voluntary blinded reporting of AI safety events so the field can learn from failures. For a United States health system, this is among the most usable starting points available today.

Internationally, the **EU AI Act** uses a risk-based approach in which most medical AI is high-risk, though its timeline is shifting. The AI-literacy obligation applies from August 2025, while a Digital Omnibus adopted in July 2026 deferred the heaviest high-risk obligations for AI embedded in medical devices to August 2028. The **World Health Organization guidance on large multi-modal models** adds further structure. United States federal regulation remains comparatively light-touch, which places more of the burden on internal governance.



Framework	Origin	What it does
NIST AI RMF	US, voluntary	Govern, map, measure, and manage functions for AI risk.
FDA (devices, PCCP, GMLP, QMSR)	US, mandatory for devices	Authorizes AI devices and governs updates, development, and quality systems.
ONC HTI-1	US, mandatory for certified health IT	Transparency requirements for predictive algorithms.
HIPAA	US, mandatory	Protects patient data regardless of how software is built.
Joint Commission and CHAI	US, voluntary, certification planned	Practical self-governance across seven areas, including safety-event reporting
EU AI Act	EU, mandatory	Risk-based obligations; medical-device high-risk deadline now 2028.
WHO LMM guidance	Global, advisory	Ethics and governance for large multi-modal models in health.

4.3 Proportional Oversight and Real Human Accountability

Human oversight is necessary but not sufficient on its own. Automation bias is well documented. When a decision aid is wrong, clinicians **often follow it anyway**, and the effect extends beyond novices. In one randomized study, even physicians trained in AI use produced worse diagnostic reasoning when

given deliberately flawed model advice. Oversight must be designed, not assumed. That means explicit escalation paths, confidence thresholds that trigger human review, and a named person accountable for the output, not simply a human somewhere in the loop.

4.4 Audit the Data Before the Model

The bias failures share a root cause in the data, so the most effective control sits upstream of the model. That means transparent data sourcing, dataset audits performed before model audits, and the affected patient populations represented in the process. It means naming the blind spots directly. Unpublished trials remove evidence the model will never learn from. Training data skewed toward a few academic centers in high-income countries will not represent every population the model meets. A model that works for 70 percent of patients is not impressive if the other 30 percent are the ones most likely to be harmed.

4.5 Transparency and Disclosure

Users deserve to know when AI is involved, what its limits are, and how confident it is. This is now partly a regulatory expectation. The [ONC HTI-1 rule](#) requires certified health IT to expose source attributes for predictive algorithms so that users can judge whether a tool is fair, appropriate, valid, effective, and safe. Transparency is not a courtesy. It is a precondition for the informed human judgment that responsible use depends on.



4.6 AI Literacy: Teaching People to Direct and Challenge AI

Users deserve to know when AI is involved, what its limits are, and how confident it is. This is now partly a regulatory expectation. The ONC HTI-1 rule requires certified health IT to expose source attributes for predictive algorithms so that users can judge whether a tool is fair, appropriate, valid, effective, and safe. Transparency is not a courtesy. It is a precondition for the informed human judgment that responsible use depends on.

4.7 Protecting Clinical Judgment Over Time

Skills that are not used tend to fade. A [2025 study in The Lancet Gastroenterology and Hepatology](#) provided the first real-world evidence of clinical de-skilling, finding that after routine exposure to AI assistance, endoscopists' adenoma detection rate on unassisted procedures fell from 28.4 percent to 22.4 percent. Physicians are aware of the danger, and the concern is sharpest among those earliest in their careers. The safeguard is deliberate. Preserve unassisted practice, monitor core skills over time, and treat AI as augmentation of clinical judgment rather than a replacement for it.

A Practical Path to Adoption

The principles above reduce to a four-step sequence any healthcare organization can follow, regardless of size or resource level.

5.1 Define the Use Case and Its Risk Tier

Name the problem precisely and place it on the risk-tier model. High-consequence, high-autonomy uses earn heavier oversight from the outset, and low-stakes uses should not be burdened with processes they do not need. A clear use case also makes success measurable, which the next step depends on.

5.3 Pilot With Human Oversight and Workflow Redesign

Pilot in a bounded setting with real oversight, and expect to redesign the workflow rather than simply insert the tool. Most pilots fail on process, not on the model, so the redesign is the work. Define who reviews the output, when the tool escalates to a human, and who is accountable when it is wrong, before the pilot goes live.



5.2 Establish Your Current Baseline

Measure the human-only workflow before changing it. An organization cannot demonstrate that AI beats what it does today if it never measured what it does today, and that baseline is what every later claim is judged against. Baseline metrics should include quality, safety, time, cost, and subgroup performance, so that equity is measured from the start rather than discovered later.

5.4 Monitor, Then Scale

Watch for drift, subgroup performance, and override rates in production. Scale only what continues to beat the baseline under real conditions, and keep monitoring after scaling, because a model that passed yesterday can degrade tomorrow. Voluntary reporting of safety events, as the Joint Commission and CHAI framework encourages, lets the wider field learn from what goes wrong.



