Paper ML26

Validating AI in GxP: Test Harnesses, Guardrails, and Compliance Controls

Bill Qubeck, Sycamore Informatics, Cary, NC
Heather Gorr, Sycamore Informatics, Boston, MA

ABSTRACT

AI adoption fails without credible validation. This article provides a fit-for-purpose validation approach for LLM-assisted authoring and review tools: requirements decomposition, risk assessment, testable acceptance criteria, and a repeatable test harness with seeded scenarios. We capture prompts, inputs, outputs, and adjudications as validation evidence and monitor model behavior over time to detect drift. The design integrates model cards, usage policies, and access controls so responsibilities are clear. Attendees learn how to convince QA that AI sits inside a controlled process, which risks must be mitigated, and how to keep validation lean enough to evolve with rapid AI change while meeting inspection expectations.

INTRODUCTION: THE REAL PROBLEM WITH AI IN CLINICAL TRIAL SUBMISSIONS

AI adoption is not failing in clinical trial regulatory submissions because it is technically infeasible. It is failing because AI systems are being introduced into controlled processes without sufficient evidence, boundaries, or lifecycle oversight. Sponsors and CROs are discovering that fluency is not a quality standard and that "the model usually gets it right" is not validation. In a submission context, the requirement is straightforward but demanding: demonstrate control, traceability, risk mitigation, and managed change. The core mismatch lies in the nature of large language models. They are probabilistic systems capable of producing convincing text even when underlying claims are incorrect, incomplete, or untraceable. Their behavior can also change over time due to vendor updates, configuration changes, retrieval corpus updates, or subtle prompt modifications. In non-regulated content creation, this variability may be acceptable. In clinical trial analysis and regulatory reporting, it represents a direct threat to data integrity and inspection readiness. Every statement in a submission artifact must be supported by traceable evidence. Plausibility without proof is not acceptable. This is why validating the model itself is the wrong objective. GxP expectations are not about proving that an AI system is universally correct. They are about proving that the computerized system and the business process using it remain in a validated state throughout their lifecycle. Practically, this means AI must operate inside a controlled workflow with a clearly defined intended use, explicit responsibilities, enforceable controls, and objective evidence that the system behaves acceptably under both expected and adverse conditions. Lifecycle thinking, transparency, and ongoing monitoring are not optional add-ons. They are prerequisites for defensible use. Clinical trial submissions amplify these challenges because the artifacts are inherently interdependent. A Clinical Study Report is not a standalone document. It is an integrated narrative tied directly to statistical methods, analysis populations, endpoint definitions, and tables, figures, and listings derived from analysis datasets. Reviewers and health authority auditors focus precisely on consistency across these components. When AI is introduced, the primary risk is not a single incorrect sentence. It is subtle cross artifact inconsistency that is difficult to detect without engineered controls.

This is where most AI programs fail. They begin with pilots that impress business stakeholders and immediately alarm Quality Assurance. QA is right to be concerned when basic inspection questions cannot be answered. What is the intended use? What risks have been identified and mitigated? How are errors detected? Where is the objective validation evidence? What controls prevent misuse? How is drift detected as system behavior changes over time? Who approves changes, and what change control governs prompts, models, and data sources? What audit trail exists for prompts, inputs, outputs, and human adjudications? If the only evidence consists of screenshots, anecdotes, or one time demonstrations, the system is not inspectable. In a GxP environment, undocumented behavior is indistinguishable from uncontrolled behavior. Another frequent failure mode is poor use case selection. When AI is positioned as an autonomous author or approval surrogate, it triggers justified procedural and cultural resistance. A more defensible approach is to define a constrained assistant role with explicit boundaries and testable behavior. In this model, AI supports the generation and review of analysis outputs and Clinical Study Report content with full traceability to CDISC SDTM and ADaM datasets, tables, figures, listings, and analysis specifications, while decision authority remains with human reviewers. This approach still delivers significant value, but it is controllable. Every factual claim can be required to reference a specific source, and anything that cannot be traced is blocked or escalated. In regulated submission workflows, traceability is not an enhancement. It is the control.

The regulatory environment reinforces this direction. Governance expectations for AI are increasing, with greater emphasis on structured risk management, documentation, transparency, and accountability. Even when a specific internal use case does not fall squarely into a formal high risk category, the trajectory is clear. More scrutiny, not less. The real problem is not whether AI can help with clinical reporting. It is that without a fit for purpose validation approach, AI introduces unmanaged variability into one of the most scrutinized artifacts in pharmaceutical development. The solution is not to argue for AI's benefits, but to design controlled processes where AI is constrained, testable, auditable, and continuously monitored with evidence that withstands inspection.

WHY CLINICAL TRIAL DATA ANALYSIS AND REPORTING IS THE RIGHT ENTRY POINT FOR AI

Starting within the Clinical Trial Data Analysis and Reporting process is an ideal candidate. Not because it is easy, but because it is already designed for control: structured inputs, standardized outputs, heavy review processes, and established accountability. Those characteristics let you apply fit for purpose validation methods to AI with far less ambiguity than other GxP processes.

Analysis and reporting are already structured, metadata driven, and review process heavy

Clinical analysis and reporting is one of the most "systems oriented" workflows in the clinical development lifecycle. The work is not an unbounded exercise. It is a chain of linked artifacts designed to be reproducible and auditable: protocol and endpoints, statistical analysis plan, analysis datasets, tables figures listings, and the clinical study report narrative that explains and contextualizes the results. ICH E3 explicitly frames the CSR as an integrated report of an individual study, which is a polite way of saying that the narrative is expected to be consistent with the analysis outputs and traceable to the underlying work products.

That structure matters because AI becomes validateable when the environment is structured. In practice, "structured" means you can define and lock down the inputs that the AI is allowed to see and reference. For submissions, this typically includes analysis specifications, CDISC SDTM & ADaM datasets, define.xml metadata, and finalized TFL outputs. FDA's Study Data Technical Conformance Guide is explicit that sponsors should submit study data in specified formats and with supporting data definition and reviewer guides. That expectation has driven the adoption of standardized datasets and metadata which becomes an advantage when designing AI guardrails and test harnesses.

Analysis and reporting is also review and process heavy by design. The work products are routinely reviewed by statistical leads, programming leads, clinical leads, medical writers, and quality functions. There are formal review cycles, comment resolution logs, and final approvals. In other words, the process already assumes that no single author, human or machine, is trusted to be correct without scrutiny. That is exactly the environment where AI can be introduced in a bounded way: as a drafting and review accelerator that feeds a controlled review workflow. The point is that you are not asking QA to accept a new philosophy. You are aligning AI to how analysis and reporting already works. AI becomes another contributor whose outputs must be traceable, reviewable, and evidence based.

Practical implication for validation design

You can define testable requirements around structure and traceability that are realistic, enforceable, and inspectable, for example:

- Every generated claim must reference a specific TFL identifier or analysis specification section.
- Numerical statements must be traceable to an approved output and must be rejected if untraceable.
- Any ambiguity in populations or endpoints triggers escalation, not inference.
- The system captures the prompt, the retrieved context, the output, and the human adjudication as part of the audit trail.

These are not theoretical controls. They are compatible with how submissions are already assembled and reviewed.

