SUMATOSOFT

Compliance and Audit Barriers to AI Adoption in Regulated Industries

The evidence gap: compliance and audit barriers to the adoption of artificial intelligence in healthcare and financial services

33 firms in healthcare, medical devices, fintech, and adjacent regulated sectors described how compliance rules affected a specific AI project.

SUMATOSOFT RESEARCH | KATERINA MERZLOVA | SEPTEMBER 8, 2026

Contents

01 TL;DR
02 Abstract
03 Key findings about compliance and audit barriers to AI adoption
04 Introduction
05 Sample description
06 Findings
07 Sector analysis
08 Regulatory context
09 Discussion
10 Conclusion
11 References
12 How to cite this research
13 Appendix A: Respondent roster
14 Appendix B: Framework glossary
15 About this study

Cite as: Merzlova, K. (2026). *The evidence gap: compliance and audit barriers to the adoption of artificial intelligence in healthcare and financial services.* SumatoSoft Research.

Read online:
sumatosoft.com/blog/compliance-and-audit-barriers-to-ai-adoption-in-regulated-industries

TL;DR

- 33 firms in healthcare, medical devices, fintech, and adjacent regulated sectors described how compliance rules affected a specific AI project.
- 22 redesigned the system to satisfy a rule, 5 delayed it and then delivered, 3 encountered no obstacles at all, and only 2 abandoned the work.
- 19 of the 33 blocking constraints concerned producing evidence rather than the accuracy of a model.
- 8 respondents named buyer-side security review as the binding constraint, placing customer procurement ahead of both GDPR and SOC 2.
- 30 of 33 expect a heavier compliance burden within 24 months, and none expect an overall reduction.

Abstract

- **Objective.** Published surveys on AI adoption show that regulated firms adopt it more slowly than technology firms, and they attribute this gap to compliance requirements. Few studies identify which rules cause delays or what happens to a project once a rule applies. This study addresses that gap.
- **Design.** Qualitative study using a structured open-response instrument of three questions, run through media response platforms and direct outreach between June and August 2026.
- **Participants.** 67 submissions arrived from 65 distinct respondents. After screening, 33 responses were included in the analysis. 23 respondents work in healthcare, in the Internet of Medical Things, in financial technology, or supply AI systems to buyers in those sectors. 10 work in adjacent regulated sectors and are reported as a comparison group.
- **Main measures.** Each response received one primary outcome code and one primary barrier code, together with framework codes and remediation-practice codes derived inductively from the corpus.
- **Results.** 22 of 33 firms redesigned an AI system to satisfy a compliance rule, while 2 abandoned one. 19 of 33 primary barriers concerned evidence production. 8 respondents named buyer-side security review as their binding constraint, placing customer procurement above every named framework except 4. 7 respondents reported that retrofitted records failed under review, and 5 reported the same outcome after treating a supplier's certification as internal validation. 30 of 33 expect their compliance burden to grow within 24 months, and none expect it to fall.
- **Conclusions.** In this sample, compliance rules reshaped AI projects far more often than they stopped them. The binding constraint was the company's capacity to produce reconstructable evidence. Organizations that designed evidence capture into a system before deployment reported materially shorter remediation cycles than firms that reconstructed it afterward.
- **Limitations.** The sample is self-selected and weighted toward the United States. Reported figures are self-reported.

Key findings about compliance and audit barriers to AI adoption

1. **Redesign was the default outcome, and abandonment was rare.** 22 of 33 firms rebuilt an AI system to comply with a rule. 5 delayed a launch and then delivered. 3 reported no obstacles because they had excluded AI from regulated decisions before starting. 1 responded from an advisory position. 2 abandoned the project.
2. **Evidence, not accuracy, was the blocking constraint.** 12 respondents named the audit trail and decision reconstruction as barriers. 7 named the confidential-data boundary, and 4 named data lineage and provenance. 2 each named change control, explainability of individual decisions, model validation and generalization, and professional accountability. A further 2 encountered no barrier.
3. **The EU AI Act led the framework counts, and procurement outranked most statutes.** 17 respondents named the EU AI Act, 9 named FDA guidance, 7 named HIPAA, and 7 named model risk management rules. GDPR drew 4 mentions, as did SOC 2 and professional conduct rules. 3 respondents named IRS reporting rules. 8 named buyer-side security review, which is a customer questionnaire rather than a statute.
4. **Two remediation practices carried the successful cases.** 10 respondents named recorded human sign-off, and 9 named evidence captured at the moment of decision. 5 described a deterministic core with a generative layer confined to surrounding language.
5. **Retrofitting was the dominant failure.** 7 respondents attempted to add records after building a system, and 5 accepted a supplier certification in place of their own validation. 5 supplied periodic reports where auditors wanted continuous evidence, and 4 offered accuracy metrics as compliance evidence.
6. **Remediation cost ran from six weeks to more than a year.** 6 respondents quantified it, and 1 attached a figure of roughly $140,000 in engineering spend. Every quantified case describes evidence work on a system that already functioned.
7. **The forecast was close to unanimous.** 30 of 33 expect a heavier burden within 24 months. 3 are anticipating eventual relief, locating it in standardization of the evidence format, and not in any relaxation of the underlying rule.

Introduction

Background

Organizations operating under regulatory supervision adopt AI at lower rates than firms that do not. A 2025 survey of support operations reported adoption rates of 92% among technology companies and 58% among firms in healthcare, financial services, government administration, and similar regulated sectors (Deskpro, 2025). The same study found that 64% of financial services respondents and 61% of healthcare respondents named compliance as a large driver of their security posture.

The governance capacity behind that adoption remains thin. A survey of more than 350 firms found that fewer than 20% had implemented model cards, dedicated incident reporting tools, regular red teaming, or similar controls (Pacific AI, 2025). Among senior finance leaders, 92% reported that their firm had adopted AI or intended to do so within 12 months, while only 43% reported having a formal governance framework (BDO, 2025).