Conclusion:

The Right Comparison Changes Everything



The question was never whether AI is perfect. It is whether AI, deployed well, beats what an organization is already doing. Healthcare has always run on imperfect people, imperfect data, and imperfect tools, and it has always governed those imperfections rather than waited for them to disappear. AI belongs in that same tradition.

The evidence supports a measured optimism. In narrow tasks, AI already outperforms typical human effort. It reduces the documentation burden that drives burnout. It exposes the process and data problems an organization needed to fix anyway. Its risks, including bias, hallucination, opacity, and de-skilling, are real, but each has a known control. Governance, not hesitation, ties the controls together.

→ **Healthcare does not need perfect AI.** It needs AI that is useful, measured, governed, monitored, and better than the status quo.



Why Estenda Solutions

The Gap Between Buying AI and Deploying It Well

Most organizations do not struggle to buy AI. They struggle to deploy it responsibly, in a way that is validated, governed, and genuinely better than what came before. That gap, between a capable model and a trustworthy clinical system, is where value is won or lost, and it is where most of the work actually lives.

What Estenda Does

Estenda helps medtech, life sciences, healthcare, and digital health organizations turn AI from a promising tool into a governed, production-ready capability. Our work spans the areas this paper describes.



AI and machine learning

designing and validating models that hold up under real clinical conditions.



Custom software development

building the systems that carry AI into production safely and maintainably.



Data analytics

auditing and preparing the data that determines whether a model helps or harms.



Implementation and support

redesigning workflows and establishing the oversight responsible use requires.



Healthcare, medical, and software research

grounding decisions in evidence rather than hype.



Strategy

prioritizing use cases by risk and value so effort goes where it pays off.

What sets our approach apart is consistent across every engagement. We increase value quickly, we reduce risk, and we bring empathy to the clinicians and patients who live with what we build.



How Engagements Start

Every engagement begins with a conversation, not a commitment. We offer a free 30-minute consultation with no obligation to understand the problem, the risk, and whether there is a responsible path worth pursuing.

Contact



info@estenda.com



estenda.com

→ **AI does not have to be perfect to be worth adopting.**

It has to be governed, measured, and better than today. That is work worth doing well.

References

- 01 Newman-Toker DE, et al. Burden of serious harms from diagnostic error in the USA. *BMJ Quality & Safety*, 2024. <https://doi.org/10.1136/bmjqs-2021-014130>
- 02 Singh H, et al. The frequency of diagnostic errors in outpatient care. *BMJ Quality & Safety*, 2014. <https://doi.org/10.1136/bmjqs-2013-002627>
- 03 Van Such M, et al. Extent of diagnostic agreement among medical referrals. *Journal of Evaluation in Clinical Practice*, 2017. <https://doi.org/10.1111/jep.12747>
- 04 Sinsky C, et al. Allocation of physician time in ambulatory practice: a time and motion study. *Annals of Internal Medicine*, 2016. <https://doi.org/10.7326/M16-0961>
- 05 Omar M, et al. Large language models are vulnerable to adversarial hallucination in clinical decision support. *Communications Medicine*, 2025. <https://doi.org/10.1038/s43856-025-01021-3>
- 06 Asgari E, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine*, 2025. <https://doi.org/10.1038/s41746-025-01670-7>
- 07 Doshi AR, Hauser OP. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 2024. <https://doi.org/10.1126/sciadv.adn5290>
- 08 Obermeyer Z, et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 2019. <https://doi.org/10.1126/science.aax2342>
- 09 Sjoding MW, et al. Racial bias in pulse oximetry measurement. *New England Journal of Medicine*, 2020. <https://pubmed.ncbi.nlm.nih.gov/33326721/>
- 10 Andrist E, et al. Association of race with pulse oximetry accuracy in hospitalized children. *JAMA Network Open*, 2022. <https://doi.org/10.1001/jamanetworkopen.2022.4584>
- 11 Seyyed-Kalantari L, et al. Underdiagnosis bias of AI algorithms applied to chest radiographs in under-served populations. *Nature Medicine*, 2021. <https://doi.org/10.1038/s41591-021-01595-0>
- 12 Daneshjou R, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 2022. <https://doi.org/10.1126/sciadv.abq6147>
- 13 Wong A, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 2021. <https://doi.org/10.1001/jamainternmed.2021.2626>
- 14 Liang S, Garg S, Moghaddam RZ. The SWE-bench illusion: when state-of-the-art LLMs remember instead of reason. *arXiv*, 2025. <https://arxiv.org/abs/2506.12286>
- 15 Wolters Kluwer Health. Survey finds broad presence of unsanctioned AI tools in hospitals and health systems, 2025. <https://www.wolterskluwer.com/en/news/wolters-kluwer-survey-finds-broad-presence-of-unsanctioned-ai-tools-in-hospitals-and-health-systems>
- 16 Black Book Market Research, reported in Today's Hospitalist. Shadow AI use among front-line clinicians, 2025. <https://todayshospitalist.com/are-you-using-shadow-ai-in-your-practice/>
- 17 Abramoff MD, et al. Pivotal trial of an autonomous AI-based diagnostic system for diabetic retinopathy. *npj Digital Medicine*, 2018. <https://doi.org/10.1038/s41746-018-0040-6>
- 18 Lang K, et al. AI-supported screen reading versus standard double reading (MASAI trial). *Lancet Oncology*, 2023. [https://doi.org/10.1016/S1470-2045\(23\)00298-X](https://doi.org/10.1016/S1470-2045(23)00298-X)
- 19 Soleymanjahi S, et al. AI-assisted colonoscopy for polyp detection: systematic review and meta-analysis. *Annals of Internal Medicine*, 2024. <https://doi.org/10.7326/M24-0885>
- 20 Nori H, et al. Sequential diagnosis with language models (MAI-DxO). *Microsoft AI / arXiv*, 2025. <https://arxiv.org/abs/2506.22405>
- 21 Goh E, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open*, 2024. <https://doi.org/10.1001/jamanetworkopen.2024.40969>
- 22 Use of ambient AI scribes to reduce administrative burden and professional burnout across six health systems. *JAMA Network Open*, 2025. <https://pubmed.ncbi.nlm.nih.gov/41037268/>
- 23 Mass General Brigham. Ambient documentation technologies reduce physician burnout, 2025. <https://www.massgeneralbrigham.org/en/about/newsroom/press-releases/ambient-documentation-technologies-reduce-physician-burnout>

