WHITE PAPER

The Insurability Problem in Healthcare AI

A Standards-Based Framework for Underwriting Risk Assessment

Jennifer Shannon, MD Glacis Technologies
Sarah Gebauer, MD Validara Health

March 2026

Abstract

A single failure in a clinical AI system can generate claims across product liability, professional malpractice, and enterprise risk policies simultaneously — separate carriers, separate actuarial models, none designed for the technology at issue. The data that would ordinarily resolve this problem does not yet exist. Healthcare AI has been deployed at scale for fewer than five years; usable loss experience will not stabilize for at least another decade. Carriers today must decide whether and how to insure these systems without the inputs their underwriting models require.

This paper proposes a framework for that interim: Standards-Proof, a three-layer methodology that repurposes the international medical device standards healthcare AI companies already implement for regulatory clearance. Layer 1 evaluates risk management depth across seven standards (ISO 14971, IEC 62304, ISO 13485, and others), scoring implementation quality rather than binary compliance. Layer 2 adds healthcare-specific validation — clinical studies, automation bias testing, supervision model assessment — tiered by whether the product directly informs care decisions, enters the clinical record, or operates administratively. Layer 3 requires runtime verification: tamper-evident audit records, adversarial stress testing, and continuous monitoring that keeps underwriting assumptions aligned with operational behavior.

Four case studies illustrate the methodology across patient-facing chatbot, prior authorization, diagnostic imaging, and clinical documentation AI. In each case the framework surfaced risks that compliance certification alone would not — a prior authorization system classified as "administrative" carrying direct clinical risk with embedded demographic bias, a documentation AI reporting 95% accuracy generating 120 clinically consequential hallucinations per month — and in two cases identified measurable risk reduction that justified more favorable terms.

The addressable premium pool for healthcare AI insurance is measured in billions annually. The carriers and AI companies that develop structured underwriting capability now will define the market's standards, accumulate its first loss experience, and compound those advantages as the regulatory environment tightens.

Disclosure: The authors have commercial interests in companies that provide healthcare AI risk assessment and runtime verification services. This paper describes capability requirements that the authors' companies, among others, address. The framework is designed to be vendor-neutral and implementable through multiple providers.

Executive Summary

The rapid deployment of healthcare AI, a market projected to reach $500 billion or more by the mid-2030s, has created an urgent insurance challenge. Healthcare AI uniquely converges product liability, professional malpractice, institutional negligence, and regulatory compliance risks, a combination that fragments across multiple insurance policies. Yet no specialized underwriting framework exists to quantify or price healthcare AI's unique risks.

Traditional insurance frameworks cannot adequately price healthcare AI risk because they were designed for fundamentally different exposures. Medical device coverage relies on regulatory classifications and actuarial history that do not exist for AI. Professional liability measures individual physician risk, not AI-amplified caseloads and AI-assisted workflows. Enterprise policies assume bounded failures rather than population-scale cascade events. The claims data that would ordinarily inform underwriting decisions will not stabilize for an undetermined amount of time.

During this actuarial gap, insurers face three unacceptable paths:

- Denying coverage entirely and stifling adoption
- Pricing blindly with premiums that either exclude startups or catastrophically underprice risk
- Waiting for loss experience to accumulate while patients remain exposed to inadequately validated systems

This paper proposes a fourth path. The Standards-Proof framework is a three-layer underwriting methodology that builds on proven international medical device risk management standards, extends them with AI-specific adversarial testing, adds healthcare-specific clinical validation requirements, and introduces runtime enforcement and verification to bridge the gap between vendor claims and tamper-evident evidence. Rather than prescriptive scoring, we propose a framework that produces a structured evidence assessment: the same documentation that demonstrates safety to regulators can demonstrate to an insurer what they need to know to make a coverage decision.

This framework approaches AI risk and insurability from the perspective of clinical medicine and AI governance — where the risks originate and where the controls either hold or fail — but does not attempt actuarial analysis. It addresses the clinical and technical risk factors that actuarial models will need to account for as healthcare AI becomes insurable.

Why Healthcare AI is Currently Uninsurable

The Convergence Problem

Healthcare AI is unique among insured technology products in one fundamental respect: it does not create risk within a single insurance domain. A single deployment event can generate a product liability claim against the vendor, a malpractice claim against the supervising physician, and an institutional

liability claim against the health system, all simultaneously, all arising from the same failure. Each domain has its own policyholder, its own carrier, its own claims logic, and its own actuarial tradition. Healthcare AI breaks those silos.

One Event = Three Simultaneous Claims

No existing healthcare AI insurance framework is designed for this convergence.

CLINICAL EVENT

An AI clinical documentation system fabricates a medication allergy during a hospital admission. A covering physician, relying on the AI-generated record, avoids first-line antibiotics for a wound infection. The patient develops sepsis requiring ICU transfer.

Product Liability	Professional Liability	Enterprise Risk
POLICYHOLDER **CARRIER**	**POLICYHOLDER** **CARRIER**	**POLICYHOLDER** **CARRIER**
Alleged algorithm defect: hallucinated a medication allergy absent from the encounter, with no warning of hallucination risk in safety-critical fields.	Alleged negligence: admitting physician cosigned AI note without verifying the allergy; covering physician relied on it without independent confirmation.	Alleged governance failure: no hallucination controls on safety-critical fields, inadequate cosigning workflows, potential False Claims Act liability.

Figure 1. A single clinical AI failure generates claims across three independent insurance domains.

Case Scenario: Clinical Ambient Documentation

A physician uses an AI scribe for a hospital admission. The system accurately captures the chief complaint, history, and physical exam, but fabricates a penicillin allergy that the patient never mentioned. The note is cosigned during a busy overnight shift. Three days later, the patient develops a postoperative wound infection. The covering physician, relying on the AI-generated note, avoids first-line penicillin-based antibiotics and selects a broader-spectrum alternative. The infection worsens under suboptimal treatment. By the time the allergy error is discovered, the patient has developed sepsis requiring ICU transfer.

In this single event, product liability attaches to the vendor whose system generated the hallucination. Professional liability attaches to two physicians — the one who cosigned the note containing fabricated content, and the covering physician who relied on it. The deploying institution faces enterprise risk from HIPAA breach exposure if the scribe's redaction controls failed, False Claims Act liability if the fabricated documentation inflated billing codes, and reputational harm when the pattern surfaces in discovery. No existing insurance framework addresses this convergence systematically.

Product Liability and the Actuarial Data Gap

Product liability covers claims arising from failures in the AI product itself: algorithmic errors, model drift, adversarial vulnerabilities, hallucinations, and design defects that cause or contribute to patient harm. The policyholder is the company that built the system.

The problem is not that product liability insurance does not exist. It is that existing underwriting has no adequate basis for pricing it. Medical device product liability insurance was built on four pillars: regulatory classification establishing baseline risk tiers, actuarial tables derived from decades of claims history, device-specific risk profiles identifying known failure modes, and proven engineering controls with documented effectiveness. For a cardiac pacemaker, insurers have decades of claims history establishing failure rates by device type, severity distributions for lead fracture litigation, and documented relationships between specific design controls and failure reduction. The actuarial infrastructure exists because the device has been implanted at scale long enough to generate it.

For healthcare AI, none of these pillars exist in usable form. Most clinical AI systems clear through pathways equivalent to Class II designation, but that single classification now spans everything from appointment schedulers to critical care decision support tools with fundamentally different harm potential. Claims history does not yet exist in any meaningful sense. Healthcare AI has deployed at scale for fewer than five years, the first major wrongful death lawsuits remain in discovery, and loss data may not stabilize for another decade.

The product liability exposure created by AI-enhanced medical devices is already moving from theoretical to documented. A Reuters investigation published in February 2026 illustrates the dynamic directly: after Johnson & Johnson's Acclarent unit added a machine-learning algorithm to its TruDi Navigation System, a surgical navigation device used in sinus procedures, the FDA received reports of at least 100 malfunctions and adverse events, compared to seven malfunction and one injury report in the three years the device had been on the market before AI was added. At least ten patients were allegedly injured between late 2021 and November 2025, including two stroke victims who filed lawsuits in Texas alleging the AI contributed to their injuries; one suit contends that "the product was arguably safer before integrating changes in the software to incorporate artificial intelligence than after the software modifications were implemented." A separate research letter published in JAMA Health Forum found that 60 FDA-authorized medical devices using AI were linked to 182 product recalls, with 43% of those recalls occurring less than a year after authorization — approximately twice the recall rate of comparable non-AI devices. The TruDi cases illustrate precisely the product liability gap this framework is designed to address: when AI is added to an existing device, the actuarial baseline established for the pre-AI product becomes unreliable, the failure modes shift in ways that existing risk profiles do not capture, and the question of whether the AI component caused or exacerbated the harm with no established underwriting methodology to price that uncertainty in advance.

The most consequential absence is in device-specific risk profiles. Adversarial perturbations can cause diagnostic AI to systematically miss disease through image modifications invisible to human reviewers. Prompt injection attacks can cause prior authorization systems to approve medically unnecessary procedures. Clinical documentation AI can fabricate medication allergies that become part of the legal medical record. These are documented failure modes demonstrated in controlled testing and, in several cases, in production deployments. They have no direct comparisons in the mechanical device history from which existing risk profiles were derived.

The financial exposure from these failures is substantial. Defense costs for AI-related medical malpractice litigation are estimated at $500,000 to $2 million per claim, higher than traditional med-mal due to

the technical discovery required. When an AI system affects decisions at population scale, even modest per-claimant damages create enormous aggregate exposure.

Professional Liability and the Attribution Gap

Currently much of the liability for AI falls onto the physician overseeing the patient's care. Professional liability covers malpractice claims arising when a clinician's decision, informed or influenced by an AI tool, causes patient harm. The policyholder is usually the physician, physician group, or health system.

The problem is that the conditions under which a human in the loop operates and are assumed — such as a physician with adequate time, adequate training, and adequate contextual information to evaluate the AI's output critically — are not the conditions under which clinical AI always actually operates.

AI fundamentally alters the risk calculus on which professional liability pricing is built. Professional liability prices on specialty classification, individual claims history, and practice volume. Under current models, a radiologist reading eighty scans daily and a radiologist reading five hundred scans daily with AI-assisted triage receive the same malpractice premium structure. The profession has not built pricing mechanisms that account for AI-expanded caseloads, novel supervision models, or the liability questions that arise when a physician acts on an AI recommendation that proves incorrect.