Sector-specific research points to auditability as the pressure point. Grant Thornton surveyed 950 executives and found that 18% of banking respondents were confident that their firm could pass an independent audit of its AI controls (Grant Thornton, 2026). Banking respondents were more likely than respondents in any other surveyed industry to describe their controls as untested.

Problem statement

The published literature measures compliance as a sentiment variable. Respondents are asked whether compliance concerns them, and the proportion answering affirmatively is reported. That design establishes that a constraint exists without describing how it operates.

Three questions remain unanswered by the survey literature. Which rule, applied to which class of system, produces the delay? What firms change in response, and whether those changes survive external review. What the change costs in elapsed time and engineering effort.

Research questions

This study addresses three research questions, each corresponding to one item in the instrument.

RQ1. Which compliance and audit rules have delayed, altered, or stopped AI projects in regulated firms, and what outcome followed?

RQ2. Which remediation practices did firms adopt to satisfy a regulator or an auditor, and which practices failed under review?

RQ3. Where do practitioners expect the compliance burden to increase or decrease over the following twelve to twenty-four months?

Research design

This study uses a qualitative, cross-sectional design with a structured open-response instrument. The design suits the research questions because the object of interest is a causal sequence within a firm, and that sequence resists capture through closed-response items. Respondents were asked to narrate a specific episode, allowing inductive coding against categories generated by the corpus itself.

The design carries an evident cost. A qualitative instrument administered to a self-selected sample does not support any inference about population prevalence. The section on limitations sets out the resulting limits, and the comparison with published research checks every main finding against large-sample studies.

Instrument

The instrument comprised three open-response items and a structured company profile. Respondents received the following text.

1. Which specific compliance, validation, or audit requirement has slowed or blocked an AI project at your company? What happened: delayed, redesigned, or shelved?
2. What did you change in how you document, validate, or monitor AI to satisfy a regulator or auditor? Name one thing that worked and one that didn't.
3. Over the next 12 to 24 months, where does the compliance burden for AI get heavier or lighter for you (EU AI Act, FDA guidance, model risk rules, other)? How are you preparing?

The first item names three candidate outcomes to establish a common response frame and to allow outcome coding without further interpretation. Item 2 requests a matched pair, which discourages promotional responses by requiring an admission of failure alongside any claim of success. The third item lists four regulatory instruments to signal the expected level of expertise and anchor forecasts to identifiable rules.

The profile captured respondent name, job title, company, industry, company size, operating jurisdictions, and stage of deployment. The latter carries analytical weight because it distinguishes firms that have undergone external review from those that have not.

Recruitment and sampling frame

SumatoSoft published an open call through Connectively and supplemented it with direct outreach to practitioners in the target sectors. Recruitment ran from June to August 2026.

The recruitment channel introduces a documented bias. Media response platforms compensate participants in publicity, and that incentive attracts submissions optimized for quotation rather than for accuracy. The screening section describes the procedure adopted in response, and the limitations section treats the residual bias.

Screening and exclusion

The call returned sixty-seven submissions from sixty-five distinct respondents. Two respondents submitted twice, and one respondent distributed a single answer across three separate messages.

We excluded 32 submissions based on five criteria, applied in the order below.

- **Criterion 1: coordinated template submission (10 excluded).** One sending address transmitted an identical response body twice, attributed to two different individuals at two different companies. Five further submissions reproduced that template's structure, heading pattern, argumentative sequence, and closing request, and originated from sending addresses unconnected to the named individual. Several closed with an explicit request for a hyperlink.
- **Criterion 2: duplicate or internally contradictory submission (2 excluded).** One respondent supplied two accounts of a single three-month delay and attributed each account to a different regulation. A second respondent transmitted substantially identical text on two occasions.
- **Criterion 3: no substantive response (2 excluded).** Two communications agencies offered to arrange an interview without answering any item.
- **Criterion 4: unverifiable figures or attribution failure (5 excluded).** One submission asserted an 18-month delay and a $500,000 loss on a project at an unnamed former employer, providing no verifiable details. One stated the respondent's role in two incompatible ways within a single message and appended an affiliate referral offer. Three presented secondary reading or unrelated credentials as first-hand operational experience.
- **Criterion 5: no AI deployment in a regulated sector (13 excluded).** Marketing agencies, a vehicle manufacturer, a mechanical contractor, and similar respondents described governance in sectors outside the sampling frame. Several answered candidly and in good faith, and their material falls outside the population under study.

Thirty-three responses met all five criteria and were included in the analysis.

Coding scheme

Each accepted response received one primary outcome code, one primary barrier code, one or more framework codes, and one or more remediation-practice codes. The coding scheme published at the end of this report reproduces the categories with their definitions.

Outcome codes were specified before coding began. Three values come from the instrument itself, and two values are required by the corpus. The five codes are redesigned, delayed then delivered, shelved, never blocked, and advisory response. The fourth accommodates respondents who reported no barrier because they had excluded AI from regulated decisions before starting. The fifth accommodates respondents who advise regulated firms without operating a system themselves.

Barrier codes were derived inductively. Eight categories emerged from an initial pass across the corpus and were then applied uniformly. Where a respondent described several obstacles, we coded the obstacle the respondent named as blocking. Where a respondent described several

projects, we coded the project described in most detail.

Framework codes and remediation-practice codes allow multiple values per respondent, and counts in those categories therefore exceed thirty-three.

A single researcher performed all coding. The study reports no inter-coder agreement statistic, and the limitations section treats that absence.

