WHITE PAPER · AI & INNOVATION

AI Governance for Clinical Research Networks: A Strategic Framework

Clinitiative Health Research April 2026

CONTENTS

Executive Summary

1. The Five Pillars of AI Governance

2. Phased Implementation Roadmap

3. Use-Case Tiering in Detail

4. Operational Metrics for AI Governance

5. Network Performance Outcomes

6. Conclusions

Executive Summary

From Tool Adoption to Governed Capability

Artificial intelligence has moved decisively from pilot to production across clinical research. AI-driven patient identification, eligibility screening, dose-response modeling, safety signal detection, and post-market pharmacovigilance are now operational realities at sites and sponsors across the industry. The question facing clinical research networks is no longer whether to adopt AI, but how to govern it: how to extract operational value while satisfying the documentation, credibility, and oversight expectations now forming in the regulatory record.

The FDA's January 2025 draft guidance "Considerations for the Use of Artificial Intelligence to Support Regulatory Decision Making for Drug and Biological Products" introduced a 7-step risk-based credibility framework that sponsors must apply when AI outputs influence regulatory decisions. The joint FDA/EMA Guiding Principles published in January 2026 aligned international expectations around ten high-level

commitments — most notably that AI should support, not replace, human regulatory and clinical decision-making. The final FDA guidance, expected in Q2 2026, will operationalize these principles into specific submission expectations.

This paper presents the AI governance framework the Clinitiative network has developed for its own operations and shares it as a reference architecture for the industry. The framework spans use-case classification, credibility documentation, demographic bias monitoring, human oversight protocols, vendor due diligence, and ongoing operational measurement. It is designed to be practical: each component maps to a specific operational artifact that a network can implement, audit, and improve.

1 The Five Pillars of AI Governance

The framework rests on five operational pillars. Each pillar addresses a category of governance risk and produces specific artifacts that anchor the network's ability to demonstrate credibility, oversight, and equity.

Use-Case Classification

Every AI use case in the network is classified by regulatory impact tier. Tier A applications directly influence regulatory submissions and require full credibility documentation under the FDA 7-step framework. Tier B applications inform operational decisions with regulatory implications and require lighter but documented credibility evidence. Tier C applications are operationally internal and require ethics and bias documentation but not formal credibility files. Classification determines what governance investment a use case warrants.

Credibility Documentation

For Tier A and Tier B use cases, the network maintains standardized Credibility Assessment Plans aligned to the FDA 7-step framework: question of interest, context of use, model characterization, risk assessment, credibility plan, execution results, and adequacy determination. CAPs are versioned, signed off by responsible parties, and made available to sponsors and regulators on request.

Demographic Bias Monitoring

All AI-augmented patient recruitment activities are subject to demographic bias monitoring. Network sites compare the demographic composition of AI-flagged candidates against the broader EHR population on a defined cadence and report deviations to the sponsor. Persistent skew triggers model recalibration or eligibility logic review before bias compounds across enrollment.

Human Oversight Protocols

AI outputs that influence clinical decisions — eligibility determinations, safety assessments, adverse event adjudication — flow through documented human verification protocols. The protocol specifies the qualified reviewer, the verification criteria, the documentation standard, and the audit trail. Investigators cannot delegate clinical judgment to AI outputs; the protocols make the human review visible and auditable.

Vendor & Operational Discipline

AI vendors and platforms are evaluated against a standardized due diligence rubric covering training data provenance, model versioning, change management, documentation quality, and audit access. Operational metrics — accuracy against human verification, demographic representation, processing latency — are tracked continuously and reviewed quarterly. The vendor relationship is governed as infrastructure, not as a procurement transaction.

2 Phased Implementation Roadmap

Implementing AI governance is not a single project but a sequenced organizational change program. The Clinitiative network deployed the framework across five phases over an 18-month window, and we recommend a similar staging for networks beginning their governance build.

Inventory & Classification (Months 0–3)

The first phase produces a complete inventory of AI applications operating across the network. Most networks discover during this exercise that AI usage is broader than centrally tracked — coordinator tools, scheduling assistants, and screening platforms acquired study-by-study often escape enterprise visibility. The inventory is then classified by the Tier A/B/C scheme to focus governance investment where regulatory risk is highest. Inventories should be refreshed semi-annually.

Credibility Documentation Build (Months 3–9)

Tier A and Tier B use cases receive structured Credibility Assessment Plans aligned to the FDA 7-step framework. Networks should expect 60-90 days of documentation work per Tier A use case, less for Tier B. The phase is also when standardized CAP templates are developed, sign-off authorities are established, and version control infrastructure is deployed. Networks that skip the template phase and write each CAP from scratch consistently report 2-3x higher documentation cost.

Bias Monitoring & Reporting (Months 6–12)

Demographic bias monitoring is operationalized across AI-augmented recruitment workflows. Sites begin reporting AI-flagged candidate demographics alongside enrollment demographics as a standard part of monthly sponsor reporting. The phase typically surfaces at least one use case where AI screening systematically under-represents a demographic group; the discovery is the first proof point that the governance framework is doing its job.

Human Oversight Protocols (Months 9–15)

Human oversight protocols for AI-influenced clinical decisions are formalized, embedded in CTMS workflows, and trained across investigator and coordinator staff. The protocols typically need 2-3 rounds of refinement based on operational feedback before they reach stable form. Networks that under-invest in this phase find that their CAPs document a model's technical credibility but cannot demonstrate the human-in-the-loop oversight that the joint FDA/EMA principles emphasize.