Outputs are regulated, but decision authority already sits with humans

This is the second reason this area is the right entry point. The outputs of analysis and reporting are regulated artifacts, but the regulated process already assigns responsibility to people, not tools. ICH E9 emphasizes statistical rigor across design, conduct, analysis, and evaluation, with responsibility and oversight embedded in the trial team and the statistician function. You are not trying to replace that responsibility with AI. You are trying to reduce cycle time and improve consistency in how humans produce and review those artifacts. This is a decisive difference between defensible and indefensible AI use cases. The moment AI is positioned as an approval surrogate, it becomes a governance and accountability problem that most quality systems are not prepared to handle. In contrast, when AI is positioned as a constrained assistant inside an approval workflow, the accountability model stays intact:

- Humans remain accountable for approvals and final content.
- AI outputs remain draft material and review inputs.
- The system is validated to ensure AI stays within guardrails and that deviations are detected and handled.

The EMA guideline on computerized systems and electronic data in clinical trials is explicit about the need for controls such as validation, SOPs, security, change control, and data handling practices when computerized systems are used in clinical trials. That expectation naturally extends to AI enabled components embedded in computerized systems used to support regulated activities. So the "regulated output" aspect is not an argument against AI. It is an argument for designing the right controls and evidence.

Errors are costly, visible, and inspectable

The third reason is blunt: this is where errors are significantly impactful. Errors in analysis and reporting are expensive because they create rework across multiple functions, delay submissions, and consume additional effort and time. They are also visible because the artifacts are interlinked. If a CSR says one thing and the table says another, it stands out. If population counts do not reconcile across outputs, it stands out. If terminology shifts between the SAP, TFLs, and CSR sections, it stands out. And because submissions are reviewed by regulators these inconsistencies are not merely internal quality issues. They can become regulatory questions about data quality, reliability, traceability, and control of the process. This visibility is an advantage for validation. In many other GxP areas, errors are latent and difficult to detect until downstream. In analysis and reporting, inconsistency is often

detectable through structured checks. That enables a fit for purpose AI validation strategy that emphasizes detectability and containment:

- **Detectability**: the system must be able to identify when AI output is ungrounded, inconsistent, or outside scope.
- **Containment**: the workflow must prevent such output from entering regulated artifacts without explicit review and resolution.
- **Evidence**: the system must capture what happened, why it happened, and how it was resolved.

The EMA computerized systems guideline reinforces that computerized systems used in clinical trials should be designed and controlled to ensure data integrity and reliability, including through validation and documented processes. These principles map directly to AI output controls and drift monitoring. If AI is generating or reviewing narratives tied to clinical data, you must be able to demonstrate that the system continues to behave in a controlled manner over time.

This makes it ideal for fit for purpose AI validation

Put the three points together and the conclusion is straightforward. Analysis and reporting is the right entry point because it already has:

1. Structured artifacts and metadata that can constrain AI behavior and make test scenarios realistic.
2. Established human oversight and approval authority that supports a bounded assistant model.
3. High visibility inconsistency patterns that enable automated checks, objective evidence capture, and repeatable validation.

Fit for purpose validation in this context means you validate what actually matters for compliance and inspection readiness, without pretending you can prove a language model is universally correct. Your validation objective becomes: the AI enabled workflow is controlled, traceable, auditable, and monitored. That is a GxP aligned objective. Concretely, this entry point supports the most defensible primary use case described in the outline: AI assisted generation and review of CSR content and analysis output with full traceability to CDISC SDTM and ADaM and TFL specifications. It also supports two lower risk expansions: AI assisted analysis specification and TFL shell authoring, and AI assisted review and consistency checking across submission artifacts. The common theme across all three is that the surrounding process provides the structure, accountability, and evidence model that QA expects.

INTENDED USE IS THE VALIDATION BOUNDARY

Intended use is the foundation of validation in GxP systems. It defines the scope of regulated activity, the assignment of responsibility, and the limits of acceptable system behavior. When AI is introduced into clinical trial analysis and reporting, intended use becomes the single most important control. Without a precise and enforceable definition of intended use, no amount of testing or documentation can make the system defensible during inspection.

AI is used to assist. Not decide. Not approve.

In regulated clinical trial workflows, authority matters more than capability. The question regulators and QA functions care about is not what the AI can do, but what the AI is allowed to do, and how the system prevents it from doing anything else. When AI is positioned as a decision maker, approver, or release mechanism, it immediately collides with established accountability models embedded in GxP quality systems. Those models assign responsibility to defined roles, supported by computerized systems, not replaced by them. This distinction is well aligned with regulatory expectations for computerized systems. The EMA guideline on computerized systems and electronic data in clinical trials makes clear that computerized systems may support regulated activities, but responsibility and oversight remain with qualified individuals operating within defined processes and SOPs. The same principle appears repeatedly in FDA guidance on computerized systems and data integrity, even though AI is not explicitly mentioned. The system can assist, it cannot own accountability. In Clinical Trial Data Analysis and Reporting, this principle is already deeply ingrained. Statistical results are approved by statisticians. Tables, figures, and listings are approved through controlled review cycles. Clinical study reports are approved by designated signatories. Introducing AI does not change those responsibilities. A defensible system design explicitly preserves them. Therefore, the foundational validation claim must be explicit and testable:

- AI assists humans by generating draft content, comparisons, and flags.
- AI does not make decisions that carry regulatory or scientific authority.
- AI does not approve or release submission artifacts.

If this cannot be enforced technically and/or procedurally, then it is not a claim, rather it is an aspiration and therefore by definition it is not GxP compliant/validated.

The system supports controlled authoring, review, and quality checks

Once the assist only boundary is established, the validation focus shifts from the AI model to the system that surrounds it. In a compliant design, AI is embedded inside existing analysis and reporting workflows, not bolted on as an external content generator. That distinction is critical. Controlled authoring means that AI generated content enters the workflow in a draft state, clearly labeled, with provenance captured. Controlled review means that AI outputs are subject to the same or stronger review expectations as human authored content, including documented adjudication. Controlled quality checks mean that the system actively enforces constraints, such as traceability and scope limitations, rather than relying on users to remember them. This approach aligns with long-standing expectations for computerized systems used in clinical trials. Both FDA and EMA guidance emphasize that systems must be validated

in the context of their intended use, with controls appropriate to the risk of the activity they support. When AI is embedded as a subsystem inside a validated analysis and reporting platform, the validation objective becomes clear: demonstrate that the system consistently enforces boundaries, captures evidence, and prevents inappropriate use. In practical terms, this means the system must do more than generate text. It must:

- Restrict AI access to approved inputs such as finalized TFLs, analysis specifications, and metadata.
- Prevent AI from accessing raw analysis datasets unless explicitly permitted.
- Enforce human review and approval checkpoints before content can be incorporated into regulated artifacts.
- Capture an audit trail of prompts, inputs, outputs, and reviewer decisions.

These are system behaviors that can be tested, documented, and defended.

Explicit intended use statement

A defensible AI validation effort includes a written intended use statement that QA can approve; one that clearly defines its scope and limits. For Clinical Trial Data Analysis and Reporting, an explicit intended use statement might read:

- AI may draft, summarize, compare, and flag content related to analysis outputs, tables, figures, listings, and clinical study report sections, using approved source materials.
- AI may support consistency checking across analysis specifications, outputs, and report narratives.
- AI may not approve any regulatory content.

The power of these statements is in how they drive design and validation. Every one of these statements can be enforceable through technical controls, procedural controls, or both. If the system technically allows AI to modify datasets, then the intended use statement is false. If the system allows AI generated content to bypass review, then the intended use statement is false. In a GxP environment, a false intended use statement is a compliance issue. This approach mirrors how intended use is treated for other regulated computerized systems. FDA guidance on software and electronic records consistently ties validation scope to intended use and risk, not to the underlying technology. AI does not get a special exemption from that logic.