Limitations and threats to validity

- **Sampling.** The sample is self-selected and not random. Percentages reported throughout describe thirty-three firms and estimate nothing about healthcare or financial services at large. Readers should treat every proportion as descriptive.
- **Self-selection bias.** Practitioners who have encountered a compliance obstacle have stronger reason to respond than practitioners who have not. The distribution of outcomes reported below almost certainly overstates the frequency of obstruction relative to the underlying population.
- **Recruitment incentive.** Participants were compensated in attribution. That incentive rewards vivid narration and creates pressure toward exaggeration. The fourth screening criterion removed the submissions where that pressure produced unverifiable claims, and residual overstatement within accepted responses cannot be excluded.
- **Self-report.** No figure attributed to an individual respondent was independently verified. Elapsed timelines, cost figures, accuracy rates, and headcount details reflect respondent recollection.
- **Single-coder design.** One researcher assigned all codes, and the study therefore reports no measure of coding agreement. The coding scheme published at the end of this report lists every category definition, so that a second coder may replicate it.
- **Geographic concentration.** Twenty-seven of thirty-three respondents named the United States among their operating jurisdictions. European respondents supply most of the material on the EU AI Act. Respondents in the United States supply most of the material on the FDA and the tax authority. Regional findings inherit that concentration.
- **Small cell sizes.** Several barrier categories rest on two respondents. Such a count establishes that a pattern occurs, and establishes nothing about how often it occurs.
- **Temporal instability.** The regulatory environment shifted during data collection. The European Parliament approved amendments to the EU AI Act on 16 June 2026, and the Federal Reserve replaced its model risk management guidance during the same period. The regulatory context section documents both changes, and the forecasts reported here should be read against the state of the rules at the time of response.

Conflict of interest

SumatoSoft provides custom AI and software development services to firms in the sectors under study, including compliance evidence and system validation. The company therefore holds a commercial interest in the salience of the problem this study documents. To limit the effect of that interest, the study publishes its screening criteria, coding scheme, full respondent roster, and source list. Each main finding is checked against published research from other authors.

Sample description

Sector composition

Nine respondents work in healthcare delivery or medical device making. Nine work in financial technology. Five supply AI products to buyers in regulated sectors without operating in those sectors directly. The remaining 10 work in adjacent regulated sectors and serve as the comparison group.

Figure 1. Who answered

Sector	Respondents
Healthcare and the Internet of Medical Things	9
Financial technology	9
Legal and professional services	6
AI suppliers selling into regulated buyers	5
HR technology, payroll, ERP and delivery partners	4

The comparison group covers legal services, tax practice, human resources technology falling under Annex III of the EU AI Act, payroll software, and enterprise resource planning. These firms face similar evidence rules under different instruments, and their responses help assess whether the main findings are specific to healthcare and finance.

Jurisdiction

Respondents reported their operating jurisdictions and could name several.

Figure 2. Where they operate

Jurisdiction	Respondents
United States	27
European Union	10
United Kingdom	7
Australia	2
Canada	1
India	1
Ukraine	1
Switzerland	1
South Africa	1
Gulf states	1

Deployment maturity

Twenty-one respondents operate AI in production. Seven are conducting pilots. Two describe deployment as scaled across multiple functions. Two responded from an advisory position without operating a system, and one has entirely excluded AI from regulated decisions.

The concentration in production deployment strengthens the study relative to surveys that pool aspirational and operational respondents. Twenty-three of thirty-three respondents have submitted a system for some form of external review.

Findings

AI project outcomes (RQ1)

Twenty-two of thirty-three firms redesigned an AI system to comply with the rule. Five delayed a launch, then delivered. Two stopped the project. Three reported no obstacle because they had excluded AI from regulated decisions before starting, and one responded from an advisory position.

Figure 3. Project outcomes

Outcome	Respondents	Share
Redesigned	22	67%
Delayed, then delivered	5	15%
Never blocked	3	9%
Shelved	2	6%
Advisory response	1	3%

The redesigns follow a consistent structure across sectors. System autonomy narrows, a named individual enters the approval path, and the evidence trail shifts from records produced after the fact to the system's architecture itself.

Sergiy Fitsak, Managing Director at Softjourn, described a representative episode. His team evaluated an AI tool against a financial technology onboarding flow that distinguished prepaid card users from credit card users. The output appeared correct on inspection. The tool had reclassified a step carrying mandatory compliance weight as optional, and the reclassification went undetected until manual review. The team rebuilt the affected section by hand and revised its validation procedure. Fitsak drew a general lesson from the episode:

> *Sequencing and classification decisions with compliance consequences cannot be delegated to a model working from a written description alone. The context that prevents that kind of error lives in the team, not in the prompt.*
>
> SERGIY FITSAK, SOFTJOURN

Egiziago Cioffi, Chief Executive Officer of the Microsoft partner SynSphere, encountered Article 9 of the General Data Protection Regulation while constructing a healthcare assistant for an Italian client. Special-category health data cannot enter a model whose scope never contemplated holding it. His team narrowed the functional scope, classified the data before any model received it, and restricted processing to a European region. The project proceeded in modified form.

The three respondents reporting no obstacles deserve a separate note, because their responses reflect a strategy rather than an absence. Dr. Maria Knöbel, Medical Director at MedicalCert UK, described a rule that predates any regulatory inquiry. Every certificate issued through her platform is approved by a physician on the General Medical Council register.

> *We haven't had an AI project blocked because we set clear limits from the beginning. We don't use AI to approve certificates or make clinical decisions. That wasn't a rule imposed on us later. It was a decision we made from day one.*
>
> DR. MARIA KNÖBEL, MEDICALCERT UK

Heath Squier, Chief AI Officer at Equity Edge Lending, made a parallel observation from consumer lending. No regulator has formally blocked a project at his firm, and the friction appears earlier, when a proposed workflow proves difficult to secure internal approval.

> *A workflow becomes difficult to approve when the team cannot show what data entered the model, what the model changed, and who reviewed the result. That pushes higher-risk uses away from autonomous decision-making and toward assisted workflows with a named human owner.*
>
> HEATH SQUIER, EQUITY EDGE LENDING AND EVKII

Barrier taxonomy (RQ1)

Eight barrier categories emerged from the corpus. The distribution identifies where the company's effort is concentrated.