The liability attribution problem at the center of this gap is unresolved. When AI recommends a treatment and a physician follows that recommendation and harm results, the question of whether the physician exercised appropriate clinical judgment, whether the vendor delivered a product fit for purpose, or whether the deploying institution fulfilled its duty of care does not have settled legal precedent in most jurisdictions.

In practice, "responding appropriately" means the physician must independently evaluate each AI output using their own clinical knowledge, identify cases where the AI's recommendation conflicts with the clinical picture, and override or reject AI outputs when their clinical judgment warrants it — even when the AI has a high documented accuracy rate and even when institutional workflows create pressure to accept AI recommendations efficiently.

The Federation of State Medical Boards issued recommendations in May 2024 which state that physicians accept responsibility for responding appropriately to AI recommendations and must uphold the standard of care. The American Law Institute approved its first restatement of medical malpractice law in May 2024, shifting the legal standard from what most physicians currently do in practice to what the best available evidence says they should do.

Enterprise Risk: Visibility and Boundedness

Enterprise risk covers the institutional consequences of AI failures: regulatory penalties, enforcement actions, litigation costs, reputational damage, and operational disruption. The carrier is typically a general liability or cyber liability insurer.

The problem is that enterprise risk from healthcare AI is both more visible and more difficult to bound than enterprise risk from most other technology deployments. A software bug in a customer relationship management system might affect one company's customer interactions, but a failed safety

control in a healthcare AI system deployed across multiple health systems can affect thousands of patients simultaneously, generating class-action exposure that scales with deployment volume.

The litigation environment illustrates this directly. In Estate of Gene B. Lokken v. UnitedHealth Group, plaintiffs allege that UnitedHealthcare's AI tool nH Predict denied Medicare Advantage coverage for post-acute care with a 90% error rate on appealed denials. UnitedHealth disputed both that nH Predict made coverage determinations and that appeal reversal rate equals error rate. Judge Tunheim's February 2025 ruling used "futile" as a legal standard for waiving exhaustion requirements (procedural), not a merits finding on AI denial accuracy. In Kisting-Leung v. Cigna Corp., a California court partially allowed a class action over Cigna's PxDx algorithm after reporting documented more than 300,000 denials over two months with physician reviews averaging 1.2 seconds per claim.

The financial scale is instructive for underwriters. In the Cigna matter, 300,000 denials over two months implies an annual run rate of 1.8 million affected decisions. If even a fraction result in compensable harm, delayed treatment, disease progression, preventable hospitalization, the aggregate exposure dwarfs typical technology E&O claims. HIPAA violations carry penalties of $100 to $50,000 per violation, up to $1.5 million annually in Tier 4. False Claims Act violations carry treble damages.

Federal enforcement is intensifying independently of the litigation trajectory. The Department of Justice subpoenaed pharmaceutical and digital health companies in 2024 over generative AI in electronic medical records. The False Claims Act has become an active vehicle for investigating AI-generated documentation that inflates billing. A Conference Board and ESGAUGE analysis of S&P 500 filings found that 72% of companies now disclose at least one material AI risk, up from 12% in 2023; reputational risk is the most frequently cited specific category, disclosed by 38% of firms.

Privacy and data security risks add a further dimension. Healthcare AI systems process protected health information under HIPAA. For example:

- A clinical ambient scribe with "guaranteed PHI redaction before cloud processing" may have a redaction service that silently fails for a subset of encounters when audio exceeds specific thresholds, transmitting patient names and sensitive health information unredacted to a cloud-based language model. Such failures have persisted for months before accidental discovery during unrelated audits, resulting in breach notifications, attorney general investigations, and class action litigation.
- The Sharp Healthcare class action filed in November 2025 illustrates how the deployment of clinical documentation AI exposes health systems to enterprise liability that existing insurance frameworks are not equipped to price. According to the complaint, Sharp deployed an AI-powered ambient clinical documentation tool (Abridge) that automatically recorded doctor-patient conversations and generated draft notes for the electronic health record without obtaining all-party consent as required under California's strict wiretapping law, which provides statutory penalties of $5,000 per violation, per call, per recording.

Three Pillars Collapsing at Once

One event in healthcare AI disrupts every input that existing underwriting was built on.

Product Liability The Actuarial Gap	Professional Liability The Attribution Gap	Enterprise Risk The Visibility and Boundedness Gap
TRADITIONAL INPUTS	**TRADITIONAL INPUTS**	**TRADITIONAL INPUTS**
Regulatory classification	Standard of care precedent	Defined breach typologies
Device-specific risk profiles	Human error typologies	Bounded network perimeter
Decades of claims history	Peer consensus on practice	Known threat actors
WHY IT BREAKS	**WHY IT BREAKS**	**WHY IT BREAKS**
Class II now spans schedulers to critical care AI	AI acts autonomously, blurring human vs. machine error	AI failure can be subtle and systemic, not a discrete breach
Claims history does not exist	Standard of care for AI reliance is undefined	Perimeter includes external data dependencies
AI failure modes have no mechanical device analogs	No peer consensus and no legal precedent	Threat actors include non-malicious data shifts

Figure 2. Each insurance domain's traditional underwriting inputs are disrupted by healthcare AI.

Three Paths of Market Dysfunction

In the absence of systematic risk assessment, three paths exist, and all produce dysfunction:

- **Coverage denial.** Carriers decline to cover AI tools due to insufficient actuarial basis. AI companies deploy uninsured or self-insured. Health systems refuse to contract without insurance, stifling adoption. Uninsured deployment removes the oversight function that insurance provides.
- **Unreasonably high premiums.** Carriers price conservatively, making insurance available only to well-funded companies. Startups face market exclusion. Consolidation reduces innovation. Health systems absorb premium costs passed through by vendors.
- **Unrealistically low premiums.** Carriers underprice due to insufficient understanding of AI risk. Catastrophic losses follow when claims materialize. Market exit returns the industry to coverage denial.

Standards-based underwriting provides the fourth path: the only systematic alternative during this data gap.

The Market Opportunity for Insurers

The Premium Pool

The healthcare AI market is projected to reach $500 billion or more by the mid-2030s, with estimates from major research firms (Precedence Research, Grand View Research) clustering between $300 billion and $600 billion depending on scope. If insurance costs represent even 1–2% of AI company revenue consistent with typical technology E&O and product liability premium-to-revenue ratios the addressable premium pool is measured in billions of dollars annually.

Healthcare AI adoption is accelerating at a pace that outstrips most specialty insurance markets. As of mid-2025, more than 1,250 AI-enabled devices had received FDA clearance, a number that had more than doubled since 2022, and health systems are deploying AI across clinical documentation, diagnostic imaging, clinical decision support, prior authorization, and revenue cycle management simultaneously. Each of these deployment categories creates liability exposure that sits at the intersection of product liability, professional malpractice, and institutional negligence in ways that existing policy forms were not written to address.

Insurance as a Procurement Requirement

As health system procurement evolves to reflect the convergence of AI risk and institutional liability, insurance is moving from a negotiated exception to a standard contractual term. Healthcare providers are increasingly requiring AI vendors to maintain specified minimum coverage across multiple policy types: cyber liability, technology errors and omissions, professional liability, and commercial general liability, with tiered structures that impose higher limits for clinical applications than for administrative ones.

This creates a market dynamic in which insurance is not merely a risk transfer mechanism but a market access requirement. AI companies that cannot obtain adequate coverage face exclusion from the largest and most desirable customer segments. Insurers who develop the capability to underwrite healthcare AI are not waiting for demand to materialize — they are meeting demand that already exists and is growing.

First-Mover Dynamics

Several carriers have begun positioning in adjacent markets. Munich Re's aiSure program addresses AI performance guarantees. Beazley and Chubb have partnered with Google Cloud on AI-augmented cyber coverage. Coalition and CFC offer technology E&O products that touch healthcare AI peripherally. The Artificial Intelligence Underwriting Company's AIUC-1 framework is structured around standards, independent audits, and insurance tied to audit outcomes, representing a parallel effort to build confidence infrastructure for AI agents across enterprise contexts; it addresses security, reliability,

and governance risks that apply broadly. But none of these programs addresses the specific convergence of product liability, professional liability, and enterprise risk that healthcare AI creates.

The carrier that develops specialized healthcare AI underwriting capability first will capture three advantages: premium volume from a growing market, the data advantage that comes from writing early policies (claims experience that informs future pricing), and influence over the emerging standards that govern insurability. These advantages compound. Early entrants will understand healthcare AI risk better than late entrants because they will have priced it, defended claims against it, and learned from the gap between their underwriting assumptions and actual loss experience.

Regulatory Tailwinds

The regulatory environment is creating demand for exactly the kind of structured risk assessment that this framework provides. The National Association of Insurance Commissioners (NAIC) approved a Model Bulletin on the Use of Artificial Intelligence Systems by Insurers in December 2023, now adopted by 24 states as of mid-2025 per Holland & Knight analysis. The NAIC itself stated in December 2025 that over half of all states had adopted this or similar guidance. While the bulletin primarily governs insurers' own use of AI in underwriting, pricing, and claims, it establishes the regulatory vocabulary and governance expectations that healthcare AI underwriting will increasingly need to satisfy: documented AI programs, bias testing, transparent governance, and risk management controls. These are the same categories of evidence the Standards-Proof framework generates.

State-level AI regulation is proliferating. Colorado's AI Act, multiple state bills modeled on the EU AI Act, and sector-specific regulations targeting healthcare AI are creating compliance requirements that insurers can help their policyholders meet. The EU AI Act's classification of healthcare AI systems as "high risk" creates parallel demand in international markets. Carriers with healthcare AI underwriting expertise are positioned to serve a global market, not merely a domestic one.

The FDA's evolving guidance on AI/ML devices, expected to finalize in the 2026–2028 timeframe, will further standardize the evidence requirements that healthcare AI companies must produce. As these requirements crystallize, the alignment between regulatory evidence and underwriting evidence will tighten, making standards-based underwriting more efficient and more defensible with each regulatory development.

The Case for Standards-Based Underwriting

The argument for basing underwriting on existing international standards rather than developing purpose-built criteria rests on three pillars, each of which addresses a structural problem that alternative approaches cannot solve.

Proven at Scale

The standards on which this framework builds ISO 14971, IEC 62304, ISO 13485, ISO/IEC 23894, IEC 62366-1, and IEC 80001-1 are used by more than 10,000 medical device manufacturers globally and

have generated decades of implementation experience that reveals both their strengths and limitations. A specialized framework developed from scratch would lack this track record entirely, and insurers would be asked to price risk against criteria that have never been tested in practice.