- 24 American Medical Association. Physician Survey on Augmented Intelligence, 2026. <https://www.ama-assn.org/practice-management/digital-health/physician-survey-augmented-intelligence>
- 25 American Medical Association. Augmented intelligence research, 2024 wave. <https://www.ama-assn.org/practice-management/digital-health/2-3-physicians-are-using-health-ai-78-2023>
- 26 Early X-ray workers: numbers, risk, and radiation harm. European Society of Radiology / PMC, 2015. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4805624/>
- 27 Lehman CD, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Internal Medicine, 2015. <https://doi.org/10.1001/jamainternmed.2015.5231>
- 28 NIST. AI Risk Management Framework (AI RMF 1.0), 2023. <https://www.nist.gov/itl/ai-risk-management-framework>
- 29 US FDA. Artificial intelligence-enabled medical devices (authorized-device list). <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices>
- 30 US FDA. Predetermined change control plans for AI-enabled device software functions, 2024. <https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles>
- 31 US FDA. Good machine learning practice for medical device development: guiding principles, 2021. <https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles>
- 32 US FDA. Quality Management System Regulation (QMSR), effective February 2, 2026. <https://www.fda.gov/medical-devices/postmarket-requirements-devices/quality-management-system-regulation-qmsr>
- 33 HHS / ONC. HTI-1 final rule: algorithm transparency for predictive decision support interventions, 2024. <https://www.healthit.gov/regulations/hti-rules/hti-1-final-rule/>
- 34 Joint Commission and Coalition for Health AI. Guidance on the responsible use of AI in healthcare, 2025. <https://www.jointcommission.org/en-us/knowledge-library/news/2025-09-jc-and-chai-release-initial-guidance-to-support-responsible-ai-adoption>
- 35 European Union. Artificial Intelligence Act, Regulation (EU) 2024/1689. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- 36 World Health Organization. Ethics and governance of AI for health: guidance on large multi-modal models, 2024. <https://www.who.int/publications/i/item/9789240084759>
- 37 Gaube S, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. npj Digital Medicine, 2021. <https://doi.org/10.1038/s41746-021-00385-9>
- 38 Budzyn K, et al. Endoscopist de-skilling risk after exposure to AI in colonoscopy. Lancet Gastroenterology & Hepatology, 2025. [https://doi.org/10.1016/S2468-1253\(25\)00133-5](https://doi.org/10.1016/S2468-1253(25)00133-5)
- 39 Zhou L, et al. Analysis of errors in dictated clinical documents assisted by speech recognition software and professional transcriptionists. JAMA Network Open, 2018. <https://doi.org/10.1001/jamanetworkopen.2018.0530>
- 40 Wong A, et al. Quantification of sepsis model alerts in 24 US hospitals before and during the COVID-19 pandemic. JAMA Network Open, 2021. <https://doi.org/10.1001/jamanetworkopen.2021.35286>
- 41 US FDA. 510(k) premarket notification K253281, UpDoc, cleared December 23, 2025. <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/pmn.cfm?ID=K253281>
- 42 Innolitics. First FDA-cleared AI agent and LLM-enabled device confirmed, 2026. <https://innolitics.com/articles/updoc-fda-cleared-ai-agent/>