Figure 4. What blocked it

Barrier	Respondents
Audit trail and decision reconstruction	12
Confidential-data boundary (health records, privilege)	7
Data lineage, consent and provenance	4
Change control and revalidation	2
Explainability of individual decisions	2
Model validation and generalization	2
Professional accountability and licensure	2
No barrier encountered	2

Twelve respondents named audit trail and decision rebuilding as the blocking constraint. Seven named the confidential-data boundary governing protected health information and legal privilege. Four named the lineage and provenance of training data. Those three categories

account for twenty-three of thirty-three responses, and nineteen of the thirty-three primary barriers concern evidence production rather than model behavior.

The distinction carries operational weight. Reviewers in this sample did not ask whether a model performed accurately. They asked whether the firm could show what had occurred.

Ken Herron, co-founder of VCONify, described the deficiency precisely. A transcript and a final output establish neither which source data the system consulted nor which configuration was active. They also omit whether a person reviewed the recommendation, and what business action followed. His team concluded that a conventional logging architecture would not yield a defensible audit record, and rebuilt around durable records of the complete interaction. Structured delivery frameworks address the same gap, and SumatoSoft has published its own Agentic Development Lifecycle, which covers human-in-the-loop controls and private deployment boundaries. His account of the failed approach was direct:

> *What did not work was trying to reconstruct that sequence afterward from transcripts, screenshots, application logs, ticketing systems, and separate approval records. Each artifact could be accurate, but the collection still failed to establish a clear chain of custody. A folder full of artifacts is not necessarily a defensible record.*
>
> KEN HERRON, VCONIFY

Viktor Bulanek, Founder and Chief Technology Officer of Penetrify, encountered the same constraint from an unusual position. His product operates autonomous agents that conduct penetration tests. Security reviewers at regulated buyers did not evaluate the quality of the AI. They requested the artifact and then asked the firm to produce it again. An agent is non-deterministic by design, and two executions against a single target therefore diverge. That property conflicts directly with the evidentiary standard applied by auditors. The company now checkpoints scan executions so that a specified run can be replayed, and stores reports as immutable, versioned artifacts.

> *Security reviewers and auditors at regulated buyers do not ask "is your AI good"; they ask "show me the artifact and prove you can produce it again." An AI agent is non-deterministic by nature, so two runs against the same target do not look identical. That collides directly with how audit evidence works. It did not shelve the product; it redesigned it. That work was not on the roadmap. Compliance put it there.*
>
> VIKTOR BULANEK, PENETRIFY

Framework salience and the procurement channel (RQ1)

Respondents named the regulatory instruments governing their work. The EU AI Act leads by a large margin, including among respondents without European operations.

Figure 5. Which rules they named

Framework	Respondents
EU AI Act	17
FDA guidance (SaMD, change control, device lifecycle)	9
HIPAA	7
Model risk management	7
GDPR	4
SOC 2	4
Bar and professional conduct rules	4
IRS reporting rules	3
DORA, NIS2, PCI-DSS, 42 CFR Part 2 and MHRA guidance	5 (one each)
Buyer-side security review	8

Seventeen respondents named the EU AI Act and nine named FDA guidance. Seven named HIPAA, and a further seven named model risk management rules.

One constraint falls outside the framework taxonomy. Eight respondents named buyer-side security review as the mechanism that delayed their work. That count positions a customer procurement questionnaire above both the General Data Protection Regulation and SOC 2 in observed salience, behind only the four leading instruments.

Chase W. Hughes, who founded and later sold the AI product ProAI, articulated the mechanism directly. The binding constraint in financial services was buyer diligence, which arrived well in advance of any statutory rule. His team presented AI-generated financial projections to certified accountants and institutional buyers, and the blocking question never varied. Reviewers required proof of how a figure was derived, followed by proof that the derivation could be reproduced.

Richard Schreiber, Chief Executive Officer of RAS Consulting Group, named the associated failure. Organizations that accepted a supplier's SOC 2 report as their own diligence were exposed when the supplier revised its terms. A certification attests that the supplier maintains controls. It establishes little about how a model handles a given data class, whether prompts are retained, what a subprocessor change produces, or how long any of it persists.

For suppliers selling AI to healthcare and financial services, this finding has commercial implications. Auditability operates as a product attribute. Bulanek reached that conclusion and constructed an integration that transmits findings into the customer's own compliance platform, so that evidence arrives in the system the auditor already consults. That change shifted the conversation away from explaining the AI and toward a control with attached evidence.

Remediation practices (RQ2)

Item 2 requested one practice that succeeded and one that failed. The distribution across both columns is consistent.

Figure 6. What worked, and what did not

Practice that worked	Respondents
Human sign-off recorded as a step, not asserted in policy	10
Evidence captured at the moment of the decision	9
Deterministic core, generative layer only at the edge	5
Controls placed at the data and identity layer	3
One evidence package per use case, with a named owner	3
Change control written before the model goes live	2
Evidence pushed into the auditor's own system	1

Practice that failed	Respondents
Records retrofitted after the build	7
Periodic reports where auditors wanted live evidence	5
Supplier certification treated as internal validation	5
Accuracy metrics offered as compliance evidence	4
A prompt or a written policy used as a control	3
One generic checklist applied to every use case	1
Synthetic test suites standing in for production data	1

Ten respondents named recorded human sign-off as the successful practice. Nine named evidence captured at the moment of decision. Five described an architecture in which a deterministic component computes any value subject to audit while a generative component is confined to surrounding language.

David Kemmerer, co-founder and Chief Executive Officer of CoinLedger, supplied the most detailed account of the third pattern. His firm constructed a language model to parse decentralized finance transactions and match them to taxable events. Auditors observed that the generative output lacked supporting documentation, and filings with the Internal Revenue Service require that every figure be manually verified. A model with 99.2% accuracy produced about 8 erroneous transactions per 1,000, each error generating a tax liability for a customer.