Regulatory Alignment

These standards are already required or referenced by the regulatory bodies that govern healthcare AI. The FDA recognizes them as consensus standards. The EU requires demonstration of conformity to harmonized standards as the basis for CE marking. Health Canada, the Therapeutic Goods Administration in Australia, and the PMDA in Japan all reference or require subsets of them. This means healthcare AI companies deploying in these markets must already be implementing these frameworks, which makes standards-based underwriting assessable and enforceable without imposing requirements that fall outside existing industry practice.

Evidence Trail Generation

Most consequentially for the insurance question, these standards create the documentation and evidence trail that underwriting actually needs. Risk management files, software development records, usability engineering files, and clinical validation studies are all artifacts these standards require manufacturers to produce. The underwriting question is not whether this evidence exists but whether its quality is sufficient to support responsible pricing.

Why Alternative Approaches Fail

Building from scratch produces a framework with no implementation track record, no regulatory recognition, and no overlap with the evidence that companies already generate for regulatory clearance. It adds cost and complexity without leveraging decades of accumulated understanding about how risk management frameworks function and fail in practice.

Binary compliance assessment certified or not certified, compliant or not compliant produces no useful differentiation between products. ISO 14971 requires hazard identification but does not specify depth. A company with minimal compliance that lists fifteen generic hazards receives the same binary assessment as one with detailed analysis identifying dozens of specific clinical failure modes. The standards tell you whether a company has done something; they do not tell you whether what the company has done is adequate.

Waiting for actuarial data produces the market dysfunction described above. Claims data will not stabilize for ten to fifteen years. During that period, health systems will either refuse to adopt AI products that lack coverage or accept products from vendors with inadequately priced policies. Neither outcome serves patient safety or the long-term interests of the insurance market.

Single-standard approaches typically ISO 14971 alone treat one standard as sufficient. But ISO 14971 does not specify AI-specific hazard identification, clinical validation requirements, adversarial testing scope, or healthcare equity thresholds. A company can achieve full compliance while addressing none of the risks that actually distinguish a safe healthcare AI product from an unsafe one.

The Standards-Proof framework avoids each of these failures by starting with proven frameworks and extending them only where the evidence record shows they are insufficient.

Why Every Alternative Fails

The case for Standards-Proof is structural: every alternative approach fails for reasons traceable to the specific nature of healthcare AI risk.

Approach	Result	Why
Purpose-Built Framework (from no existing standards)	FAIL	No implementation track record. Not recognized by any regulatory body. Evidence trail doesn't overlap with regulatory clearance — adds cost without benefit. No decades of accumulated understanding of how risk management frameworks function and fail.
Binary Compliance Check (certified/not)	FAIL	Cannot differentiate implementation depth. ISO 14971 doesn't specify how many hazards to identify or how granular the analysis should be. A 15-generic-hazard file receives the same treatment as a 42-specific-hazard analysis.
Wait for Actuarial Data	FAIL	10–15 year wait produces either refusal to adopt (uninsured, unmonitored deployment) or acceptance of products from vendors who found coverage through policies that don't adequately price risk. Neither ensures patient safety.
Single-Standard Approach (ISO 14971 alone)	FAIL	ISO 14971 alone does not specify AI hazard identification, clinical validation statistics, adversarial testing scope, or equity thresholds. Full compliance possible while not addressing any of the risks that differentiate safe from unsafe healthcare AI.
Standards-Proof	WORKS	Starts with frameworks proven at scale across 10,000+ manufacturers. Evidence quality scoring replaces binary compliance. Designed to evolve toward actuarial pricing as claims data accumulates. Creates economic incentives that reward actual risk reduction.

Figure 3. Comparison of underwriting approaches for healthcare AI.

The Standards Ecosystem

Understanding how the relevant standards relate to one another, where each contributes value, and where each leaves gaps is essential to understanding why a multi-standard approach is the right foundation for healthcare AI underwriting. The table below summarizes the portfolio; the subsections that follow describe each standard's role and limitations.

General AI governance frameworks, including NIST AI RMF and ISO 42001, describe how organizations should manage AI risk internally but do not constitute underwriting inputs. A health system can maintain a fully documented AI governance program and still deploy a clinical AI product that causes patient harm. The Standards-Proof framework therefore does not include general AI governance frameworks as underwriting criteria. Their role is narrower: they define the compliance environment that runtime enforcement is designed to satisfy, requiring organizations to demonstrate adherence to AI governance policies at the transaction level, not merely to adopt them.

Standard	Scope	Gap for Healthcare AI	Insurance Value
ISO 14971	Risk management process: hazard identification, risk control, residual risk evaluation	Does not specify AI-specific hazards, clinical validation requirements, or equity thresholds	Provides structural backbone for assessing risk management quality and control hierarchy
AAMI CR34971 and TIR 34971	AI-specific hazard categories within the ISO 14971 structure	Technical information report, not certifiable; does not specify clinical validation depth	FDA-recognized; populates AI-specific risks that ISO 14971 treats generically
ISO 13485	Quality management systems; certifiable via third-party audit	Certifies the system, not the product; does not evaluate specific AI performance	Organizational maturity signal; reduces underwriting burden through ongoing surveillance
IEC 62304	Software lifecycle; safety classification (A/B/C) drives testing rigor	Predates ML systems; does not address continuous learning or model retraining	Determines verification depth; misclassification creates systematic underpricing risk
ISO/IEC 23894	AI-specific risk management: bias, drift, interpretability, data governance	Does not specify healthcare thresholds for performance disparity or hallucination tolerance	Names AI risks as structured categories requiring specific assessment, not generic hazards
IEC 62366-1	Usability engineering; use error analysis and prevention	Not designed for AI-specific use errors: automation bias, cognitive deskilling	Critical for healthcare AI, where most consequential failures are use-related
IEC 80001-1	Network integration risk; responsible organization collaboration	Does not address AI-specific data pipeline risks or model behavior under degraded inputs	Addresses gap between product testing and production reality

Table 1: Standards portfolio for healthcare AI underwriting

ISO 14971: The Structural Backbone

ISO 14971:2019 is the foundational standard for medical device risk management. It is referenced by the FDA as a consensus standard, required under the EU Medical Device Regulation, and referenced by regulatory bodies in Canada, Australia, Japan, Brazil, and other jurisdictions. It provides a structured seven-step process covering hazard identification, risk estimation, risk evaluation, risk control, residual risk evaluation, risk management review, and post-production surveillance.

The strength of ISO 14971 for underwriting lies in its technology-agnostic structure and its control hierarchy. The standard requires manufacturers to identify hazards specific to their product, estimate severity and probability, and implement controls following a defined hierarchy that prioritizes design-level solutions over training-based mitigations. This hierarchy is critical for underwriting because it creates a structured basis for differentiating between vendors who have addressed their risk profile and those who have not.

The hierarchy operates at three levels. Design controls eliminate or reduce hazards at the source through engineering choices and carry the greatest demonstrated effectiveness. Protective measures reduce the probability or severity of harm without eliminating the underlying hazard. Information controls training, warnings, and labeling provide the least protection and, when relied upon as the primary mitigation strategy, offer minimal basis for underwriting confidence.

For example, a sepsis prediction system that will not generate alerts unless its positive predictive value exceeds a clinically meaningful threshold has decreased the alert fatigue hazard at the design level. A

diagnostic AI that will not deploy until subgroup performance falls within acceptable ranges across demographic groups has minimized algorithmic bias at the design level. A system that relies primarily on clinician training to prevent automation bias has addressed the hazard only at the information level, the least effective category.

AAMI CR34971 and TIR 34971: AI-Specific Risk Identification

AAMI CR34971:2022 is the first AI-focused guidance document recognized by the FDA as appropriate for meeting requirements for medical devices. AAMI subsequently published TIR34971:2023. It follows the same seven-step structure as ISO 14971 and populates that structure with hazard categories specific to AI and machine learning: data management, feature extraction, algorithm training and evaluation, bias, health inequity, and cybersecurity as they apply to ML systems.

ISO 13485: Quality Management Systems

ISO 13485:2016 establishes the quality management system framework within which risk management is conducted. It is one of the few certifiable standards in this group. Certification provides a signal about organizational maturity documented design controls, validated processes, corrective and preventive action systems, and management review that goes beyond any individual risk management file. The limitation is that certification confirms organizational capacity, not product-specific adequacy.

IEC 62304: Software Safety Classification

IEC 62304 establishes a classification system determining verification and validation rigor. Class A covers software where no injury is possible (administrative tools). Class B covers software where non-serious injury is possible (documentation assistants, screening tools). Class C covers software where death or serious injury is possible (sepsis prediction, treatment dosing, diagnostic AI in critical care). This classification directly determines testing depth and is distinct from regulatory device classification — a distinction that produces errors in underwriting when conflated.

ISO/IEC 23894, IEC 62366-1, and IEC 80001-1

ISO/IEC 23894 addresses AI-specific risks that traditional standards treat generically: algorithmic bias, model drift, training data governance, interpretability, and autonomous decision-making boundaries. IEC 62366-1 addresses usability engineering, particularly critical for healthcare AI, where automation bias, alert fatigue, and cognitive deskilling are among the most consequential failure modes. IEC 80001-1 addresses network integration risks, relevant because healthcare AI systems that function correctly in testing can fail when they receive degraded or incomplete data from production EHR interfaces.

Each standard addresses dimensions of risk that the others do not cover, and the gaps between them are precisely the dimensions where healthcare AI risk is most consequential. The Standards-Proof framework assembles these standards into an integrated underwriting methodology.

The Standards-Proof Framework

The preceding sections have established three concepts:

1. Healthcare AI creates multi-domain liability
Risk exposure spans clinical safety, cybersecurity, privacy, and professional liability simultaneously.

2. Existing underwriting frameworks lack the required inputs
Traditional insurance models were not designed for AI systems and therefore lack the operational evidence needed to price this risk.

3. Relevant standards exist but remain incomplete
International standards cover many technical dimensions of AI governance, but gaps remain and the standards have not been integrated into an insurable assurance framework.

The question is how to assemble these standards in a format an underwriter can use.

> **The Standards-Proof framework answers that question through a three-layer structure in which each layer addresses a different category of risk and the layers are sequenced to mirror the way healthcare AI risk materializes over a product's lifecycle.**

Layer 1: Foundation Assessment Against the Standards Portfolio

> **Primary Domain: Product Liability**
>
> Risk management file quality, design control implementation, and software lifecycle evidence. Establishes the baseline risk profile.