Ambiguity in intended use expands audit scope instantly

One of the most common and costly mistakes teams make is leaving intended use ambiguous. Phrases such as “AI assisted authoring” or “AI supported review” sound reasonable, but they are insufficient in an inspection context. Ambiguity invites questions, and questions expand audit scope. When intended use is unclear, health authority auditors are forced to infer how the system might be used, not just how it is intended to be used. That inference almost always lands on worst case scenarios. Could the AI generate content without review? Could it influence statistical interpretation? Could it introduce untraceable claims? Could it be used outside its original context? Each unanswered question becomes a line of inquiry. This is why intended use must be narrow, explicit, and documented. It is also why intended use must be enforced by design. Regulators and health authority auditors do not rely on user intent. They rely on system behavior. If a system allows misuse, then misuse is in scope, regardless of policy statements. This principle is consistent with regulatory expectations for data integrity. Both FDA and EMA emphasize that systems should be designed to prevent inappropriate use and detect deviations, rather than relying solely on procedural controls. In the context of AI, that means the system must make it difficult or impossible for AI to operate outside its validated role.

The word assistant is meaningless without controls

Calling an AI system an assistant does not make it compliant. Assistant is not a control. It is a label, a description. Without technical and procedural enforcement, the word assistant provides no assurance to QA or regulators. A meaningful assistant role is defined by constraints. These constraints include:

- **Scope constraints:** what content types and data can the AI can access?
- **Functional constraints:** what actions the AI can and cannot perform?
- **Workflow constraints:** where AI outputs enter the process and where they are blocked?
- **Accountability constraints:** who must review, approve, and resolve AI outputs?

These constraints are what validation actually tests. Validation does not test whether the AI is helpful. It tests whether the system behaves according to its intended use under normal and adverse conditions. That includes attempts to push the AI outside its role, whether intentionally or accidentally. When these constraints are in place, the term assistant becomes meaningful. It describes a system that supports human work without undermining control. When they are absent, the term is cosmetic, and cosmetic controls do not survive inspection.

Why intended use defines everything that follows

Intended use is the boundary because it determines: 1) What risks must be assessed? 2) What controls must be designed and tested? 3) What evidence must be collected? 4) What changes trigger revalidation? 5) Etc. In Clinical Trial Data Analysis and Reporting, a narrowly defined assistant role allows teams to apply AI where it adds value, while preserving the integrity of the submission process. It also allows the validation of behavior within boundaries; this is the starting point for AI in GxP.

RISK ASSESSMENT FOR LLMS IN CLINICAL TRIAL ANALYSIS AND REPORTING

For large language models used in clinical trial analysis and reporting, traditional computerized system risk models are insufficient. LLMs introduce failure modes that do not resemble deterministic software defects and are not reliably

detected by casual review. A credible risk assessment must therefore focus on how LLMs fail in practice, how those failures manifest in regulated artifacts, and how the system detects and contains them before they impact submission quality or data integrity. In this context, risk is not defined by the likelihood that the model produces an error in isolation. Rather risk is defined by the likelihood that an error is produced, appears plausible, and escapes detection long enough to enter a regulated submission workflow.

LLM specific hazards mapped to clinical reporting

1. Hallucinated efficacy or safety claims

Hallucination in clinical reporting is not random nonsense. It is the generation of content that is syntactically correct, clinically plausible, and statistically framed, but not supported by the underlying analysis outputs. In the context of clinical study reports, this can include invented trends, implied treatment effects, or narrative interpretations that are not present in the TFLs or other analyses. This risk is particularly acute because CSRs are interpretive documents. They contextualize results rather than merely restating numbers. Without strict traceability enforcement, an LLM can introduce interpretive language that subtly alters the scientific message while remaining difficult to challenge during review. FDA and ICH guidance emphasize that clinical study reports must accurately reflect the study data and analyses performed. Any narrative that cannot be traced back to those analyses represents a data integrity risk, not merely a wording issue.

2. Fabricated citations or table references

A well documented LLM failure mode is the fabrication of citations. In clinical reporting, this failure manifests as references to non-existent tables, incorrect table numbers, or mismatched listings. Unlike hallucinated facts, fabricated references can appear correct at a glance, especially in large reports with many outputs. This hazard directly undermines traceability, which is a core expectation in submission review. Regulatory reviewers rely on internal consistency across datasets, outputs, and narratives to assess reliability. An AI system that invents references, even occasionally, introduces a failure mode that must be explicitly mitigated through controls such as enforced citation validation and automated cross checks.

3. Inconsistent population language across TFLs and CSR sections

Clinical reporting relies heavily on consistent population definitions such as intent to treat, safety population, and per protocol population. These definitions are specified in the SAP and reflected in CDSIC SDTM/ADaM metadata and TFLs. LLMs, however, are prone to paraphrasing and semantic drift. Without constraints, an AI may describe the same population differently in different sections of a CSR, or subtly shift terminology in ways that introduce ambiguity. This is not a cosmetic issue. Population inconsistencies are a common source of regulatory questions because they raise concerns about which subjects were included in which analyses. When AI is involved, this risk increases unless the system enforces strict alignment between narrative language and approved metadata.

4. Prompt injection leading to data leakage

Prompt injection is a security failure mode in which user supplied input manipulates the model into revealing information or bypassing controls. In clinical trial environments, this risk intersects directly with confidentiality and data protection obligations. An LLM embedded in an analysis and reporting platform may have access to sensitive study data, metadata, or internal documents. If prompt injection is not mitigated, users may unintentionally or maliciously cause the system to disclose information outside its intended scope. Regulatory guidance on computerized systems emphasizes the need for appropriate security controls, access restrictions, and prevention of unauthorized data access. LLM specific threats such as prompt injection must therefore be treated as part of the system security risk assessment, not as an abstract AI concern.

5. Automation bias in reviewers

Automation bias occurs when human reviewers suspend critical judgment and/or place unwarranted confidence in system generated outputs, particularly when those outputs appear professional and authoritative. In clinical reporting, this can lead to reviewers accepting AI generated narratives without sufficient scrutiny, particularly under time pressure. This is a human factors risk, not a model defect, and it must be addressed explicitly in risk assessment and control design. FDA guidance on human factors and computerized systems consistently emphasizes that systems should be designed to support appropriate user behavior and reduce the risk of misuse. For AI assisted reporting, this means clearly labeling AI generated content, enforcing mandatory review steps, and training reviewers to treat AI output as draft material, not authoritative interpretation.

6. Silent behavior change after vendor model updates

One of the most critical and least intuitive risks associated with LLMs is silent behavior change. Model providers routinely update models to improve performance, safety, or efficiency. These updates can change LLM output behavior even when the system configuration appears unchanged. In a GxP environment, this creates a unique risk. The system may no longer behave as it did at the time of validation, even though no internal change was made by the sponsor. Regulatory expectations for computerized systems require that validated systems remain in a controlled state and that changes are assessed for impact. For AI systems, this requires explicit drift detection mechanisms and vendor change surveillance as part of the risk mitigation strategy.

7. Plausible error is the dominant risk

The most dangerous failure mode for LLMs in clinical trial analysis and reporting is not obvious error. It is plausible error that escapes detection. Obvious errors are often caught quickly. Plausible errors are not. Plausible errors align with domain expectations, use correct terminology, and fit naturally into the narrative flow of the CSR. They are difficult to identify through casual review and may only surface during regulatory questioning, when the cost of remediation is highest. This is why risk assessment for LLMs must prioritize detectability over correctness. The system must be designed to surface uncertainty, enforce traceability, and block ungrounded content before it reaches regulated artifacts. From a validation perspective, this insight drives everything that follows. Test scenarios must include plausible but incorrect cases. Controls must be designed to detect subtle inconsistencies. Monitoring must focus on behavior change, not just error rates. This approach aligns with regulatory emphasis on data integrity, lifecycle control, and inspection readiness rather than theoretical model performance.