The rebuild introduced a procedure Kemmerer terms deterministic shadow logging. Every recommendation executes in parallel with a rules engine, and the system commits the input, the prompt version, the model release hash, and the final output to an immutable store. When an

auditor asks how a transaction was parsed six months previously, the team replays the exact prompt state against raw on-chain data.

> *The IRS requires 100% mathematical auditability on tax filings. Every lot ID must map back to a verified blockchain hash with zero variance. A model returning 99.2% accuracy meant 8 out of every 1,000 transactions contained subtle hallucinations on gain/loss figures. That created an immediate tax liability for our users.*
>
> DAVID KEMMERER, COINLEDGER

Hughes described an equivalent architecture from a different starting point. The model proposed and composed the language, and the engine computed any value requiring defense. His formulation is exact: a figure that had to survive an audit was computed deterministically, and only the prose surrounding it was generative.

The failure column shows the same pattern. Seven respondents attempted to retrofit records after constructing a system. Five accepted supplier certification instead of internal validation. Five supplied periodic reports where auditors required continuous evidence. Four presented accuracy metrics as evidence of compliance.

Three respondents named a subtler failure mode. Cioffi instructed a model to withhold a category of data, and an auditor rejected the prompt as a control. His conclusion extends beyond the project: a prompt provides guidance, and the boundary must lie in identity and retrieval. Hughes reached an equivalent position independently. A prompt cannot be attested; it is not versioned in a form reviewers accept, and its behavior changes when the underlying model changes.

Jose Bohorquez, President and Founder of the medical device cybersecurity firm CyberMed, named the version of this failure that recurs in regulated device work. Teams attempted to satisfy reviewers with general machine learning performance metrics. The Food and Drug Administration requires traceability from requirement through risk to test evidence, and a performance dashboard does not substitute for that chain.

Ritwick Dey, co-founder and Chief Technology Officer of Panto AI, reported a failure that contradicts common guidance. His team produced extensive post-hoc explainability artifacts as the primary compliance deliverable. Auditors found the output noisy and unstable across retraining cycles, and preferred concise provenance records and standardized metric attestations to dense explanation dumps.

Remediation cost (RQ2)

Six respondents quantified the time spent on remediation.

Figure 7. What the redesign cost

Respondent	Remediation work	Reported duration
Richard Schreiber, RAS Consulting Group	Client intake automation moved off a general-purpose cloud model	6 weeks
David Ghozland, Credentiable	Human reviewer returned to the provider-verification approval step	A quarter
David Kemmerer, CoinLedger	Transaction reconciliation rebuilt on a deterministic engine	4 months, about $140,000
Kirill Meshyk, Unidata	Consent and provenance metadata rebuilt into the training pipeline	6 months
Ritwick Dey, Panto AI	Clinical document triage moved to an on-premises enclave	9 to 12 months
Russell Sean, QuanMed AI	Change-control records rebuilt for FDA review	More than a year

Reported durations range from six weeks to more than a year. Every reported figure represents the cost of adding evidence to a system that already functioned as intended. No respondent reported time expended on improving model accuracy.

Kirill Meshyk, Head of AI Data Collection at Unidata, spent six months rebuilding a data collection system so that every training record carried immutable consent metadata. His team was constructing a clinical predictive model for connected medical devices. Auditors rejected the initial training sets because the firm could not produce consent evidence for individual patient records.

Russell Sean, Chief Executive Officer and Chief Technology Officer of QuanMed AI, reported the longest delay in the sample. His firm planned an adaptive diagnostic feature and suspended it for more than a year. The team spent that period rebuilding its records to satisfy FDA expectations for a Predetermined Change Control Plan. His summary is the sharpest observation in the corpus: the AI functioned, and the records surrounding it did not.

Two additional respondents described remediation lasting several weeks but did not provide a figure.

The economic implication is direct. Evidence suggests that infrastructure imposes modest costs when designed into a system and high costs when reconstructed under external pressure.

Forward expectations (RQ3)

Item 3 asked where the compliance burden would increase or decrease over the next 12 to 24 months.

Figure 8. Where the burden goes next

Expectation	Respondents
Heavier	30
Heavier first, then lighter	3
Lighter overall	0

Thirty of thirty-three respondents expect an increase. Three expect an initial increase followed by relief as standards mature. No respondent expects an overall reduction.

The domains are named as an expanding cluster. Continuous monitoring displaces one-time pre-launch approval. Evidence that a named individual reviewed an output displaces a policy asserting that review occurs. Explanation of individual decisions displaces aggregate performance reporting. Supplier accountability propagates down the supply chain as buyers transmit their own rules outward.

Hughes named a fifth domain that published surveys rarely capture. Agent systems execute many steps, invoke tools, generate intermediate decisions, and record almost none of the sequence of events. A single model invocation is, at a minimum, loggable as both input and output. Reconstructing the behavior of a multi-agent system, and the reasoning behind it, is an engineering problem before it is a compliance problem.

The three respondents anticipating eventual relief agree on its source. Dey expects harmonized checklists and commercially available compliance tooling to render operational compliance repeatable. Herron expects clearer risk tiers to reduce the effort required to resolve classification disputes. A third respondent, writing from the supplier side, made a similar observation about standardized audit formats, as each customer currently reconstructs its own review procedure.

Relief in this sample comes from the standardization of the evidence format. No respondent anticipates relaxation of the underlying rule.

Sector analysis

Healthcare and the Internet of Medical Things

Nine respondents work in healthcare delivery or medical device making, the terrain covered by custom healthcare software development. Their constraints separate into two layers with distinct consequences.