The first layer establishes the baseline risk profile through a structured evaluation of the quality and depth of the company's risk management implementation across the full standards portfolio. The purpose is not to determine whether the company has achieved certification or met minimum regulatory requirements. It is to evaluate implementation quality — the difference between a risk management file that lists fifteen generic hazards and one that identifies dozens of specific clinical failure modes with quantified severity and probability estimates grounded in clinical data.

The assessment covers four domains, each reflecting a distinct dimension of risk:

- **Risk management evidence** drawn from ISO 14971 and TIR 34971. The risk management file is the primary document through which a vendor demonstrates systematic hazard identification.
- **Clinical validation rigor:** In the absence of actuarial data, the depth of clinical validation is the single most consequential indicator of whether the product will perform safely in the populations it will serve.
- **Workflow integration evidence:** Capturing the human factors and deployment context risks that IEC 62366-1 and IEC 80001-1 address in principle but that require product-specific evaluation.

- **Quality management maturity** reflecting the signal that ISO 13485 certification and organizational risk management capability provide about institutional capacity to sustain quality over time.

The relative importance of these domains varies by clinical risk tier and deployment context. For a diagnostic AI in critical care, clinical validation rigor and adversarial testing evidence carry greater weight. For an administrative scheduling tool, quality management maturity and workflow integration evidence may be more determinative. The framework provides structured assessment across all domains while allowing underwriter judgment to calibrate emphasis to the specific product.

The control hierarchy from ISO 14971 translates directly into underwriting confidence. Design controls — engineering decisions that eliminate or reduce hazards at the source — receive the greatest weight. A diagnostic AI that will not deploy until subgroup performance falls within acceptable ranges across demographic groups has addressed algorithmic bias at the design level. Protective measures, such as mandatory physician review before high-risk actions, receive moderate weight. Information controls, including training programs and warnings, receive minimal weight when they constitute the primary risk reduction strategy, because evidence on their effectiveness in clinical environments is consistently weaker.

Products that cannot demonstrate adequate risk management depth across the standards portfolio do not qualify for coverage under the framework, because the evidence base required for responsible pricing does not exist.

Layer 2: Healthcare-Specific Validation

> **Primary Domain: Professional Liability**
>
> Validates whether the product performs as intended when used by real clinicians with real patients. Tiered by use-case: full clinical validation for direct clinical risk products; modified requirements for indirect risk; no clinical validation for administrative tools.

The second layer addresses whether the product actually performs as intended when used by real clinicians with real patients in real clinical environments. The standards require validation against intended use and usability testing, but they do not specify what constitutes adequate clinical validation for healthcare AI.

Use-Case Customization

Layer Two is customized to use case, with validation requirements modulated to match the risk profile of the specific product:

Tier	Risk Level	Examples	Validation Required	Primary Liability Domain
Tier 1	Direct Clinical Risk	Clinical decision support, diagnostic AI, sepsis prediction, medication dosing, triage support	Full pathway: clinical validation, blinded physician panel, supervision model, automation bias, explainability	Professional Liability (PRIMARY)

Tier	Risk Level	Examples	Validation Required	Primary Liability Domain
Tier 2	Adjacent Clinical Risk	Clinical documentation AI, ambient scribes, prior authorization AI	Modified: hallucination rate testing, decision accuracy vs. criteria, demographic parity in decisions, audit trail and verification	Enterprise Risk & Regulatory Liability (PRIMARY)
Tier 3	Administrative / Operational	Revenue cycle, billing AI, workforce scheduling, supply chain, facility management	No clinical validation. Quality management maturity (Layer 1) and runtime enforcement (Layer 3) apply.	General Liability

Table 2: Use-case risk tiers and validation requirements

- **Tier One: Direct clinical risk.** Products whose outputs directly inform patient care decisions, clinical decision support, diagnostic AI, sepsis prediction, medication dosing, triage tools. The full validation pathway applies: clinical validation, workflow integration, supervision model validation, automation bias quantification, and clinical explainability validation.
- **Tier Two: Indirect clinical risk.** Products whose outputs enter the clinical record or affect clinical workflows without making clinical recommendations. Clinical documentation AI, prior authorization systems, and clinical coding tools. Validation requirements are modified to match the pathway to harm which exists but is longer and more mediated than in Tier One.
- **Tier Three: Administrative and operational risk.** Products operating in healthcare infrastructure without a direct pathway to patient care decisions. Revenue cycle, scheduling, supply chain. Clinical validation does not apply, but quality management, adversarial testing, and runtime verification requirements remain.

Clinical Validation

For Tier One products, the gold standard requires prospective testing of patients drawn from the populations the system will serve, evaluated by an independent physician panel blinded to the AI system's outputs. The sample must be representative of the clinical diversity the intended population presents. Subgroup performance must be stratified across demographic groups, with performance within each subgroup falling within clinically justified ranges of overall performance. Statistical power must be documented and must exceed 80% for all primary performance claims.

Structured alternatives provide a pathway for products that have not yet reached the gold standard: predicate-supported validation for products in domains with established performance records; staged prospective studies with time-limited coverage and a commitment to complete gold-standard validation; institution-led validation where the deploying health system oversees the study; and retrospective validation with enhanced monitoring for lower-risk products in Tier One where the clinical evidence supports a retrospective approach. The framework does not set the gold standard as a coverage gate, because doing so would exclude precisely the early-stage products that most need the safety discipline insurance incentivizes.

Workflow Integration Validation

Workflow integration validation measures whether the AI system improves clinical efficiency or just redistributes cognitive work. A documentation AI claiming to reduce documentation time from ten minutes to two minutes may show actual results of two minutes for AI draft generation plus eight

minutes for physician review and hallucination detection, totaling ten minutes unchanged from baseline. The efficiency claim lacks support, and underwriting should reflect that.

Supervision Model Validation (Pre and Post Deployment)

Supervision model validation is the Layer Two requirement most directly relevant to professional liability. When AI enables expanded caseloads, radiologists reading several hundred scans daily with AI assistance compared to eighty without — evidence of quality maintenance at the expanded ratio becomes essential. The gold standard establishes through controlled testing the maximum patient caseload at which physician quality metrics remain at baseline levels and requires explicit documentation that the physician's malpractice carrier is aware of and has accepted the supervision model.

This requirement establishes a factual technical baseline for liability attribution analysis. An AI vendor that has documented the supervision model its product enables, validated the caseload at which that model remains safe, and confirmed that the malpractice carrier has accepted the model has created the evidence trail that both the product liability underwriter and the professional liability underwriter need to price their respective exposures consistently.

Automation Bias Quantification

Automation bias quantification tests whether physicians actually detect the cases where the AI system is wrong. A system that is 95% accurate but produces errors difficult for a physician to recognize presents a fundamentally different risk profile than one whose errors are obvious. The gold standard requires testing confirming physicians detect more than 90% of incorrect AI outputs under conditions approximating actual clinical workflow.

Clinical Explainability Validation

Clinical explainability validation confirms that the AI's reasoning is sufficient for a physician to justify the clinical decision to patients, peer reviewers, and, when necessary, in litigation. Inadequate explanations include model confidence percentages, feature importance lists, and attention weight visualizations — these provide algorithmic transparency without clinical utility.

Adequate explanations offer clinical rationale: "High sepsis risk because elevated lactate at 4.2 mmol/L, hypotension with blood pressure 85/50, fever at 101.2 degrees F, and increased heart rate at 115 bpm" provides defensible reasoning. Not all AI systems can provide this level of explanation, and the framework accounts for that limitation but the distinction between clinically useful and clinically useless explainability is a material underwriting consideration.

Layer 3: Continuous Operational Assurance (Pre- and Post-Deployment)

> **Supporting Evidentiary Layer for all Layers**
>
> Continuous control loop that keeps underwriting assumptions aligned with real-world performance and behavior in real time through adversarial testing and runtime enforcement and verification.

Pre-deployment assurance: stress testing before go-live and before any material model, workflow, or environment change.

Pre-deployment stress testing addresses risk categories that the standards portfolio covers only in principle. The standards identify adversarial manipulation and post-market surveillance requirements, but none specify testing methodology, frequency, or adequate scope for healthcare AI. This part of Layer Three fills that gap with a structured protocol designed to identify the specific vulnerabilities that healthcare AI products present in production environments.

Each evaluation runs more than one thousand test cases organized across six domains adapted for healthcare contexts:

Domain	What Is Tested
Security	Adversarial image perturbation for diagnostic AI; prompt injection for prior authorization systems; jailbreak testing of clinical scope controls
Safety	Harmful clinical outputs under edge-case conditions; scope creep beyond validated domains; high-risk recommendations without safeguards
Reliability	Clinical hallucination detection (fabricated lab values, medication allergies, clinical findings); performance consistency under load
Data & Privacy	PHI exposure in clinical explanations; training data leakage; re-identification of de-identified data through inference attacks
Accountability	Audit trail completeness; clinical governance explainability; AI failure incident response protocols; physician notification procedures
Equity	Subgroup performance across race, ethnicity, age, sex — an initial internal underwriting screen flags disparities exceeding 5% for further clinical review, not as a universal safety threshold; prior auth parity testing

Table 3: Pre-deployment adversarial testing domains

Results feed directly into coverage determination. A product that identifies vulnerabilities above defined thresholds triggers a remediation requirement before broader deployment or major release expansion. If remediation is not completed within a defined period, coverage remains conditional or constrained.

The Dynamic Test Suite

Sufficient testing at deployment does not guarantee sufficiency six months later. Patient populations shift, workflows change, and new failure modes emerge. Pre-deployment and post-deployment assurance connect as a single system: when runtime monitoring detects a new failure pattern — a drift in subgroup accuracy, a hallucination pattern not covered by existing tests — that finding is fed back into the adversarial test suite for the next model version. This transforms the question from "did you test enough before deployment" to "does your testing infrastructure systematically discover and incorporate new failure modes as they emerge, and can you demonstrate that it has done so." The second question is answerable with operational evidence.

Post-Deployment Assurance: Runtime Verification and Enforcement

Post-deployment assurance serves a fundamentally different purpose. It is the infrastructure that makes the other two layers verifiable in operation. Layers One and Two tell the underwriter what should happen. Post-deployment assurance tells the underwriter what did happen, for each consequential inference.