CORE VALIDATION STRATEGY FOR AI IN CLINICAL TRIAL ANALYSIS AND REPORTING

As stated previously in this article, a defensible validation strategy for AI in clinical trial analysis and reporting does not attempt to prove that a language model is correct, intelligent, or reliable in the abstract. That framing is misaligned with GxP expectations. Validation in this context is about demonstrating that a regulated process remains controlled when AI is introduced, that risks are mitigated, and that objective evidence exists to support continued use over time. The core strategy therefore shifts the validation target away from the model itself and toward the workflow, controls, and evidence framework that surround it.

1. Validate the workflow, guardrails, and evidence capture. Not the model itself.

Traditional computerized system validation assumes deterministic behavior, where identical inputs yield identical outputs. Large language models do not meet this criterion by design, as their outputs are probabilistic, context sensitive, and can vary even under stable configurations. Attempting to validate an LLM as a deterministic algorithm therefore leads to over validation, false assurance, or both. Regulatory guidance does not require validation of internal algorithms, but rather validation of the computerized system as used for its intended purpose. This principle is well established in FDA and EMA guidance, which emphasize fitness for intended use, data integrity, and controlled operation over exhaustive proof of internal mechanisms. For AI assisted analysis and reporting, this requires validation claims to be explicitly scoped to the end to end workflow. The questions validation must answer are practical and inspectable:

- Does the system constrain AI behavior to its approved intended use?
- Do guardrails prevent or detect ungrounded or out of scope output?
- Is there objective evidence showing how AI outputs were generated and reviewed?
- Does the system reliably surface errors and uncertainty rather than masking them?

If these questions can be answered with documented evidence, the system can be defended regardless of the internal complexity.

2. Treat the LLM as a configurable component inside a validated system

A critical architectural decision is to treat the LLM as a component, not the system. In regulated environments, components change. Systems remain validated through change control. This distinction is essential for maintaining a validated state while allowing AI technology to evolve. In this model, the validated system consists of:

- Defined workflows for authoring, review, and quality checks
- Guardrails that constrain inputs, outputs, and actions
- An evidence capture layer that records prompts, context, outputs, and adjudications
- Change control processes governing configuration and dependencies

The LLM itself is treated as a configurable dependency whose behavior is bounded by the system. Model version, provider, parameters, and enabled tools are configuration items subject to impact assessment and regression testing. This aligns with regulatory expectations that changes to computerized systems are assessed for impact on the validated state and documented appropriately. This approach also addresses one of the most challenging realities of AI adoption: vendor driven change. Sponsors do not control when commercial model providers update models. By treating the model as a replaceable component and validating behavior through system level tests, organizations can respond to vendor changes without re-validating the entire platform. The FDA's Good Machine Learning Practice principles emphasize lifecycle management, transparency, and monitoring rather than static one-time validation. While developed in the context of medical devices, the underlying principles translate directly to AI enabled components in regulated clinical workflows. The system must be designed so that change is expected and managed.

3. Build a repeatable, inspectable test harness

The centerpiece of a credible AI validation strategy is a repeatable, inspectable test harness. This is the mechanism by which abstract claims about control and risk mitigation are turned into objective evidence. A test harness in this context is not a one-time acceptance test. It is a controlled execution framework that repeatedly exercises the AI enabled workflow under predefined conditions and captures results in a form that QA and health authority auditors can review. At a minimum, a compliant test harness should:

- Execute approved scenarios using fixed inputs representative of real analysis and reporting tasks

- Capture the full execution context including prompts, system instructions, retrieved content, and model configuration
- Record outputs in structured form suitable for comparison and adjudication
- Support human review with defined pass and fail criteria and documented rationale

The scenarios themselves must be designed to reflect both normal and adverse conditions. Normal scenarios demonstrate that the system performs its intended functions. Adverse scenarios test failure modes such as ambiguous inputs, conflicting outputs, missing data, or attempts to elicit out of scope behavior. This approach aligns with long standing validation principles for computerized systems, where testing is designed to demonstrate that the system performs as intended and handles error conditions appropriately. The difference is that for LLMs, the focus is on behavioral acceptability rather than exact output matching.

4. roving acceptable behavior under normal and adverse conditions

Acceptable behavior must be defined before it can be tested. In clinical trial analysis and reporting, acceptable behavior typically includes:

- Only producing content that can be traced to approved sources
- Flagging uncertainty or inconsistency rather than resolving it implicitly
- Refusing to generate content when inputs are insufficient or out of scope
- Preserving consistency in terminology and population definitions

The test harness must demonstrate that these behaviors hold in ideal cases and when the system is stressed. For example:

- When a table is missing, does the AI refuse to summarize results
- When population definitions conflict, does it flag the issue
- When prompted to interpret results beyond the data, does it decline

Documented outcomes from these tests form the core validation evidence. They show not that the AI is perfect, but that the system reliably enforces boundaries and surfaces risk. When these artifacts are available, discussions about AI move out of the realm of novelty and into the realm of controlled systems. The system can be inspected using the same logic applied to other computerized systems used in clinical trials. This is the practical value of the core validation strategy described here. It allows organizations to adopt AI where it adds value, without redefining regulatory expectations or changing quality standards.

AI ASSISTED GENERATION, REVIEW & VALIDATION OF ANALYSIS OUTPUTS AND CLINICAL STUDY REPORT CONTENT WITH FULL TRACEABILITY TO CDISC SDTM AND ADAM AND TFL SPECS

This use case is the anchor because it sits directly on the regulated submission path while remaining technically and procedurally controllable. It addresses a persistent industry pain point: transforming structured analysis outputs into consistent, traceable, regulator ready narrative content. Unlike exploratory analyses or operational use cases, this workflow already operates under strict standards and review expectations. That makes it uniquely suitable for fit for purpose AI validation.

Scope within Clinical Trial Data Analysis and Reporting

The scope of this use case is intentionally narrow and explicit. It targets activities where AI can add material value while remaining bounded by existing regulatory controls. Within Clinical Trial Data Analysis and Reporting, the system supports:

1. Drafting and reviewing CSR sections related to efficacy and safety

AI assists in drafting descriptive and interpretive text for efficacy and safety sections of the clinical study report, using approved TFL outputs and analysis specifications as inputs. The AI does not determine statistical significance or clinical relevance. It drafts narrative content that must be reviewed and approved by qualified humans.

2. Generating narrative summaries aligned to TFL outputs

AI generates structured summaries that correspond directly to specific tables, figures, or listings. Each summary is explicitly tied to one or more TFLs, ensuring that narrative content cannot exist independently of approved outputs.

3. Validating consistency between CDISC SDTM and ADaM metadata, TFL shells, and report text

AI is used to compare population definitions, endpoint descriptions, parameter names, and analysis descriptions across metadata, outputs, and report text. Inconsistencies are flagged, not resolved autonomously.

4. Flagging discrepancies for human review

The system identifies mismatches, omissions, and potential inconsistencies and presents them to reviewers with traceable evidence. Resolution authority always remains with humans.

This scope aligns directly with how submissions are assembled and reviewed in practice. It also aligns with regulatory expectations that sponsors maintain consistency and traceability across submission artifacts.

Validation objectives

Validation objectives for this use case are behavioral and define what the system must consistently enforce to remain in a validated state.