The first layer governs data. The Health Insurance Portability and Accountability Act appeared in seven responses across the full sample. Jonathan Freed, Chief Executive Officer of the residential detoxification facility Reprieve House, described the most restrictive variant. His AI-assisted intake records workflow required a redesign because 42 CFR Part 2 imposes consent and redisclosure rules that are stricter than those governing ordinary medical records. Francisco Ortiz, Lead Forensic Mental Health Evaluator at District Counseling, reported a related constraint producing an unusual failure. Machine translation of trauma narratives from Spanish into English flattened idiom and culturally specific meaning, and in visa evaluations, a single imprecise sentence alters how large mental abuse is understood.

The second layer governs change, and it produces the longest delays recorded in this study. Bohorquez stated the operative principle: a model that a manufacturer retrains is, in the view of the Food and Drug Administration, a device that the manufacturer changed. Any update to the substance can trigger fresh validation and the creation of new records. He has observed teams reduce AI features to static, locked models to make a submission achievable, and others abandon cloud retraining plans entirely.

The Food and Drug Administration established a mechanism addressing this constraint. A Predetermined Change Control Plan allows a manufacturer to pre-authorize specific model updates without a new marketing application. The agency finalized guidance in December 2024 and updated it in August 2025 (FDA, 2025a).

Adoption of that mechanism is instructive. The agency had authorized more than 1,350 AI-enabled devices by early 2026, about double the 2022 total (FDA, 2026). As of September 2024, 1.9% of authorized AI and machine learning devices carried an authorized change control plan (Wu et al., 2025). The regulatory accommodation exists and attracts minimal traffic.

Bohorquez grounded a forecast in precedent. The agency applied its 2023 cybersecurity guidance permissively during the first year and grew much stricter as manufacturers had opportunity to adapt. He anticipates an equivalent trajectory for AI-enabled devices.

Two clinicians described a validation constraint independent of either layer. Dr. Raymond Douglas, an oculoplastic surgeon at Cedars-Sinai, sought to apply AI to patient scans in order to track disease progression. Demonstrating that the model held across different imaging equipment and patient populations consumed much more time than planned. His team now logs every training source, every test, every known blind spot, and every case in which the model and the physician disagree. Dr. Alexander Acosta, Medical Consultant at SonderCare, suspended a charting pilot for gynecology and obstetrics notes after clinical staff found errors. He reports

that about one-third of drafted notes needed physician correction.

Clinician caution finds support in larger research. The 2026 Future Ready Healthcare Survey found that 74% of clinicians rank hallucination among the main clinical risks of AI. Only 27% report awareness of their own firm's governance rules (Wolters Kluwer and Ipsos, 2026).

Financial technology and financial services

Nine respondents work in financial technology, where fintech software development carries auditability duties from the first design review. Their constraints converge on reproducibility, defined here as the capacity to execute a decision procedure twice and obtain identical output.

Robin Lahiri, founder of the tax compliance firm einSearch.IO, described the most acute instance. Auditability standards applied by the Internal Revenue Service to B-Notice prevention compelled redesign of an AI-driven vendor exception system, relocating it from automatic resolution to deterministic controls with human review. Returning a probabilistic match score proved unacceptable to accounts payable teams exposed to financial penalties for improperly filed tax identities.

Arpit Mittal, an independent researcher and IEEE Senior Member, named a technical failure mode that compliance discussions rarely address. Deep learning models frequently detect more fraud than simpler alternatives, but deployment stalls because auditors reject a decision that is hard to explain. Early architectures often require redesign to eliminate temporal look-ahead bias in the feature set before they satisfy audit standards.

Yogesh Kansal, founder and Chief Executive Officer of Sliq Pay, described the architectural consequence. Banking partners and auditors required an explanation for every flagged transaction, which moved his team away from deep networks toward simpler ensemble models. The team now logs feature weights and decision paths for every alert.

Sivakumar Dhanasekar, an enterprise architect with 18 years in financial services, provided the most complete specification of what an external review demands. He listed the questions his team could not answer following a technically successful pilot. Which version of the data produced a given result. Which model version, prompt, rule set, and threshold were active at that moment. Whether the same decision is repeated months later. How the firm detects performance degradation. What occurs when a third-party model changes without notification. The project was delayed and redesigned. His conclusion holds beyond his own firm: an accurate model and a production-ready model are distinct artifacts.

The regulatory position in the United States shifted during data collection. The Federal Reserve replaced SR 11-7, its long-standing model risk management guidance, with SR 26-02. The agencies excluded generative and agentic AI from the scope of the current revision, calling those technologies novel and fast-moving. They also signaled a later request for information (Risk.net, 2026). One respondent cited SR 11-7 as operative, and it is no longer operative.

That exclusion has not reduced pressure on the firms in this sample, and the finding on procurement explains why. Buyer diligence and banking partner requirements operate independently of whether a supervisory letter names generative AI.

Larger surveys converge with these findings. The Cambridge Centre for Alternative Finance surveyed 628 respondents, among them 130 regulators. Some 79% of regulators rate explainability as critical or important, while 37% of industry firms treat model opacity as an operational risk (CCAF, 2026). Wolters Kluwer surveyed 148 financial institutions and found explainability the most acute regulatory concern, at 28.4%. A further 58.8% of banks named clearer guidance as their main barrier (Wolters Kluwer, 2026).

Regulatory context

Three regulatory developments frame the forecasts recorded above.

- **The EU AI Act deferred its high-risk rules.** The European Parliament approved amendments on 16 June 2026 that postponed application of the high-risk regime. Standalone Annex III systems, a category that includes credit scoring and recruitment tools, now apply from 2 December 2027 rather than 2 August 2026. Artificial intelligence embedded in regulated products under Annex I, a category that includes medical devices, applies from 2 August 2028 rather than 2 August 2027 (Morgan Lewis, 2026). The absence of harmonized standards drove the deferral, and CEN-CENELEC deliverables are anticipated toward the end of 2026. Transparency rules under Article 50 retain their original schedule (Gibson Dunn, 2026).