Deployed clinical AI systems are dynamic and require continuous prospective monitoring tied to the specific subgroups and clinical contexts for which performance claims were made at validation, not

aggregate accuracy tracking against a static set. For underwriting purposes, this distinction is critical: it defines the difference between a monitoring program that can actually detect when an insured system has drifted outside its validated operating thresholds and one that provides the appearance of oversight without any verifiable proof.

Underwriters evaluating healthcare AI risk need a concrete mechanism for assessing whether a deployed system is still behaving as it was when the policy was written. The most defensible approach is to require a documented baseline intent — a verifiable statement of what the AI system is authorized to do, and critically, what it is not — established at the point of deployment. Model drift is then measured as deviation from this baseline over time, not as a vague concept but as a quantifiable signal: the rate at which the system's outputs diverge from its stated operational boundaries. This reframes the insurability question from "is this AI safe?" (which is static) to "is this AI still operating within the behavioral envelope it was underwritten against?" (which is continuous and auditable). Without a baseline, there is no drift measurement. Without drift measurement, there is no distinction between a system that has been performing safely for twelve months and one that has been silently degrading since week five or since an LLM is updated. For underwriters, the presence or absence of continuous drift monitoring against a cryptographically anchored baseline is the single clearest indicator of whether a healthcare organization's AI risk posture is improving, stable, or deteriorating — and it should be priced accordingly.

Every major AI governance framework now in force or imminent — NIST AI RMF, ISO 42001, the EU AI Act, HIPAA as applied to AI-assisted clinical workflows, and the Colorado AI Act — shares a common structural requirement: operational evidence of governance in practice, not documentation of policy adoption alone. Organizations will need to produce that evidence, whether through audit logs, attestation architectures, third-party verification, or other means. Documentation-only programs cannot close this gap.

Without runtime verification capability, insurers face a fundamental challenge. They are pricing risk based on design documents and test results, but when a claim arrives they have no objective record of what actually happened during the specific inference that caused harm. This creates several critical gaps:

Runtime Verification and Enforcement	Why It Matters
Tamper-evident audit records within declared monitoring scope	Without tamper-evident inference records, cryptographic attestation makes post-incident log modification independently detectable; guardrails can be bypassed in practice while existing on static checklists or PDFs.
Defensible Chain of Custody	"Our model should not have done that" has no litigation value. Layer 3 preserves contemporaneous, tamper-evident records of which controls fired during the specific inference in question.
Measurable actual risk	Layers 1–2 tell the underwriter what should happen. Layer 3 tells the underwriter what did happen enabling loss ratio analysis based on operational reality.
Materially strengthen incident and audit response	FDA, FTC, and state regulators require proof of compliance at the time of the incident. Layer 3 provides contemporaneous, tamper-evident records that materially strengthen incident response, audit response, and regulatory fact development.

Table 4: Runtime verification requirements and underwriting rationale

The framework requires runtime verification infrastructure[1] that provides tamper-evident audit records within the declared monitoring scope, real-time enforcement evidence for safety policies and guardrails, and third-party attestation capability that does not depend solely on vendor self-reporting.

Runtime verification infrastructure can produce independently verifiable risk signals: governance coverage ratios, exposure windows (unattested durations), statistical violation rate bounds with confidence intervals, override frequency, and consent coverage metrics. Most of these signals are derivable directly from attestation logs without access to the operator's protected content. Two — retention integrity and consent coverage or review completion rate — additionally require operator-published policy artifacts against which log entries are evaluated. Together they address the privacy-versus-accountability dilemma that has stalled insurer engagement.

Toward Parametric Coverage for Observable AI Risk

Conventional insurance responds to events after they happen. A claim is filed, an investigation follows, and the parties reconstruct what went wrong from whatever evidence remains. For healthcare AI, this model creates a specific problem: the most consequential failures are rarely discrete events. A model drifts out of its validated performance range. A safety control silently stops executing on a subset of cases. A detection rate declines for a specific patient population over weeks before anyone notices. By the time a traditional claim surfaces, the exposure has already compounded across thousands of patient encounters and the investigation requires expensive technical discovery that is often inconclusive when the only available evidence is the vendor's own logs.

The Standards-Proof framework generates continuous operational evidence about how healthcare AI systems behave in deployment. That evidence infrastructure opens the door to a coverage mechanism that traditional healthcare liability insurance cannot offer: parametric triggers that respond to verified performance failures in near-real time, before those failures compound into patient harm and traditional claims.

How Parametric Coverage Works in This Context

Traditional indemnity insurance responds after harm has occurred. A patient is injured, a claim is filed, liability is adjudicated, and a payout is made. The process takes months or years. During that interval, the conditions that caused the harm may persist, additional patients may be affected, and the financial exposure compounds.

Parametric insurance operates on a fundamentally different mechanism. Instead of indemnifying proven losses after the fact, it pays a predetermined amount when a defined, objectively measurable event occurs. The payout is triggered by the event itself, not by a claims investigation that establishes the extent of resulting harm. Parametric structures are already established in insurance markets for natural catastrophe risk (earthquake intensity, hurricane wind speed, rainfall volume), where the triggering event is measurable and the delay inherent in traditional claims adjustment is itself costly.

[1] Open standards specifying runtime attestation infrastructure for AI governance are emerging. See, e.g., OVERT (Observable Verification Evidence for Runtime Trust), an open standard defining requirements for generating, storing, and verifying cryptographic proof of AI control execution (overt.is). OVERT specifies a risk signal architecture (Section 4.6, Annex D) that produces independently verifiable governance, operational, and agentic risk signals — including coverage ratios, violation rate bounds, override frequency, and behavioral drift rates — directly applicable to the parametric triggers and continuous monitoring this framework requires. Section 21 (Legal Preservation and Production) addresses retention, legal hold, and chain-of-custody requirements for attestation artifacts. OVERT supports evidence generation and does not determine compliance, admissibility, coverage, or safe harbor.

Healthcare AI runtime assurance creates the conditions under which parametric structures become feasible for technology risk. When an infrastructure layer generates tamper-evident, time-stamped evidence of system behavior for each inference, specific measurable thresholds can serve as trigger events:

- **Performance threshold breach.** A clinical documentation AI's hallucination rate for clinically consequential content exceeds 0.5% over a 30-day rolling window, verified against the baseline established during Layer 2 validation.
- **Equity threshold breach.** A prior authorization system's approval rates diverge by more than 3% between demographic subgroups, verified by inference-level logging of each authorization decision and its associated demographic parity check.
- **Model drift beyond validated bounds.** A diagnostic imaging AI's sensitivity drops below the threshold established in prospective clinical validation, detected through continuous monitoring against a reference standard.
- **Safety control execution failure.** The audit trail reveals that a required safety control — hallucination detection, scope boundary enforcement, human-in-the-loop verification — failed to execute for a defined number or percentage of inferences within a monitoring period.

When a trigger fires, a predetermined payout is released. The payout amount is calibrated to the estimated cost of the response the trigger demands, not to the eventual magnitude of patient harm, which may not be known for months or years.

Remediation and Financial Response in Parallel

A natural question arises: if the monitoring infrastructure detects a failure, why not simply correct it rather than pay a claim? The answer is that parametric coverage funds the correction but does not replace it. The trigger initiates two parallel tracks — one operational and one financial — that run simultaneously.

The operational track is the immediate response to the detected failure. Depending on the severity and nature of the trigger, this may include reverting to a prior model version with known performance characteristics, activating fallback protocols that route affected decisions to human review, notifying the deploying institution and affected clinicians, initiating root cause analysis, and suspending the system in the affected clinical context until the issue is resolved. The runtime verification infrastructure that detected the failure also verifies the remediation: the same audit trail that showed the threshold breach documents the corrective action and its effect on system behavior.

The financial track is the parametric payout, which provides the resources to execute the operational response without the delays inherent in traditional claims processing. A performance threshold breach in a clinical documentation AI, for example, triggers immediate costs: the deploying institution must conduct a retrospective chart review to identify patients whose records may contain hallucinated content, clinicians must be notified and allocated time for record correction, patients may require notification depending on the nature and severity of the hallucinated content, and regulatory reporting obligations may be triggered. These costs are real, quantifiable, and time-sensitive. Waiting for a traditional claims process to determine liability and authorize payment delays the response at exactly the moment speed matters most.

A Case Example

Consider the clinical documentation AI described in Case Study 4. The system processes 40,000 patient encounters per month and has a validated baseline hallucination rate of 0.3% for clinically

consequential content — approximately 120 notes per month containing fabricated clinical information with potential to affect care decisions.

Under a parametric structure, the policy defines the following trigger: if the clinically consequential hallucination rate exceeds 0.5% over a 30-day rolling window, verified by inference-level audit records showing each note's hallucination detection results, the trigger fires.

In Month 8 of the policy, a model update introduces a regression that increases the hallucination rate to 0.7% for encounters exceeding thirty minutes with patients who have complex medication lists. The runtime monitoring infrastructure detects the threshold breach within 72 hours, verified against the tamper-evident audit trail.

Two responses activate. **Operational response:** The system reverts to the prior model version for encounters matching the affected profile. The deploying institution is notified with specific information about which encounter types are affected. A retrospective review of notes generated during the breach window identifies approximately 90 notes requiring physician review and potential correction. The audit trail documents each step. **Financial response:** The parametric payout releases a predetermined amount calibrated to cover the estimated costs of retrospective chart review, physician time for note correction, patient notification where clinically indicated, regulatory reporting, and operational disruption during the remediation period. The payout occurs within days of trigger verification, not months or years after a claims investigation.

The model update is corrected and retested against the adversarial test suite, including new test cases derived from the specific failure pattern discovered in production. The system is redeployed once performance returns to validated thresholds, with the entire sequence documented in the audit trail.

What Parametric Coverage Does Not Replace

Parametric coverage addresses the observable, measurable dimension of healthcare AI risk. It presumes a functioning monitoring infrastructure that can detect threshold breaches and verify them against tamper-evident records. It is not a substitute for traditional indemnity coverage, and it is not designed to be.

Three categories of risk remain outside the parametric structure and require traditional coverage:

Unobservable failures. Monitoring infrastructure has a defined scope. A hallucination detection system may catch fabricated allergies and invented symptoms but miss subtler forms of clinical inaccuracy — a note that correctly identifies the patient's symptoms but omits a critical finding the physician mentioned during the encounter. Failures that fall outside the monitoring system's detection capability will not trigger a parametric threshold and will surface only through traditional channels: patient complaints, malpractice claims, retrospective quality reviews, or regulatory investigations. Traditional indemnity coverage remains the backstop for these scenarios.