1. No numerical or categorical claim appears without a traceable source
Any number, percentage, subject count, or categorical statement must be traceable to a specific TFL or analysis reference. If traceability cannot be established automatically, the content is blocked or flagged.

2. All summaries reference specific tables or listings
Narrative summaries must include explicit references to the TFLs they describe. Free floating narrative is not permitted within the regulated workflow.

3. Population language matches CDISC SDTM and ADaM metadata
Population definitions used in narratives must match approved metadata terms and definitions. Synonyms and paraphrasing are not allowed unless explicitly mapped and approved.

4. Uncertainty or ambiguity triggers escalation, not inference
When inputs are incomplete, conflicting, or ambiguous, the system must escalate the issue for human review rather than inferring or inventing content.

These objectives are directly testable and align with core data integrity principles emphasized by both FDA and international guidance.

Test harness design for this use case

This approach applies established validation principles for computerized systems, adapted to probabilistic AI.

1. Seeded scenarios using real world CDISC SDTM and ADaM and TFL patterns
Seeded scenarios are pre-defined, approved test cases that reflect realistic analysis and reporting situations. They are derived from real world CDISC SDTM and ADaM datasets, common TFLs, and typical CSR narrative patterns. Each scenario includes known expected behaviors, such as required references, acceptable phrasing boundaries, and escalation conditions. Seeded scenarios should be version controlled, approved as validation assets, and serve as a stable baseline against which behavior is assessed over time.

2. Adversarial cases including conflicting inputs and incomplete tables
In addition to nominal scenarios, the harness includes adversarial cases designed to trigger known failure modes. Examples include mismatched population definitions, missing tables, conflicting endpoint descriptions, and prompts that attempt to elicit unsupported interpretation. These scenarios test detectability and containment, not just functionality.

3. Capture of prompts, inputs, outputs, adjudication, and model configuration as validation evidence
Every test execution captures the full execution context: prompts, system instructions, retrieved metadata, outputs, human adjudication decisions, and model configuration details. This evidence is retained as part of the validation record and supports inspection readiness.

Guardrails specific to analysis and reporting

Guardrails are the controls that make the intended use enforceable. For this use case, they are designed around data integrity, traceability, and accountability.

1. Read only access to datasets
AI components have read only access to approved datasets and outputs. They cannot modify CDISC SDTM or ADaM datasets, TFLs, or metadata under any circumstances.

2. Citation enforcement for factual claims
The system enforces citation requirements for all factual claims. Unsupported statements are blocked or flagged automatically.

3. Output schemas that block unsupported assertions
Generated outputs must conform to predefined schemas that restrict content types, enforce references, and prevent unsupported interpretation. This reduces reliance on reviewer vigilance alone.

4. Mandatory human review before inclusion in submission artifacts
No AI generated content can enter regulated submission artifacts without documented human review and approval. This control is enforced by workflow, not policy alone.

Together, these guardrails transform AI from an unbounded text generator into a controlled subsystem within a validated analysis and reporting platform.

VALIDATION TEST HARNESS ARCHITECTURE

With a harness, AI behavior becomes measurable, repeatable, and governable. This section applies across all AI use cases described in this paper. Whether AI is used for CSR drafting, analysis specification authoring, or cross artifact consistency checking, the harness architecture is what keeps the system in a validated state over time.

1. Controlled execution runner

The controlled execution runner is the environment in which all validation and regression testing occurs. Its role is to eliminate uncontrolled variability and ensure that each test execution is reproducible and attributable. Key characteristics include:

- Fixed and versioned execution context, including model identifier, configuration parameters, enabled tools, and system prompts
- Explicit control over inputs, including which CDISC SDTM & ADaM datasets and TFLs are accessible
- Isolation from production usage to prevent contamination of validation evidence
- Automatic capture of timestamps, user identity or service account, and execution identifiers

From a GxP perspective, the execution runner is analogous to a validated test environment for a computerized system. Regulatory guidance does not require that production and validation environments be identical, but it does require that validation testing be performed in a controlled and documented manner representative of intended use. The runner provides that control. Critically, the runner enforces the intended use boundary. If the AI attempts to access unauthorized data, generate unsupported content types, or bypass workflow constraints, the execution fails visibly and is logged as such.

2. Scenario library with versioning and approval

The scenario library defines what the system is expected to handle and is a validated asset. Each scenario represents a realistic and relevant use case derived from actual clinical trial analysis and reporting patterns. Scenarios include:

- Approved inputs such as CDISC SDTM and ADaM datasets, TFLs & their shells, and CSR excerpts
- Defined intent, such as drafting a summary for a specific table or checking consistency across artifacts
- Expected behavioral outcomes, including required traceability, escalation conditions, and refusal cases

All scenarios are version controlled and approved. Changes to scenarios follow change control, just as changes to test scripts do in traditional computerized system validation. This aligns with regulatory expectations that validation artifacts themselves are controlled and auditable. The library must also include negative scenarios. These are intentionally designed to provoke failure modes identified in the risk assessment, such as conflicting population definitions or missing outputs. Their purpose is not to demonstrate success, but to demonstrate detectability and containment.

3. Objective adjudication rubric with severity classification

AI validation fails when outcomes are judged subjectively. A test harness requires an adjudication rubric that defines what constitutes acceptable and unacceptable behavior in objective terms. An effective rubric includes:

- Clear pass and fail criteria tied to validation objectives
- Severity classification for failures, such as critical, major, or minor
- Explicit mapping between failure types and risk categories
- Required documentation for adjudication decisions

For example, a hallucinated efficacy claim would typically be classified as a critical failure, while a stylistic inconsistency with preserved traceability might be minor. This mirrors established quality practices where deviations are assessed based on impact rather than mere occurrence. Objective adjudication is essential for inspection readiness. Health authority auditors expect to see not only that tests were run, but that results were evaluated consistently and that failures were understood and addressed. The rubric ensures that AI validation evidence is defensible and repeatable.

4. Regression testing across model and configuration changes

Regression testing is where the harness proves its value over time. Large language models are not static. Vendor updates, configuration changes, prompt adjustments, and retrieval corpus updates can alter system behavior. The harness supports regression testing by:

- Re executing approved scenarios on a defined cadence
- Triggering targeted regression after any model, prompt, or configuration change
- Comparing results against previously approved baselines
- Flagging behavioral changes that exceed predefined thresholds

This approach directly supports lifecycle control expectations for computerized systems. Regulators expect that validated systems remain under control as changes occur, and that the impact of change is assessed and documented. For AI systems, regression testing through a harness is the only practical way to meet that expectation.

COMPLIANCE CONTROLS QA EXPECTS AND HEALTH AUTHORITY AUDITORS RECOGNIZE

Quality Assurance and health authority auditors are not looking for novel AI governance concepts, instead they are looking for familiar controls applied consistently to a new tool. When AI is embedded in clinical trial analysis and reporting, the question is not whether the technology is new, but whether the controls surrounding it are recognizable, enforceable, and auditable under GxP. This section describes the baseline controls QA expects to see and health authority auditors recognize immediately. None of these controls are AI specific inventions; they are established GxP controls applied deliberately to AI enabled processes.

1. Role based access and segregation of duties

Role based access control is foundational to GxP compliance. AI does not change this requirement. In a compliant design, access to AI functionality is governed by defined roles that align with existing responsibilities in clinical trial analysis and reporting. Typical role distinctions include:

- Authors who may request AI generated draft content
- Reviewers who adjudicate AI outputs
- Approvers who authorize inclusion of content in regulated artifacts
- Administrators who manage configuration but do not author or approve content

Segregation of duties should be enforced technically and not assumed procedurally. A system that allows the same individual to configure prompts, generate content, adjudicate outputs, and approve final artifacts collapses control

boundaries and will not withstand inspection. Regulatory guidance consistently emphasizes the need for appropriate access controls and segregation of duties for computerized systems used in clinical trials. The EMA guideline on computerized systems and electronic data explicitly calls out access management and prevention of unauthorized activities as core expectations. These expectations apply equally when AI is involved.