Continuous Measurement & Refresh (Months 15+)

AI governance becomes a steady-state operating discipline. Quarterly metric reviews cover accuracy against human verification, demographic representation, and operational throughput. Vendor performance is rated annually. The use-case inventory is refreshed semi-annually. The framework is updated when FDA, EMA, or ICH publish new guidance — and at this scale of change, sponsors should expect annual framework revisions for the foreseeable future.

3 Use-Case Tiering in Detail

The Tier A/B/C scheme allows networks to scale governance investment to regulatory risk. Misclassifying a Tier A use case as Tier C is the most common — and most consequential — governance failure mode.

Tier A — Regulatory Evidence

AI outputs that directly inform regulatory submissions or clinical decisions reflected in the final clinical study report. Examples: AI-driven endpoint adjudication, dose-finding model recommendations, AI-augmented safety signal detection feeding DSMB review. Tier A applications require full FDA 7-step credibility documentation, sign-off by qualified senior staff, and inspection-ready audit trails.

Tier B — Operational with Regulatory Reach

AI outputs that shape operational decisions whose downstream effects are visible to regulators. Examples: AI-augmented patient screening (where enrollment demographics will be reviewed during regulatory submission), AI-driven query generation feeding source data verification. Tier B applications require streamlined credibility documentation focused on accuracy, bias, and human oversight — not full statistical validation of the model architecture.

Tier C — Internal Operations

AI outputs limited to internal operations with no regulatory submission impact. Examples: scheduling assistants, internal training content generation, productivity tooling. Tier C applications require ethics and bias documentation but not formal credibility files. The classification is reviewed annually because scope creep — a Tier C tool gradually being used for Tier B purposes — is a recurring governance challenge.

Common Misclassification Risks

The most frequent misclassification is treating AI-augmented patient screening as Tier C when it is properly Tier B. Networks treating screening as a back-office efficiency tool fail to monitor demographic representation, fail to document human verification protocols, and consequently find themselves unable to demonstrate credibility when a sponsor or regulator inquires. Classification reviews should specifically test screening, query generation, and adverse event triage use cases.

Cross-Tier Governance Boundaries

Some platforms span multiple tiers — a single AI vendor might provide screening (Tier B) and scheduling (Tier C). Governance must apply tier-specific documentation to specific functions rather than treating the platform as monolithic. Networks that vendor-level govern under-document the Tier B functions; networks that function-level govern from the outset avoid retroactive documentation work later.

Reclassification Triggers

A use case's tier may shift as its operational footprint expands. Triggers include: AI outputs being incorporated into sponsor reporting, AI recommendations being audited by regulators, AI outputs influencing protocol-level decisions, and AI workflows being relied upon during inspection. Each trigger should prompt a documented reclassification review with updated credibility evidence.

4 Operational Metrics for AI Governance

Governance is only credible if it is measured. A documented framework that produces no operational signals will not detect drift, bias, or capability degradation. The metrics presented below have proven the most useful in operating the framework at network scale and align with what regulators are increasingly asking sponsors to report.

Accuracy against human verification — the rate at which AI outputs agree with qualified human review — is the foundational metric. A target threshold of 88-92% agreement is appropriate for most Tier B applications; lower accuracy may indicate model drift, training data mismatch, or eligibility logic configuration error. Demographic representation ratios — comparing AI-flagged candidate demographics against the underlying EHR population — surface bias issues before they compound across enrollment. Processing latency tracks operational fitness for the workflows AI is supposed to accelerate.

Beyond these operational metrics, networks should track governance maturity indicators: percentage of AI use cases with current CAPs, percentage of bias monitoring reports completed on schedule, time-to-resolution for governance issues raised in quarterly review. These second-order metrics measure whether the framework is functioning as a discipline rather than as a paper exercise.

5 Network Performance Outcomes

The framework has been operating across the Clinitiative network for 18 months. The performance outcomes that follow illustrate the operational discipline that governed AI use enables.

Table 1. Network Performance Outcomes: key figures

Figure	Measure	Detail
100%	Tier A Coverage	Every Tier A AI use case in the network has a current Credibility Assessment Plan aligned with the FDA 7-step framework. CAPs are versioned, signed off, and available to sponsors and regulators within 48 hours of request. No Tier A use case is allowed to operate without a current CAP.
91%	Human Verification Agreement	Network-wide, AI-augmented eligibility determinations agree with investigator review at a 91% rate. The 9% disagreement rate concentrates in complex eligibility criteria involving prior treatment history and concurrent medication review — areas where the human oversight protocol is essential.
2	Bias Issues Detected & Resolved	Bias monitoring has detected two demographic skew issues across AI-augmented studies in the last 18 months — one in cardiology, one in rheumatology. Both were resolved through eligibility logic recalibration before they materially affected enrollment demographics, demonstrating the framework's preventive value.
0	Regulatory Findings	Across the AI-augmented studies that have been inspected during the framework's operating period, no AI-related regulatory findings or 483 observations have been issued. The combination of credibility documentation, human oversight protocols, and bias monitoring has held up to regulatory review.

6 Conclusions

AI is no longer a future capability; it is current infrastructure. The networks and sponsors who treat it as governed capability will extract its operational value while preserving the regulatory standing their development programs depend on. Those who treat it as an unmanaged efficiency tool will find their use cases increasingly contested as the regulatory framework matures.

The five-pillar framework presented here is not the only valid approach, but it has proven operationally robust at network scale. The most important takeaway is not the specific structure but the discipline: classify use cases, document credibility, monitor for bias, govern human oversight, and measure continuously. Each of those is achievable today, with current tools and current organizational capacity. The cost of not doing them — in regulatory findings, in operational drift, in lost sponsor trust — is far higher than the cost of doing them well.