Monitoring infrastructure failure. The monitoring system itself can fail. An audit logging service that experiences downtime, a drift detection algorithm that misses a gradual performance degradation, or a demographic parity check that malfunctions silently — all represent scenarios in which the infrastructure that should trigger parametric coverage does not operate as designed.

Long-tail liability. Some harms from healthcare AI failures do not manifest for months or years. A missed lung nodule on a screening CT, a hallucinated allergy that persists in the medical record and affects prescribing decisions across multiple future encounters, or a systematic bias in prior authorization that delays treatment for a population of patients over an extended period — all generate liability that emerges long after the triggering inference. Parametric coverage addresses the immediate

response costs when a threshold breach is detected. Traditional coverage addresses the downstream claims that emerge later.

Layer 3 of the Standards-Proof framework becomes the enabling infrastructure. It operates at the runtime and inference level, generating a continuous, tamper-evident record of what the system actually did: which inferences executed, which safety controls fired, what the model output was, and whether the deployment context matched the validated envelope. A parametric policy built on Layer 3 attestation can specify measurable thresholds — withdrawal time compliance rate, confirmed detection yield, alert volume per procedure — and verify them against an audit trail the vendor cannot retroactively alter. The trigger becomes a mathematical fact derived from the attestation record, not a judgment requiring months of adversarial discovery.

How the Three Layers Work Together

The three layers are not independent assessments conducted in sequence. They interact continuously. Layer One establishes the baseline risk profile. Layer Two validates the assumptions embedded in that profile by testing them in clinical environments with real clinicians, real patients, and real workflows. Layer Three provides continuous operational assurance: runtime enforcement, tamper-evident audit infrastructure, and the evidentiary bridge that enables liability attribution when incidents occur.

The layers also map to liability domains. Layer One, with its emphasis on risk management file quality and design controls, primarily addresses product liability. Layer Two, with its emphasis on clinical validation, workflow integration, and supervision models, primarily addresses professional liability. Layer Three provides the evidentiary bridge pre and post deployment across all three domains.

How Assessments Are Conducted

A standards-based assessment is not a checklist exercise. It is a structured evaluation conducted by reviewers with clinical, technical, and regulatory expertise who examine the full evidence portfolio a healthcare AI company has produced and evaluate its adequacy against the requirements of each layer.

Evidence Collection

The process begins with evidence collection. The company provides its risk management file (the primary deliverable of ISO 14971 compliance), software development documentation (IEC 62304), clinical validation studies, usability engineering files (IEC 62366-1), quality management system documentation (ISO 13485), and any adversarial testing results already available. For companies with mature regulatory programs, much of this documentation already exists because international standards require it.

Clinical Risk Evaluation

The assessment team evaluates the clinical risk profile of the specific product, not healthcare AI in general, but this product, in this clinical domain, serving this patient population, deployed in this clinical context. A sepsis prediction algorithm deployed in a community hospital ICU presents a different risk profile than the same algorithm deployed in an academic medical center, because the patient population, staffing ratios, specialist availability, and data quality differ in ways that affect both the probability and severity of potential failures.

The evaluation traces the clinical pathway through which harm could occur. For a diagnostic imaging AI, the assessors ask: What happens when the system produces a false negative in a patient with an aggressive malignancy? How quickly would the miss be detected through standard clinical follow-up? What is the incremental harm from the delay? What patient populations are most vulnerable to this failure mode? This clinical pathway analysis requires physician expertise that technical assessment alone cannot provide.

Evidence Quality Evaluation

The core of the assessment is evaluating evidence quality, not just evidence existence. Two companies may both have ISO 14971-compliant risk management files. One identifies forty-two specific clinical failure modes with quantified severity and probability estimates grounded in published clinical data. The other lists fifteen generic hazards "software error," "user error," "network failure" with qualitative severity ratings. Both are compliant. They represent fundamentally different levels of risk understanding. The assessment distinguishes between them.

Similarly, clinical validation studies vary enormously in rigor. A prospective study with adequate sample size, independent physician panel, subgroup stratification, and documented statistical power provides a qualitatively different signal than a retrospective analysis of convenience-sampled data with no subgroup reporting. Both may be described as "clinical validation" by the company. The assessment evaluates what was actually done.

Assessment Output

The assessment produces a structured report that maps evidence quality to each domain, identifies specific gaps, and provides the underwriter with the information needed to make a coverage and pricing decision. The report does not produce a single numerical score because the relative importance of different evidence domains varies by product, clinical context, and deployment setting, but it does provide structured qualitative assessment that the underwriter can translate into pricing through the same judgment-based process used in other specialty insurance lines.

For products with identified gaps, the report includes a remediation pathway: specific actions the company could take to strengthen its evidence portfolio, with estimated timelines and the coverage implications of completing each action. This creates economic incentives aligned with safety. A company that wants more favorable terms knows exactly what evidence to produce.

Case Studies

The following case studies are illustrative composites constructed from the authors' professional experience with healthcare AI risk assessment. They are not descriptions of specific companies, products, or engagements. Details have been generalized to demonstrate the framework's application across representative product categories and risk profiles.

Case Study 1: Patient-Facing Symptom Triage Chatbot

A consumer-facing AI chatbot had been deployed by a health system to handle patient symptom inquiries through the patient portal. The vendor characterized the product as an "informational triage assistant" that directed patients toward appropriate care settings — urgent care, emergency department, or primary care scheduling — without any clinical decision-making.

The potential for harm is two-fold. Undertriage — directing a patient with a serious condition to a lower-acuity setting or advising them to monitor symptoms at home — can delay time-sensitive intervention. Chest pain attributed to musculoskeletal strain has a narrow treatment window measured in minutes. Stroke symptoms reassured as "likely migraine" carry the same time-dependency. Overtriage creates a different harm: a chatbot that reflexively routes every ambiguous symptom to the emergency department degrades the care signal for genuinely emergent cases and imposes costs, delays, and patient anxiety that are themselves clinical harms.

Five gaps required remediation before coverage could be approved:

- No adversarial conversation-depth testing. The vendor's evaluation used single-turn exchanges. Multi-turn conversations of four or more exchanges were not tested, and scope boundary controls were not evaluated against the extended context windows that generate emergent diagnostic behavior.
- Undertriage rate not stratified by condition severity. The vendor reported an overall "appropriate routing" rate of 91%, but this figure pooled all symptom categories. When the assessment team stratified by time-sensitive conditions — chest pain, stroke symptoms, severe allergic reaction — the undertriage rate for those categories was 14%, a figure that would not have been visible without subgroup analysis.
- No usability analysis for health literacy variation. The system had been tested with users who self-reported as comfortable with medical terminology. No testing had been conducted with users with limited health literacy, limited English proficiency, or cognitive impairment — populations disproportionately represented in the patient portal's actual user base.
- Scope boundary control limitations. The system's instruction to "not provide diagnoses" was embedded in the system prompt but was not enforced through output filtering or hard scope boundaries. Prompt injection testing demonstrated that adversarial inputs could cause the system to override this instruction in approximately 7% of attempts.
- No runtime monitoring for scope boundary violations. The vendor had no mechanism to detect when the deployed system crossed from triage guidance into diagnostic statements in production. The inference-level behavior that triggered claims liability was invisible to both the vendor and the health system.

The vendor implemented output classifiers that evaluated each response against a defined scope boundary specification before delivery to the patient, with automatic response suppression and escalation to a human triage nurse when the classifier flagged a boundary violation. The undertriage rate for time-sensitive conditions was reduced to 3.2% through retraining on a stratified dataset that overrepresented the specific symptom patterns where errors had clustered. Usability testing with a representative sample that included participants with limited health literacy identified fourteen additional response patterns that were factually accurate but clinically misleading at lower reading levels. Inference-level logging was implemented to capture scope boundary classifier decisions, escalation events, and routing recommendations for every patient interaction, providing the audit trail that both the health system and the product liability carrier required for claims defense.

Post-Deployment Accuracy Evaluation

Pre-deployment testing establishes what a system can do under controlled conditions. Post-deployment accuracy evaluation answers the question that actually governs ongoing liability: what is the system doing, to which patients, in production. For a patient-facing triage chatbot, that question requires linking routing decisions to clinical outcomes in a way that the vendor's operational metrics alone cannot answer.

Case Study 2: Prior Authorization AI — Systematic Bias Remediation

A prior authorization AI trained on historical payer data had been deployed without fairness assessment. The company had characterized the product as administrative and had not conducted clinical validation. That characterization was the first problem the assessment had to address.

Prior authorization decisions directly affect patient access to care. When the system denies authorization for a cardiac catheterization in a patient with unstable angina, the consequence is not administrative inconvenience — it is delayed diagnosis of potentially life-threatening coronary disease. The appeal process typically takes five to fifteen business days. For time-sensitive conditions, that delay can result in myocardial infarction, stroke, or death. A system that makes those decisions is not administrative in any meaningful sense. It is clinical, and its risk profile is clinical.

The assessment identified a 6% disparity in authorization rates across racial groups. The source was the training data: decades of documented structural barriers in which certain populations received fewer authorizations not because they needed less care but because the healthcare system had failed to provide it. The AI had learned to replicate those barriers at scale and at speed.

Four gaps required remediation before coverage could proceed:

- No validation against evidence-based clinical criteria, and no systematic tracking of appeal reversal rates
- No documentation of what historical data the system had been trained on or who had audited it for embedded structural disparities
- No analysis of the clinical consequences of authorization delays by condition type, and no evidence that human reviewers were genuinely engaging with AI recommendations rather than rubber-stamping them

- No runtime audit infrastructure to verify that fairness controls were executing on live authorization decisions, leaving the insurer unable to confirm that a remediated algorithm was actually running in production

An independent physician panel validated the system's decisions against evidence-based clinical criteria. The algorithm was rebalanced to achieve demographic parity. A usability study and workflow redesign established that human reviewers were engaging substantively with recommendations. The company implemented inference-level logging that captured the algorithm version, input data, decision output, and demographic parity check result for every authorization processed — providing verifiable proof that fairness controls executed on each claim rather than periodic self-reporting from the vendor.

The product moved from uninsurable to insurable. The insurer recognized that a prior authorization system with validated clinical appropriateness and documented demographic parity reduces enterprise risk exposure from the class action liability pattern emerging from cases like Lokken and Cigna — making the remediated system not merely an acceptable risk but a better risk than the unvalidated version it replaced.