2. Complete audit trails for prompts, outputs, and adjudications

Audit trails and traceability are non-negotiable. For AI enabled systems, the scope of the audit trail must expand beyond traditional data changes to include AI interactions. At a minimum, a compliant audit trail captures:

- The prompt or request submitted to the AI
- The input data and retrieved context used to generate the output
- The AI generated output itself
- The identity of the user initiating the request
- Timestamps for each step

This level of detail is essential because AI outputs are not deterministic. Without capturing prompts and context, it is impossible to reconstruct how a given output was produced. From an inspection perspective, that absence is equivalent to missing lineage and traceability. FDA and EMA guidance on computerized systems emphasize that audit trails must allow reconstruction of events and support traceability of regulated activities. For AI systems, prompts and retrieved context are part of the event. Excluding them creates a gap in the audit trail that health authority auditors can challenge.

3. Record retention aligned to the quality system

AI generated content and its associated audit trail are GxP records when used to support regulated activities. As such, they are subject to the same record retention requirements as other clinical trial documentation. This expectation is consistent with long-standing regulatory requirements for electronic records in clinical trials. Health authority auditors will not accept arguments that AI outputs are transient or exploratory if they influence regulated deliverables. A compliant retention strategy ensures that:

- AI interaction records are retained for the same duration as the artifacts they support
- Records are protected from unauthorized modification or deletion
- Retrieval is possible for audits, inspections, and investigations
- Retention policies are documented and aligned with the organization's quality system

CONTINUOUS VERIFICATION AND DRIFT DETECTION

Continuous verification is not an enhancement to AI validation. It is a requirement. Large language models are inherently dynamic systems whose behavior can change over time even when no internal code change has been made by the sponsor. In regulated clinical trial analysis and reporting, the validated state must be maintained throughout system use, not assumed based on initial testing. Continuous verification is the mechanism by which that requirement is met. The objective is straightforward. The organization must be able to demonstrate that the AI enabled system continues to behave acceptably for its intended use, that deviations are detected promptly, and that appropriate actions are taken to restore control before regulated artifacts are impacted.

1. Routine execution of scenario suites

The foundation of continuous verification is routine execution of approved scenario suites through the validation test harness. These scenarios are not ad hoc checks. They are controlled, versioned assets designed to reflect real analysis and reporting tasks as well as known failure modes. On a defined cadence, such as monthly or quarterly depending on risk, the system executes the full scenario suite using the controlled execution runner. Results are compared against previously approved baselines, focusing on behavioral characteristics rather than exact wording. Routine execution demonstrates that:

- Guardrails remain effective
- Traceability requirements are consistently enforced
- Refusal and escalation behavior has not degraded
- The system still operates within its validated intended use

From a regulatory perspective, this is lifecycle validation in practice. The EMA guideline on computerized systems emphasizes that validated systems must remain in a controlled state and that ongoing verification is necessary when systems are used over time.

2. Detection of behavioral shifts after vendor updates

One of the most critical drift vectors for LLM based systems is vendor driven change. Commercial model providers routinely update models, safety policies, and infrastructure. These updates can alter output behavior even when system configuration remains unchanged. A compliant AI governance model therefore includes explicit surveillance of vendor updates and their potential impact. This includes:

- Monitoring vendor release notes and change notifications
- Identifying whether updates affect models used in regulated workflows
- Triggering targeted regression testing using the validation harness
- Documenting impact assessments and outcomes

Behavioral shift detection is based on evidence, not vendor assurances. Even when updates are described as nonfunctional, the system must demonstrate continued acceptable behavior through testing. Regulatory expectations for computerized systems are clear on this point. Changes that may affect system performance or data integrity must be assessed and controlled. This is no different for AI systems.

3. Regulatory alignment with lifecycle monitoring expectations

Lifecycle monitoring is not an emerging concept, rather it is an explicit regulatory expectation for computerized systems. As stated previously the FDA's guiding principles emphasize ongoing monitoring, transparency, and the ability to manage change throughout the system lifecycle. The core message is consistent across guidance: validation is not a one-time event. It is a continuous and ongoing obligation. For AI in clinical trial analysis and reporting, continuous verification and drift detection are the mechanisms that fulfill that obligation in a practical, inspectable way.

TECHNICAL IMPLEMENTATION EXAMPLE: CONSISTENCY CHECKING AND DISCREPANCY DETECTION ACROSS SUBMISSION ARTIFACTS

This section describes, in concrete technical and operational terms, how an AI capability can be implemented, validated, and operated within a regulated Clinical Trial Data Analysis and Reporting environment. The implementation focuses on consistency checking and discrepancy detection across CDISC SDTM/ADaM metadata and data, tables/figures/listings, and Clinical Study Report content. This use case is intentionally selected because it operates directly on regulated submission artifacts while remaining bounded, observable, and defensible under inspection. It introduces AI into a critical but well understood part of the submission workflow without transferring scientific judgment, statistical authority, or approval responsibility to the system. Unlike generative authoring use cases, this implementation does not rely on trust in AI interpretation. It relies on constraint, normalization, traceability, deterministic validation, and escalation. From a regulatory perspective, this represents the lowest ambiguity entry point for AI in submissions. From a business perspective, it targets a persistent bottleneck: manual, error prone cross artifact review required to ensure internal consistency across increasingly complex submission packages.

Intended technical behavior and operational scope

The AI functions strictly as a comparison and verification mechanism. Its role is to evaluate alignment across structured and semi structured artifacts and surface inconsistencies that require human resolution. It is explicitly designed to assist reviewers, not to replace them. Permitted behavior is limited to comparison and detection. The system may compare population usage across CDISC SDTM/ADaM, TFLs, and CSR sections. It may verify endpoint names, parameter labels, and analysis method descriptions for consistency. It may detect mismatches in table numbering, listing references, and internal citations. It may flag discrepancies with explicit, traceable evidence for human review. The system is explicitly prohibited from drafting CSR narrative content, interpreting clinical or statistical significance, resolving discrepancies autonomously, modifying datasets or metadata, or approving or releasing submission content. These prohibitions are enforced technically and procedurally. This boundary ensures the AI cannot change the scientific or regulatory position of a submission. From a GxP standpoint, it preserves the accountability model regulators expect: humans decide and approve.

Controlled inputs, retrieval architecture, and access segregation

The AI has no ambient knowledge and no autonomous access to data. All inputs are provided through a gated retrieval subsystem operating within the validated environment. Retrieval sources allowed are listed, versioned, and approved. Each retrieved content element (e.g., CDISC dataset, a TFL, CSR section) unit carries mandatory metadata, including artifact identifier, artifact version, section or row identifier, and a content hash. Retrieval configuration itself is treated as a controlled item. Changes to corpus content, chunking strategy, ranking logic, or retrieval parameters require impact assessment and targeted regression testing. Access to retrieval configuration, population registries, and validation scenarios is segregated by role. Authors, reviewers, QA, and system administrators have distinct permissions. This segregation of duties is enforced technically and aligns with expectations for computerized systems supporting regulated processes.