One respondent stated that the high-risk rules took effect in August 2026 and that medical device AI has until August 2027. Both dates precede the amendments. This study corrects the record here and excludes those claims from the analysis.

- **United States device regulation tightened.** The Quality Management System Regulation took effect on 2 February 2026, aligning device quality requirements with ISO 13485. Section 524B of the Federal Food, Drug, and Cosmetic Act governs connected devices, requiring manufacturers to submit a vulnerability management plan and a software bill of materials with premarket applications. Failure to comply is prohibited.
- **United States model risk guidance narrowed in one respect.** SR 26-02 replaced SR 11-7 and excluded generative AI from scope. The agencies have signaled that further guidance will follow, and not that these systems escape supervision.

The direction of regulatory travel varies by jurisdiction, while the rule to produce evidence appears in every jurisdiction represented in this sample.

Discussion

Interpretation

The main finding of this study concerns the location of the constraint. Compliance rules in this sample did not terminate AI projects, nor did they primarily challenge model performance. They challenged company capacity to reconstruct a decision after the fact, and firms responded by rebuilding systems until that capacity existed.

Three observations support that interpretation. Nineteen of thirty-three primary barriers concern evidence production. Every quantified remediation cost reported above represents evidence work rather than accuracy work. The most frequently reported failure, retrofitted records, is a failure of evidence architecture and not of modeling.

A second finding concerns the channel through which rules arrive. Eight respondents named buyer-side security review as their binding constraint, exceeding the counts for the General Data Protection Regulation and SOC 2. Regulated buyers transmit their own rules to suppliers through procurement, and suppliers therefore encounter regulatory pressure without falling under direct supervision. That mechanism explains why the exclusion of generative AI from SR 26-02 produced no observable relief among the financial technology respondents in this sample.

Comparison with published research

The findings align with large-sample research on three points and extend it on one.

Grant Thornton found that 18% of banking executives express confidence in passing an independent audit of AI controls (Grant Thornton, 2026). That figure and the barrier distribution reported here describe the same deficiency from opposite directions.

The Cambridge Centre for Alternative Finance recorded a gap between regulator and industry perception of explainability, at 79% and 37% respectively (CCAF, 2026). The remediation practices provide a mechanism to address that gap. Respondents who invested in explainability artifacts reported that auditors preferred provenance records, which suggests that the two populations may be describing different requirements under a shared term.

The Pacific AI governance survey recorded use below 20% for model cards and similar controls (Pacific AI, 2025). The remediation cost findings supply an explanation grounded in cost. Organizations construct such controls only after external pressure arises, and the rebuilding process consumes months.

This study extends the published literature by cross-tabulating named rules against project outcomes. No large-sample survey currently reports which rule produced which outcome, and this study supplies that distribution for 33 firms.

Implications for practice

Five implications follow from the findings, each supported by a specific result.

- **Design the evidence trail before constructing the model.** This appears twice in the data: as the most frequent successful practice and as the most frequent failure when the order is reversed. Seven respondents attempted the reverse order and rebuilt. The evidence auditors requested is a record of system behavior bound to a model version, a data snapshot, a named approver, and a timestamp.
- **Treat model identity as a controlled configuration item.** Bulanek learned this through an incident. A version pin in one location overrode the intended model selection without notice, while a provider retired a preview model beneath his product. Scans continued executing, and not in the manner the records described. A firm unable to state which model produced a given output on a given date possesses no audit position.
- **Compute any value subject to audit, and generate only the language surrounding it.** This separation is a standing design principle in custom AI solutions built for supervised buyers. Three respondents adopted this architecture independently while working in different areas of finance. The separation supplies the language quality of a generative system alongside the defensibility of a deterministic one.
- **Conduct independent diligence on suppliers.** Five respondents cited supplier certification as an accepted internal validation among their failures. A SOC 2 report or a compliance label describes the supplier's controls. It describes nothing about how a model handles a given data class, whether prompts are retained, what a subprocessor change produces, or how long the data persists.
- **Record human review as structured data.** Schreiber's warning applies broadly, because retrofitting a supervision record onto a year of unlogged output is the expensive version of this problem. Every forecast recorded under forward expectations names evidence of supervision as an expanding requirement.

Conclusion

Thirty-three firms operating under regulatory supervision described how compliance rules affected their AI projects. Twenty-two redesigned a system, and two abandoned one. The rules behind those outcomes concerned the rebuilding of decisions far more often than the accuracy of models. The firms that absorbed the highest costs tried to supply evidence after constructing a system.

The forecast recorded in this study is close to unanimous in direction. Thirty of thirty-three respondents expect their compliance burden to increase within twenty-four months, and the three anticipating eventual relief locate it in standardization of the evidence format. Two events left that expectation intact: the deferral of the EU AI Act high-risk deadlines, and the exclusion of generative AI from US model risk guidance. A large share of the pressure comes from customer procurement, not from supervision.

Future research should test whether the barrier distribution reported here holds under probability sampling, and whether the association between evidence-first design and shorter remediation cycles survives controlled comparison.

References

- BDO (2025). *Survey of 210 senior finance leaders on AI governance*. Reported in Corporate Compliance Insights, November 2025.
- Cambridge Centre for Alternative Finance (2026). *2026 Global AI in Financial Services Report*. Cambridge Judge Business School. 628 respondents, including 130 regulators.
- Deskpro (2025). *State of AI in Support Operations*. Survey of more than 220 professionals.
- Food and Drug Administration (2025a). *Marketing Submission Recommendations for a Predetermined Change Control Plan for AI-Enabled Device Software Functions*. Final guidance.
- Food and Drug Administration (2026). *AI in Software as a Medical Device*. Device authorization listing.
- Gibson Dunn (2026). *EU AI Act Omnibus Agreement: Postponed High-Risk Deadlines and Other Key Changes*.
- Grant Thornton (2026). *Banking Insights: 2026 AI Impact Survey Report*. Survey of 950 executives.
- Morgan Lewis (2026). *EU Approves Delays and Other Amendments to Certain EU AI Act Obligations*.
- Pacific AI (2025). *2025 AI Governance Survey*. More than 350 respondents.
- Risk.net (2026). *Do banks still need to validate GenAI models?*
- Wolters Kluwer (2026). *Q1 2026 Banking Compliance AI Trend Report*. Survey of 148 financial institutions.
- Wolters Kluwer and Ipsos (2026). *Future Ready Healthcare Survey*.
- Wu, K. et al. (2025). *Regulating Flexibility for AI: FDA Experience with Predetermined Change Control Plans*.