Case Study 3: AI-Assisted Colonoscopy — Model Drift Post Deployment

A software-as-a-medical-device (SaMD) system using computer-aided detection to identify polyps during colonoscopy had received FDA 510(k) clearance based on a pivotal study conducted at three academic gastroenterology programs. The system displayed a real-time bounding box around suspected polyps on the endoscopist's monitor, functioning as a second observer throughout the procedure. The company sought coverage for deployment across a network of ambulatory surgery centers and community gastroenterology practices.

Colorectal cancer is the second leading cause of cancer death in the United States. The primary mechanism of colonoscopy-based mortality reduction is adenoma detection: finding and removing precancerous polyps before they progress. The adenoma miss rate in standard colonoscopy is approximately 26%; missed adenomas are a primary contributor to interval colorectal cancer — cancers diagnosed within three to five years of a negative colonoscopy. An AI system that increases adenoma detection reduces that risk; one that performs unreliably in production or that induces endoscopists to reduce their own inspection effort can increase it.

Gaps required remediation before coverage could be approved:

- The validation population did not match the deployment population. The pivotal study was conducted at academic centers with high-performing endoscopists whose baseline adenoma detection rate was already in the top quartile nationally. The ambulatory surgery centers seeking to deploy the system had a different physician mix. A system validated on the best-performing endoscopists provides no evidence of how it performs with others.
- Accuracy figures were not broken down by outcome. The company reported a single pooled accuracy number. What underwriting requires is a stratified figure: how often does the system flag polyps that turn out to be hyperplastic benign findings requiring no removal? A high false positive rate for benign polyps increases procedure time, drives unnecessary resections, and creates complication risk without clinical benefit.

- No post-market surveillance monitoring plan tied to clinical outcomes. The post-market surveillance plan tracked software uptime and error logs. It did not track whether adenoma detection rates in deploying practices were actually improving. There was no inference-level logging: no record of what the system flagged during each procedure, what the endoscopist did in response, and what pathology confirmed afterward. Without that audit trail, a systematic detection failure would be invisible until a cluster of interval cancer claims made it visible in retrospect.

The vendor's initial submission relied entirely on the pivotal 510(k) clearance study. But the underwriter declined to quote on the clearance data alone. The vendor was informed that coverage would require a bridge validation study in the community setting and a post-deployment observability plan that went beyond adverse event reporting.

The health system network procuring the product also requested the same information. The health system's chief medical officer conditioned the enterprise agreement on deployment-context validation. Adenoma detection is highly operator-dependent, and a pivotal study conducted with academic subspecialists does not establish performance characteristics for a general gastroenterology workforce.

Requiring a bridge study — from the vendor's perspective — was asking them to generate evidence beyond what the FDA had determined necessary, and a significant resource and time commitment. The gap here is that FDA clearance, insurability, and procurement are different determinations: the 510(k) pathway establishes substantial equivalence to a predicate device under the conditions tested, not performance in conditions that were not tested, and a carrier writing product liability coverage for missed adenoma claims in a community gastroenterology network was underwriting the deployment population, not the study population.

This also highlights a 2024 scoping review of 692 FDA-approved AI and machine learning–enabled medical devices cleared between 1995 and 2023, which found that only 9% contained a prospective study for post-market surveillance, and only 1.9% included a link to a published scientific study with safety and efficacy data.

Coverage was approved conditional on completion of the bridge study and implementation of inference-level logging that captured per-procedure detection events, endoscopist response actions, and pathology correlation — providing the continuous performance evidence that both the underwriter and the deploying institutions required.

Case Study 4: Clinical Documentation AI — The Hallucination Problem

An ambient clinical documentation AI had been deployed across a multispecialty physician group. The company reported a 95% accuracy rate based on internal testing. The assessment began with a single question: what does 95% accuracy mean in clinical context?

The system processed approximately 40,000 patient encounters per month. A 5% error rate meant roughly 2,000 notes per month contained inaccuracies. But not all inaccuracies carry equal clinical weight. A misspelled medication name is an error. A fabricated penicillin allergy that changes subsequent prescribing decisions is a patient safety event. The accuracy figure, taken at face value, obscured the distinction entirely.

Stratifying errors by clinical consequence changed the picture. Approximately 0.3% of notes — roughly 120 per month — contained hallucinated clinical content with potential to affect care decisions:

fabricated allergies, invented symptoms, omitted critical findings. These were not random errors distributed evenly across encounter types. Hallucinations were more frequent in encounters exceeding thirty minutes, in visits with multiple active medical problems, and in patients with complex medication lists — precisely the patients where accurate documentation matters most.

Five gaps required remediation:

- Internal accuracy metrics did not stratify by clinical consequence, making the 95% figure meaningless for underwriting purposes
- No testing for hallucination patterns under the specific conditions — encounter length, clinical complexity — where they were most likely to occur
- No physician review time assessment; the company claimed physicians reviewed and approved every note, but no evidence existed that review time was sufficient to detect hallucinated content
- No post-deployment monitoring for hallucination rate changes after model updates
- No audit trail of safety control execution at the inference level, meaning that if a fabricated allergy reached the medical record and caused harm, neither the vendor nor the deploying institution could demonstrate which controls had fired — or whether they had fired at all

The company implemented hallucination detection that flagged clinical content requiring explicit physician verification before it could reach the medical record — a design-level control rather than a training-based one. Inference-level audit logging captured, for each note, which safety controls executed, whether hallucination detection had flagged content for review, and whether scope boundary controls had prevented the system from generating content outside its validated clinical domain. A physician review time study established that adequate review required a minimum of ninety seconds per note for straightforward encounters and longer for complex ones, giving the deploying institution a defensible basis for workflow expectations. Ongoing monitoring tracked hallucination rates by encounter type, with automatic escalation when rates exceeded established baselines. Coverage was approved with monitoring requirements tied directly to those thresholds.

The Next Chapter

Healthcare AI deployment is accelerating. More than 1,250 AI-enabled devices have received FDA clearance. Health systems are deploying AI at scale, and insurance is increasingly a procurement requirement. The first major lawsuits are establishing legal precedent. State legislatures are passing AI-specific regulations. The EU AI Act is entering its enforcement timeline. The window in which carriers can build specialized capability and capture first-mover advantage is open now and will narrow quickly.

Recommendations for Insurers

- **Invest in clinical advisory capability.** Healthcare AI underwriting requires physician expertise. Carriers should recruit or contract with clinical advisors who understand both medical device risk management and clinical workflow — not general AI consultants without healthcare domain knowledge. This capability takes twelve to eighteen months to build, creating a head start for early movers.

- **Develop standards-based underwriting frameworks.** Use the international standards portfolio as the foundation. These standards are already required by regulators and implemented by the companies seeking coverage, which means the evidence trail already exists.
- **Pilot with healthcare AI companies.** Start with three to five companies across different risk tiers (direct clinical, indirect clinical, administrative) to calibrate evidence requirements and pricing against actual products.
- **Engage reinsurers early.** Munich Re and other reinsurers are actively developing AI risk programs. Early engagement establishes the reinsurance support needed to write meaningful limits.
- **Consider the bidirectional risk model.** Healthcare AI that reduces diagnostic errors or improves treatment timing can decrease expected loss across a carrier's portfolio.
- **Require Independent Verification of Claims.** Design documents, test results, and process documentation tell underwriters what a system was built to do. Require policyholders to maintain tamper-evident real-time audit infrastructure that provides records of safety control execution within the declared monitoring scope.

Recommendations for Healthcare AI Companies

The appropriate starting point for insurance readiness depends on where a product sits in the regulatory and clinical landscape.

Tier 1: FDA-Regulated Medical Device AI

1. Implement ISO 14971 and IEC 62304 with sufficient depth, not just checkbox compliance
2. Conduct prospective clinical validation with subgroup stratification beyond what FDA clearance requires
3. Pursue ISO 13485 certification
4. Define model drift thresholds before deployment and instrument the system to detect them post deployment
5. Build inference-level audit records capturing safety control execution and subgroup performance threshold compliance

Tier 2: General-Purpose AI in Clinical Contexts

1. Apply ISO 14971 / IEC 62304 risk management discipline even without FDA regulatory obligation
2. Stratify hallucination rates by clinical consequence — not overall accuracy
3. Validate minimum physician review time and conditions required to reliably detect hallucinated content
4. Test whether physicians catch scope boundary violations, not just factual errors
5. Monitor for model drift and define escalation protocols when performance degrades
6. Maintain continuous post-deployment verification that safety controls are executing at inference

Tier 3: Administrative and Operational AI

1. Conduct harm pathway analysis before accepting administrative classification — if a plausible sequence leads to delayed or denied care for a time-sensitive condition, the product is Tier 2
2. Audit training data for embedded structural disparities

3. Conduct demographic parity testing at the subgroup level — not just overall performance metrics
4. Document software lifecycle processes and maintain a corrective action mechanism
5. Implement inference-level logging that captures fairness control execution on each individual decision — not just aggregate periodic reporting

Applicable to all categories: Build tamper-evident inference-level audit infrastructure proactively, not reactively. Companies that treat this as a foundational product requirement will have a meaningfully different claims defense posture and will build trust with their customers by differentiating from their competitors.

Recommendations for Clinicians Utilizing AI

Understand the AI tool's intended use and its clinical validation. Before incorporating any AI tool into your workflow, ask for the clinical validation study and assess whether the population it was conducted in resembles your patients in age distribution, comorbidity burden, demographic composition, and care setting.

Invest in AI literacy and require transparency from your vendors. You do not need to understand model architecture to use clinical AI responsibly, but you do need to understand enough to ask the right questions: What data was this trained on? What are its known failure modes? How does it behave when input data quality degrades?

Document your clinical reasoning independently of the AI output. When you agree with an AI recommendation, document why based on your own clinical reasoning; when you override one, document that too. This record is your primary protection when the AI system's output later becomes the subject of a claim.

Know your automation bias risk and actively counteract it. Automation bias emerges in experienced clinicians working under time pressure with tools that are usually correct, meaning the cases where the AI is wrong are precisely the cases where you are least likely to catch the error.

Treat AI-generated clinical documentation as a draft, not a record. Review AI-generated notes with the critical attention you would apply to a medical student's note, paying particular attention to medication lists, allergy fields, and clinical findings that would affect subsequent treating physicians. Your signature is the legal attestation that the content is accurate.