Population normalization and canonical mapping layer

Before any AI based comparison occurs, population references across artifacts are normalized into a canonical, controlled representation. This layer is mandatory. It exists to ensure that population comparison is deterministic, explicit, and auditable, and that the AI is never required to infer population meaning from prose or labels. In clinical submissions, population definitions appear differently depending on the artifact. ADaM metadata operationalizes analysis populations explicitly through subject level flags and derivation logic. SDTM metadata provides supporting provenance, subject attributes, and trial design context. TFLs reference populations using human readable labels or footnotes. CSR sections reference populations in narrative prose, often using synonyms or abbreviated terminology. These representations are not directly comparable without normalization. This normalization layer deliberately separates data governance from AI behavior. Population meaning is defined once, approved once, and reused everywhere. At the core of this layer is a Population Definition Registry managed under the quality system. Each population entry includes a unique identifier, canonical name, authoritative SDTM/ADaM dataset and flag variable, derivation logic, human readable definition, approved synonyms, and version and approval status. ADaM ADSL flags are the authoritative operational definitions for analysis populations. Population references are extracted deterministically prior to any AI involvement. ADaM flags are mapped directly to canonical population identifiers.

SDTM attributes are extracted to support cross checks. TFL population labels are mapped using approved synonyms. CSR population mentions are extracted only from predefined sections and normalized using the registry. Free interpretation is not permitted. All extracted references are reduced to a normalized population mapping table that becomes the sole input for comparison.

LLM configuration and execution constraints

The AI operates under a dedicated, locked configuration profile optimized for regulated consistency checking and discrepancy detection. This profile is not advisory. It is enforced technically and governed procedurally. All configuration parameters, prompts, and instruction templates are treated as controlled configuration items subject to change control, impact assessment, and regression testing. The purpose of this configuration is to minimize behavioral variability, eliminate creative generation, and ensure outputs are predictable, inspectable, and suitable for validation in a GxP context.

Explicit enforced LLM control settings

The following controls are enforced for this use case and captured as part of the execution record for every run:

Control Category	Details	Description
Sampling and determinism controls	• temperature = 0 • top_p = 1.0	These settings suppress creative variation and constrain the model to the most likely completion paths. While probabilistic behavior cannot be eliminated entirely, these controls materially reduce variability and are necessary for defensible validation.
Reproducibility and drift visibility	• seed = fixed per validation scenario backend or system fingerprint captured when provided by the platform	Seed-based determinism is treated as best effort, not a guarantee. Backend or system fingerprints are captured to detect vendor-driven model or infrastructure changes even when no local configuration has changed. These fingerprints are explicitly monitored as part of drift detection.
Output length and termination	• max_tokens = 400 to 900, sized to expected findings volume • explicit stop sentinel enforced	Token limits are sized to the expected volume of structured findings and prevent uncontrolled or truncated output. Explicit stop sentinels are enforced to terminate responses deterministically and prevent run-on generation.
Penalty configuration	• frequency_penalty = 0 • presence_penalty = 0	Novelty is explicitly undesirable in regulated review tasks. These penalties remain disabled to avoid incentivizing variation or reformulation.
Output channel enforcement	• structured output only • no free-text narrative channel • strict JSON schema enforcement with post-generation rejection	The system does not permit narrative prose intended for direct consumption. Outputs that do not conform exactly to the approved schema are rejected automatically and do not enter the review workflow.

Configuration governance and change control

All parameters listed above are treated as configuration items, not tuning preferences. Any change to these settings requires documented impact assessment and targeted regression testing using the validated harness. Changes are reviewed and approved through the quality system before deployment. This explicit control inventory ensures that AI behavior is not described abstractly, but defined concretely, enforced consistently, and inspectable at any time.

AI comparison and structured output

The AI compares normalized population objects and other structured metadata, not raw text. It detects mismatches, omissions, multiple population usage, and unapproved terminology. It emits findings only, never resolutions. Each finding includes severity classification, discrepancy category, precise description, evidence links with artifact identifiers, versions, locations, and content hashes, an impact statement framed in regulatory terms, and required human action. Evidence on both sides of any discrepancy is mandatory. Severity classification is tied to impact, not frequency. Critical findings include incorrect population mapping, missing evidence, or failure to detect known discrepancies in validated scenarios. Major findings include incorrect categorization or incomplete impact descriptions. Minor findings include wording issues that do not affect interpretation. Acceptance thresholds are defined per severity class and enforced by the harness.

Deterministic validation, failure handling, and escalation

Before any human reviewer sees an AI output, a non-AI deterministic validator executes. It confirms artifact existence and approved versions, validates identifiers, enforces exact population vocabulary matching, verifies numeric equality where applicable, and enforces schema compliance. If validation fails, the output is rejected, logged as a control failure, and blocked from the workflow. The user receives a controlled error indicating that the system could not produce a valid result. No output enters review by exception. When validation or regression failures occur, escalation is procedural. Depending on severity, this may trigger feature restriction, suspension of AI functionality, root cause analysis, corrective action, and partial or full revalidation. These responses mirror existing deviation and change management processes in GxP environments.

Test harness, regression testing, and lifecycle management

Validation is anchored to the harness, not the model, which is treated as an interchangeable component. The harness exercises the full workflow end to end against approved scenarios with explicit acceptance criteria, including baseline behavior, conflicting populations, missing references, prompt manipulation, terminology drift, and vendor change detection. Regression testing runs both on a defined cadence and in response to change events. Any modification to models, prompts, schemas, retrieval sources, population definitions, or validation logic triggers impact

analysis and focused retesting. Evidence produced by the harness is maintained independently from routine operational records to preserve inspection integrity.

Continuous monitoring, drift detection, and record retention

Operational executions generate quality signals that are trended over time, including discrepancy categories, severity distributions, escalation frequency, and validator rejection rates. Vendor driven model updates are treated as expected drift risks. Backend fingerprint changes trigger targeted regression and review. Every execution produces a complete, reconstructable audit trail capturing prompts, retrieved artifacts, model outputs, validator results, human adjudications, and configuration snapshots. Records are retained in alignment with the lifecycle of the submission artifacts they support, consistent with computerized system expectations in clinical trials.

Why this implementation is defensible

This system does not depend on AI being correct. It depends on the process being controlled. Behavior is bounded, evidence is captured, failures are detectable, and change is governed. The harness, not the model, is what is validated. That is why this implementation is inspectable, maintainable, and aligned with regulatory expectations.

AI Consistency Checking Control Matrix for Clinical Trial Data Analysis and Reporting