How to cite this research

Merzlova, K. (2026). *The evidence gap: Compliance and audit barriers to the adoption of artificial intelligence in healthcare and financial services.* SumatoSoft Research. https://sumatosoft.com/blog/compliance-and-audit-barriers-to-ai-adoption-in-regulated-industries

Reuse of figures and tables. The eight figures and the data tables accompanying them may be reproduced with attribution to SumatoSoft and a link to this page. Quotations from named participants remain the words of those participants and should bear their names and companies.

Data availability. The coding scheme and the participant roster are published in full below. Raw responses are not published because participants supplied them for attributed quotation and not for redistribution.

Appendix A: Respondent roster

Healthcare and the Internet of Medical Things

Respondent	Position	Company	Jurisdictions
Jose Bohorquez, PhD	President and Founder	CyberMed	US
Russell Sean	CEO and CTO	QuanMed AI	US
Raymond Douglas, MD	Oculoplastic surgeon	Private practice and Cedars-Sinai	US
Jonathan Freed	Owner and CEO	Reprieve House	US
Francisco Ortiz	Lead Forensic Evaluator	District Counseling PLLC	US
Alexander Acosta, MD	Medical Consultant	SonderCare	US, Canada
Cassidy Blair, PsyD	Founder	Blair Wellness Group	US
David Ghozland, MD	Co-founder	Credentiable, and Intimate Health Center	US
Maria Knöbel, MD	Medical Director	MedicalCert UK	UK

Financial technology

Respondent	Position	Company	Jurisdictions
David Kemmerer	Co-founder and CEO	CoinLedger	US, EU, UK, Australia
Chase W. Hughes	Founder	ProAI, Pro Business Plans and Equity Up	US, EU, Gulf states
Sivakumar Dhanasekar	Engineering Lead	Equifax	US
Yogesh Kansal	Founder and CEO	Sliq Pay	US, India
Arpit Mittal	Independent researcher	IEEE Senior Member	US
Sergiy Fitsak	Managing Director	Softjourn	US, Ukraine
Robin Lahiri	Founder	einSearch.IO	US
Lily Stoyanov	CEO and Founder	Transformify	EU, UK
Heath Squier	Chief AI Officer	Equity Edge Lending, and EVKII	US

Suppliers of AI to regulated buyers

Respondent	Position	Company	Jurisdictions
Viktor Bulanek	Founder and CTO	Penetrify	EU, US, UK
Egiziago Cioffi	CEO and Enterprise Architect	SynSphere	EU (Italy), UK
Ritwick Dey	Co-founder and CTO	Panto AI	US
Kirill Meshyk	Head of AI Data Collection	Unidata	US, UK, EU
Ken Herron	Co-founder	VCONify	US

Comparison group: adjacent regulated sectors

Respondent	Position	Company	Jurisdictions
Richard Schreiber	CEO	RAS Consulting Group	US
Anthony Guerriero	Co-founder	The Leveraged Years	US, EU
Yury Byalik, JD	Founder	CivilCase	US
Samuel Landis, Esq.	Tax attorney	Segal, Cohen & Landis	US
Nick Heimlich	Founder and Attorney	Nick Heimlich Law	US
Peter Jaraysi, Esq.	Founder	Slam Dunk Attorney	US
Phillip Hamnett	Founder and CEO	TalentAid	Germany
Blake Smith	Marketing Manager	ClockOn	Australia
Tobias Trost	CEO and Founder	ERP Pilot	Switzerland, EU
David Hunt	COO	Versys Media	UK, EU, US, South Africa

Appendix B: Framework glossary

Framework	Respondents	Scope as invoked in this study
EU AI Act	17	Risk classification, technical documentation, human oversight, post-market monitoring
FDA guidance	9	Software as a medical device, predetermined change control, total product lifecycle expectations
HIPAA	7	Protected health information handling, audit logging, business associate obligations
Model risk management	7	Independent validation, conceptual soundness, ongoing monitoring
GDPR	4	Lawful basis, Article 9 special-category data, Article 22 automated decisions
SOC 2	4	Service company controls, invoked by buyers more often than by regulators
Bar and professional conduct rules	4	Client confidentiality, unauthorized practice, supervision of assisted work
IRS reporting rules	3	Audit trail obligations for filings and vendor identity verification
DORA	1	Threat-led penetration testing for financial entities in the European Union
NIS2	1	Network and information security obligations
PCI-DSS	1	Cardholder data handling
42 CFR Part 2	1	Substance use treatment records, consent and redisclosure
MHRA and GMC	1	United Kingdom device regulation and clinical accountability

Framework and remediation codes

Framework codes and remediation-practice codes permit multiple values per respondent. The framework glossary lists framework counts, and the section on remediation practices lists remediation counts, which total 33 successful practices and 26 failed practices across 33 respondents.

About this study

SumatoSoft is a custom AI and software development company with headquarters in Boston and a development center in Warsaw. The company has delivered more than 350 projects across 25 countries over 14 years and holds ISO 27001 and ISO 9001 certification.

Data collection ran from June to August 2026. Figures attributed to individual respondents are self-reported and were not independently verified. Where a respondent's regulatory claim conflicted with a primary source, the primary source governs, and the correction appears in the text. The conflict of interest section states the interest applying to this study.