Participate in your institution's AI governance processes. If your institution has an AI governance committee or vendor evaluation program, engage with it; if it does not, advocate for one. Clinicians are the stakeholders best positioned to evaluate whether a tool's clinical validation is adequate for the population it will actually serve, and are often the last to be consulted and the first to be held liable when it is not.

Recommendations for Health Systems, Life Science Companies, and Payors

- **Require proof of insurance from AI vendors.** Insurance serves as a proxy for risk management maturity. Vendors who can obtain coverage have demonstrated, to an independent underwriter, that their evidence portfolio meets a minimum standard.

- **Require AI vendors maintain tamper-evident audit logs of safety control execution within the declared monitoring scope,** and that those logs are accessible to the institution in the event of a regulatory investigation, adverse event review, or litigation discovery request.
- **Demand validation evidence.** Risk management files, clinical validation studies, and adversarial testing results should be standard procurement requirements, not optional enhancements. Do not accept "we're working on it" — systematic validation is a prerequisite for patient safety.
- **Coordinate with malpractice carriers.** When deploying AI that changes physician workflows or expands caseloads, ensure that the malpractice carrier is aware of and has accepted the supervision model. This protects the institution from coverage gaps.
- **Track performance data.** Institutional performance data including outcomes before and after AI deployment will become the actuarial foundation on which future insurance pricing depends. Institutions that collect this data systematically are better positioned to demonstrate risk reduction and negotiate favorable terms.

Regulatory Evolution

The regulatory environment will continue to develop through the 2026–2028 timeframe and beyond. FDA AI/ML guidance finalization, state law proliferation, EU AI Act implementation, and the first major claims establishing case law will all shape the environment. A standards-based approach is inherently adaptable to this evolution because the underlying standards evolve with the regulatory environment rather than becoming obsolete.

As claims data accumulates over the coming decade, the evidence quality assessment methodology can be calibrated against actual loss experience, creating a progressively more accurate pricing model with the standards-based risk assessment as guide.

Limitations

We recognize that a standards-based risk assessment framework cannot resolve a problem that is still in a transitional period awaiting data and legal precedent. What it can do is organize the evidence that underwriting decisions require while that data accumulates and that precedent develops.

This framework cannot substitute for actuarial data. The three-layer methodology identifies what evidence should exist, how well it has been executed, and whether it is sufficient to support a coverage decision, but actuarial pricing requires claims history that does not yet exist for healthcare AI at meaningful scale. What the framework provides is a principled basis for decisions that must be made before the data arrives. That is a meaningful contribution, but it is not the same thing as actuarial pricing.

Standards compliance does not guarantee safety. ISO 14971 can be implemented rigorously or superficially. IEC 62304 can produce thorough software lifecycle documentation or credentialing paperwork. ISO 13485 certifies organizational capacity, not product-specific adequacy. The framework's evidence quality assessment methodology is designed to distinguish rigorous implementation from checkbox compliance — but that assessment is only as good as the reviewer conducting it, and a determined vendor can produce evidence that satisfies formal criteria without meaningfully reducing patient risk.

The framework reduces this exposure by requiring operational verification, not only documentation, but it does not eliminate it.

Legal precedent is still forming in ways that matter for coverage design, and the gaps are specific. The first major verdicts will answer some of these questions and create new ones. Coverage terms written today will be interpreted against standards that do not yet exist, and exclusions drafted to address current theories of liability may not reach the theories plaintiffs' counsel will develop once claims data is available.

The framework generates the operational evidence that will matter in those disputes: what the system was designed to do, what it actually did at the inference level, whether safety controls executed as documented. It cannot determine how courts will allocate liability across vendors, deployers, and clinicians when they do. That uncertainty is structural, not a gap the framework can close.

Conclusion

The pace of healthcare AI deployment has outpaced the insurance frameworks designed to price its risks. The resulting gap creates risk for every stakeholder: AI vendors face liability exposure, health systems and clinicians face unquantified risk from the products they deploy, insurers face a market they cannot responsibly price, and patients face the consequences of inadequately validated systems deployed without the oversight mechanisms that insurance provides.

The Standards-Proof framework offers a systematic path forward. It is evidence-based — every requirement traces to regulatory standards or documented failures — and it is implementable because the standards it builds on are already in use by thousands of medical device manufacturers and are already required by the regulatory bodies that govern healthcare AI.

The framework is designed to be adaptable, adjusting to level of risk and use case, and evolving with the regulatory environment and recalibrated against actual loss experience as claims data accumulates. Legal precedent is still forming in ways that will shape coverage design. It identifies both the AI products that create risk and the AI products that reduce it, enabling a market mechanism that rewards safety, assurance, and accountability.

The emergence of agentic AI — systems capable of taking autonomous sequences of actions across clinical workflows with minimal human oversight at each step — will introduce a meaningfully different liability structure. When no single decision point carries a physician co-signature, the attribution problem that the framework addresses across product liability, professional liability, and enterprise risk becomes substantially harder to resolve. Standards bodies are already responding, with emerging specifications extending governance and verification requirements to autonomous action sequences. The Standards-Proof framework's layered structure — grounded in standards that evolve, adversarial testing that updates as threat environments change, and runtime enforcement that captures what systems actually do rather than what they are designed to do — is built to incorporate agentic risk categories as deployment patterns, failure modes, and litigation templates develop.

The healthcare AI insurance market will mature. First movers — both insurers and AI companies — will establish market position, influence emerging standards, and build competitive advantages that compound over time. The question is not whether healthcare AI will become insurable, but which carriers and which companies will lead the market when it does.

References

1. Estate of Gene B. Lokken v. UnitedHealth Group, Case No. 0:23-cv-03514-JRT-SGE, U.S. District Court for the District of Minnesota. Ruling on motion to dismiss, February 13, 2025.
2. Kisting-Leung v. Cigna Corp., U.S. District Court for the Eastern District of California. Ruling on motion to dismiss, March 31, 2025.
3. Morgan Lewis. "AI in Healthcare: Opportunities, Enforcement Risks and False Claims, and the Need for AI-Specific Compliance." July 2025.
4. Federation of State Medical Boards. Navigating the Responsible and Ethical Incorporation of Artificial Intelligence into Clinical Practice. May 2, 2024.
5. American Law Institute. Restatement (Third) of Torts: Medical Malpractice. Approved May 2024.
6. Tonello M, Jones A. "AI Risk Disclosures in the S&P 500: Reputation, Cybersecurity, and Regulation." The Conference Board/ESGAUGE. Harvard Law School Forum on Corporate Governance, October 15, 2025.
7. Bloomberg Law. "AI, Algorithm-Based Health Insurer Denials Pose New Legal Threat." April 2025.
8. ECRI. Top 10 Health Technology Hazards for 2025. Plymouth Meeting, PA: ECRI, December 2024.
9. ECRI. Top 10 Patient Safety Concerns 2025. Plymouth Meeting, PA: ECRI, March 2025.
10. ArentFox Schiff. "AI Service Agreements in Health Care: Indemnification Clauses, Emerging Trends, and Future Risks." July 2025.
11. Chambers and Partners. "Healthcare AI 2025 — USA." Global Practice Guides, 2025.
12. Fenwick & West. "Tracking the Evolution of AI Insurance Regulation." December 2025.
13. National Association of Insurance Commissioners. Model Bulletin on the Use of Artificial Intelligence Systems by Insurers. Approved December 2023. Available at: content.naic.org/insurance-topics/artificial-intelligence
14. Grand View Research. "Healthcare Artificial Intelligence Market Size, Share & Trends Analysis Report." 2025.
15. Munich Re. aiSure: AI Performance Guarantee Solutions. 2025.
16. ISO 14971:2019. Medical devices — Application of risk management to medical devices.
17. AAMI CR34971:2022 / TIR34971:2023. Application of ISO 14971 to machine learning in artificial intelligence.
18. ISO 13485:2016. Medical devices — Quality management systems.
19. IEC 62304:2006+AMD1:2015. Medical device software — Software life cycle processes.
20. ISO/IEC 23894:2023. Information technology — Artificial intelligence — Guidance on risk management.
21. IEC 62366-1:2015+AMD1:2020. Medical devices — Part 1: Application of usability engineering to medical devices.
22. IEC 80001-1:2021. Application of risk management for IT-networks incorporating medical devices.
23. Dowdell J, Stecklow S, Terhune C, Levy R. "As AI enters the operating room, reports arise of botched surgeries and misidentified body parts." Reuters. February 9, 2026.
24. MobiHealthNews. "Patient files lawsuit against Sharp Healthcare ambient AI use." December 2025.

25. Lee B, Dai T, Guo C, et al. "Recalls of Artificial Intelligence and Machine Learning–Enabled Medical Devices in the United States." JAMA Health Forum. 2025;6(4):e250543.

26. Robbins R, Jewett C. "FDA clears high-risk AI medical devices with little data, Reuters investigation finds." Reuters. August 2024.

27. U.S. Food and Drug Administration. "Artificial Intelligence and Machine Learning (AI/ML)–Enabled Medical Devices." Updated July 2025.

28. Rucker P, Miller M, Armstrong D. "How Cigna Saves Millions by Having Its Doctors Reject Claims Without Reading Them." ProPublica. March 25, 2023.

29. Holland & Knight. "The Implications and Scope of the NAIC Model Bulletin on the Use of AI by Insurers." May 2025.

30. Aaron DG et al. "A New Legal Standard for Medical Malpractice." JAMA. Published online February 26, 2025. doi:10.1001/jama.2025.0097

31. NAIC Statement on AI Executive Order, December 16, 2025. Available at: content.naic.org

32. Adams R, Henry KE, Sridharan A, et al. "Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning–based early warning system for sepsis." Nature Medicine. 2022;28:1455–1460. doi:10.1038/s41591-022-01894-0

33. Corley DA, Jensen CD, Marks AR, et al. "Adenoma detection rate and risk of colorectal cancer and death." N Engl J Med. 2014;370(26):1298–1306.

34. Zhao S, Wang S, Pan P, et al. "Magnitude, Risk Factors, and Factors Associated With Adenoma Miss Rate of Tandem Colonoscopy: A Systematic Review and Meta-analysis." Gastroenterology. 2019;156(6):1661–1674.e11.

35. Muralidharan V, Adewale BA, Huang CJ, et al. "A scoping review of reporting gaps in FDA-approved AI medical devices." NPJ Digital Medicine. 2024;7(1):273. doi:10.1038/s41746-024-01270-x

36. OVERT (Observable Verification Evidence for Runtime Trust), Version 1.0. GLACIS Technologies, March 2026. Available at: overt.is