Control Category	Specific Control	Risk Mitigated	How the Control Works	Validation Evidence	Ongoing Monitoring
Sampling determinism	temperature = 0	Non reproducible narrative variation	Forces the model to select the most probable completion path rather than creative alternatives	Harness execution logs showing stable output classes across repeated runs	Drift analysis comparing defect category distribution over time
Sampling determinism	top_p = 1.0	Unintended output variability	Prevents probabilistic truncation of candidate tokens	Regression results demonstrating consistent findings across runs	Alert if defect frequency or severity shifts
Reproducibility	seed fixed per validation scenario	Inability to reproduce validation outcomes	Anchors randomness for test scenarios to support comparison	Harness records showing seed value per scenario	Seed changes flagged as configuration deviations
Drift visibility	backend or system fingerprint capture	Silent vendor model changes	Captures provider-supplied identifiers when models or infrastructure change	Execution metadata stored with each run	Automated detection of fingerprint changes triggering regression
Output bounds	max_tokens = 400–900	Truncated or runaway output	Constrains response length to expected findings volume	Harness results confirming complete findings within limits	Token overflow or truncation alerts
Deterministic termination	explicit stop sentinel	Partial or uncontrolled generation	Forces clean termination of output	Output validation logs showing sentinel enforcement	Rejection count of malformed outputs
Novelty suppression	frequency_penalty = 0	Creative rewording of findings	Prevents incentivization of new phrasing	Stable discrepancy phrasing across regression runs	Textual drift trend analysis
Novelty suppression	presence_penalty = 0	Inconsistent restatement of identical issues	Avoids encouraging semantic variation	Repeatability of finding descriptions	Monitoring for semantic drift
Output channel restriction	Structured output only	Plausible but unsupported prose	Disallows narrative text entirely	Schema validation logs	Zero tolerance enforcement rate
Output schema enforcement	Strict JSON schema	Missing or ambiguous findings	Blocks outputs without required fields	Validator rejection records	Schema failure rate trends
Evidence enforcement	Evidence required on both sides	Unsubstantiated discrepancies	Requires dual artifact references per finding	Validation logs showing evidence completeness	Escalation of evidence gaps
Population governance	Canonical Population Registry	Semantic population mismatch	Defines populations once and reuses everywhere	Approved registry versions	Registry change audit trail
Population authority	ADaM ADSL flags authoritative	Ambiguous population derivation	Anchors analysis populations to operational logic	Registry mapping documentation	Registry integrity checks
Population normalization	Deterministic extraction rules	AI inference of population meaning	Extracts populations before AI involvement	Extraction test cases	Extraction failure alerts
Retrieval isolation	Allow-listed corpus only	Prompt injection or data leakage	Prevents access outside approved artifacts	Retrieval config approval records	Unauthorized access attempts
Retrieval traceability	Artifact ID, version, hash	Post hoc traceability loss	Guarantees exact source reconstruction	Stored retrieval metadata	Hash mismatch alerts
Deterministic validation	Pre-review validator	Human reliance to catch errors	Blocks invalid outputs before review	Validator execution logs	Validator rejection rate
Failure handling	Automatic rejection on failure	Silent acceptance of bad output	Stops workflow on control failure	Error handling records	Failure trend monitoring
Regression testing	Scenario-based harness	Undetected behavioral drift	Replays known edge cases	Harness execution reports	Scheduled and change-triggered runs
Drift detection	Severity and category trending	Gradual degradation	Detects non obvious shifts	Trend reports	Escalation thresholds
Segregation of duties	Role-based access control	Unauthorized changes	Separates authoring, review, QA	Access logs	Access violation alerts
Audit trail	Full execution reconstruction	Inability to explain decisions	Captures prompts, inputs, outputs, adjudication	Audit trail records	Retention compliance checks
Record retention	Submission-aligned retention	Evidence loss	Retains records as long as artifacts	Retention policies	Periodic retention audits

CONCLUSION

Clinical trial analysis and reporting is the most practical entry point for controlled AI adoption in regulated environments. The work is inherently structured, SOP governed, and subject to intensive human review, with clear accountability for interpretation and approval. AI can therefore deliver value without assuming decision authority. Although errors are costly, they are explicit and traceable, which makes issues discoverable and addressable. Value emerges only when AI is confined to behaviors that are testable, monitored, and auditable. Drafting, summarization, comparison, and anomaly flagging can be engineered so every output is anchored to a specific table, listing, or analysis. Content that cannot be traced is blocked, escalated, or flagged. Approval of statistical results, dataset modifications, and interpretive conclusions remains exclusively human. This boundary defines validation scope, risk classification, and inspection exposure. Ambiguity at this boundary increases audit risk immediately. The most significant risks are not obvious failures but plausible ones that evade detection. Confident but incorrect narratives, fabricated references that appear valid, inconsistent population language, reviewer automation bias, and silent behavioral change after vendor updates pose greater risk than overt errors. Drift is an expected property of

probabilistic systems. Treating drift as a lifecycle condition reduces the likelihood of undetected degradation and inspection findings.

A defensible implementation anchors AI behavior within a controlled workflow rather than trusting the model itself. Canonical definitions are established for shared concepts, particularly analysis populations, and approved once for reuse across all artifacts. Population references are extracted deterministically and normalized before any AI involvement. The AI compares normalized objects and flags discrepancies rather than resolving them. Behavior is constrained through explicit configuration and execution controls. Variability is reduced through fixed parameters, fixed seeds, and constrained decoding. Outputs are limited to structured findings. Retrieval sources are versioned and traceable. Deterministic validators block invalid outputs prior to human review. Prompts, schemas, retrieval corpora, and configuration parameters are managed as configuration items under change control. A repeatable test harness executes approved scenarios, adversarial cases, and drift sentinels to generate objective evidence over time. Continuous monitoring trends defect type and severity so gradual degradation is detected before it affects submissions. What ultimately matters is not the model. The validated asset is the harness and the workflow that surrounds it. Models are replaceable components. Controls, evidence capture, change management, and lifecycle monitoring are what make the system defensible. This is aligned to regulatory expectations that emphasize ongoing verification and governance for AI enabled functionality rather than single event validation.

The path forward is practical and incremental. Bound a single assistant behavior in analysis and reporting. Define intended use precisely and enforce it technically. Build the harness early and let it generate evidence continuously. Treat prompts and retrieval sources as configuration, not convenience. Assume drift will occur and design for detection and response from day one. When AI is embedded inside a controlled process with clear boundaries and objective evidence, it behaves like any other validated computerized system while delivering measurable operational value without expanding your inspection risk.

REFERENCES

1. Clinical Data Interchange Standards Consortium. Analysis Data Model (ADaM) Implementation Guide. Accessed January 2026. https://www.cdisc.org/standards/foundational/adam
2. Clinical Data Interchange Standards Consortium. Study Data Tabulation Model (SDTM) Implementation Guide. Accessed January 2026. https://www.cdisc.org/standards/foundational/sdtm
3. European Medicines Agency. Guideline on Computerised Systems and Electronic Data in Clinical Trials (EMA/INS/GCP/112288/2023). March 9, 2023. https://www.ema.europa.eu/en/documents/regulatory-procedural-guideline/guideline-computerised-systems-and-electronic-data-clinical-trials_en.pdf
4. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Official Journal of the European Union, July 12, 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
5. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). E3: Structure and Content of Clinical Study Reports. November 1995. https://database.ich.org/sites/default/files/E3_Guideline.pdf
6. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). E9: Statistical Principles for Clinical Trials. September 1998. https://database.ich.org/sites/default/files/E9_Guideline.pdf
7. U.S. Food and Drug Administration. Good Machine Learning Practice for Medical Device Development: Guiding Principles. Updated December 19, 2025. https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles
8. U.S. Food and Drug Administration. Study Data Technical Conformance Guide: Technical Specifications Document. Updated December 16, 2025. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/study-data-technical-conformance-guide-technical-specifications-document
9. U.S. Food and Drug Administration. Study Data Technical Conformance Guide. Updated December 2025. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/study-data-technical-conformance-guide
10. U.S. Food and Drug Administration. Applying Human Factors and Usability Engineering to Medical Devices. February 2016. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/applying-human-factors-and-usability-engineering-medical-devices
11. U.S. Food and Drug Administration. Data Integrity and Compliance With Drug CGMP: Guidance for Industry. December 2018. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-integrity-and-compliance-drug-cgmp-guidance-industry

CONTACT INFORMATION

Contact the authors at:

Author Name: Bill Qubeck
Company: Sycamore Informatics
Email: Bill.Qubeck@SycamoreInformatics.com
Website: https://www.sycamoreinformatics.com/

Author Name: Heather Gorr
Company: Sycamore Informatics
Email: Heather.Gorr@SycamoreInformatics.com
Website: https://www.sycamoreinformatics.com/