Paper PP33

# Reproducibility, Lineage, and Control: Building a Traceability Framework for the Modern Clinical Trial

Bill Qubeck, Sycamore Informatics, United States
Shilpa Sood, Sycamore Informatics, India

## ABSTRACT

As clinical trials become more complex, global, and data-intensive, the margin for ambiguity in how data is collected, transformed, and analyzed has all but vanished. Regulatory expectations are rising—from ICH E6(R3) to 21 CFR Part 11 and EMA's emphasis on data integrity—and with them, the need for full traceability, version control, and reproducibility across the entire clinical data lifecycle has become mission-critical.

This poster presents a comprehensive, future-ready framework for establishing robust traceability from source data to submission deliverables. It details how to implement systematic lineage tracking across clinical systems—from EDCs and lab feeds through to SDTM transformations, ADaM datasets, TLF outputs, and final regulatory submission packages. Attendees will gain insight into practical, scalable techniques for maintaining audit logs, managing change control, and automating version management across both code and data assets.

By integrating metadata-driven processes, standardized workflow definitions, and lightweight automation layers, the framework supports transparency, inspection readiness, and operational scalability—regardless of platform or vendor landscape. The poster will showcase real-world use cases where gaps in lineage or version control led to downstream rework, regulatory queries, or audit findings—and how they were resolved through proactive design.

Whether you're preparing for your next inspection, trying to eliminate "black box" data steps, or building infrastructure to support real-time data review, this framework offers a blueprint for creating a verifiable, reproducible, and regulator-ready clinical data environment.

## INTRODUCTION

#### THE MANDATE FOR TRACEABILITY IN THE MODERN CLINICAL TRIAL

As clinical trials evolve into complex, global, and data-intensive endeavors, the industry faces unprecedented regulatory scrutiny regarding data integrity, reproducibility, and governance. Driven by updated guidelines such as ICH E6(R3), the requirements of 21 CFR Part 11, and the European Medicines Agency's (EMA) relentless focus on data quality, the ability to demonstrate a clear, verifiable, and reproducible path from raw source data to the final regulatory submission is no longer a best practice—it is a mission-critical necessity. This paper presents a comprehensive framework designed to embed full traceability, robust version control, and systematic lineage tracking across the entire clinical data lifecycle.

The primary focus of the clinical trial industry remains, and must always be, patient safety. This core tenet underpins the evolving industry focus, moving from initial efforts in regulatory compliance and process automation toward a modern era of process optimization. Achieving this optimization while maintaining the highest safety and quality standards requires a fundamental shift in how clinical data is managed and governed.

## NEWER & MORE STRINGENT REGULATORY REQUIREMENTS

Modern regulatory expectations demand far greater technological and process sophistication than legacy systems and manual processes can provide. The environment must be designed proactively to address a comprehensive set of requirements that are becoming non-negotiable standards.

Key necessities for compliance include:

- Better Data Lineage & End-to-End Traceability: A verifiable, end-to-end relationship must exist between the source data, all transformation programs, and the final output.
- Thorough User and Action-Specific Audit Trails: Chronological, computer-generated, and immutable records

of all events, actions, and changes are essential.
- Clarity in Access Restrictions: Granular access control, particularly with respect to sensitive unblinded data, must be consistently handled and fully auditable across the entire data lifecycle.
- Accurate Reproducibility: The ability to generate the exact same output at any point in time using the precise inputs and logic is the bedrock of scientific validity.
- Non-negotiable Version Control: Full version control is expected for all files, including programs, datasets, metadata, and outputs—versioning code alone is insufficient.
- Robust Integration Across Applications: Siloed systems inhibit compliance and efficiency. Seamless, integrated control across the statistical computing environment is required.
- Automated and Documented Change Tracking: The system must automatically manage change tracking, eliminating the high effort and risk associated with manual processes.
- Data Integrity & Security: Security controls and verifiable ingress and egress of data and files must be audited and supported.

## KEY ELEMENTS OF REPRODUCIBLE RESULTS

Reproducibility is the bedrock of scientific validity and regulatory acceptance. A result is deemed reproducible only if the exact same output can be generated at any point in time, using the same inputs and the same logic. This requires an integrated approach comprising four critical, non-negotiable elements: Version Control, Access Governance, Audit Trails, and Traceability & Lineage.

### VERSION CONTROL

**Version control** for all artifacts (e.g., data, programs, outputs) within the Clinical Study Report (CSR) programming process is the fundamental mechanism that enables reproducibility. It transforms programming actions into an **immutable, auditable history**.

- **Inadequacy of Code-Only Versioning:** While solutions like Git and SVN are optimized for text-based code files, they are typically **inefficient and unsuitable** for directly versioning large, binary, or structured clinical datasets (SDTM, ADaM). Critically, solely versioning the **code** used for transformation is **insufficient** to meet regulatory expectations. The specific version of the **data input** used for any analysis must also be managed and linked.
- **The Requirement:** A robust framework must provide automated, low-overhead version management for **both code and data assets** to guarantee that a final output can be re-created using the precise combination of code and data that produced the original result.

### ACCESS GOVERNANCE

A permanent, **immutable record** of all access permissions is an absolute must to demonstrate oversight and governance. This record ensures compliance and control over the clinical trial analysis process.

- **Auditable History:** The system must create an auditable history of policy changes, answering key regulatory questions:
    - **Who** had access, and **when**?
    - **Who** authorized the access?
    - **What** specific data or resources were they authorized to access?
- **Versioned Policy Management:** A suitable solution must manage the **versioning of the access control policy** itself and log the granular events of access and attempted access that surround the data. This is especially crucial for managing access to sensitive **unblinded data**.

### AUDIT TRAILS

An **audit trail** is a chronological, computer-generated record providing definitive proof of the sequence of events, operations, and changes executed by a user or a system process. For a truly regulated solution, audit trails must be:

- **Immutable (WORM):** The record must not be alterable or deletable once recorded (**Write Once, Read Many**), preventing tampering and upholding data integrity.
- **Time-Stamped:** Every entry must be recorded with a precise date and time using a consistent, tamper-proof system clock.
- **Attributable:** Every action must be uniquely tied back to the individual user or the specific system process that performed it.

#### TRACEABILITY & LINEAGE

Traceability and Lineage provide the **verifiable, end-to-end relationship** between the data, programs, and output. This crucial function connects every output directly back through:

1. The analysis program and its specific version.
2. The specific **ADaM dataset version** used.
3. The source **SDTM data version**.
4. Ultimately, the **raw patient data point** and the clinical **protocol** that governed its collection.

Lineage functions as the **immutable metadata linkage** across all steps, enabling the sponsor and regulatory reviewer to answer the critical question, **"How was this result produced?"** with a definitive and reproducible answer at any point in time.

**Crucially, version control is necessary but insufficient.** Sponsors require systems that **automatically manage the relationships** between programs, their inputs, and outputs, alongside managing permissions and audit trails. Leveraging this information accelerates delivery while ensuring regulatory compliance with minimal overhead.

## THE CLINICAL DATA FLOW: TRACEABILITY CHALLENGES

The clinical data lifecycle is a chain of crucial transformations, as illustrated in the data flow diagram below. With increased standardization (e.g., **CDISC SDTM** and **ADaM**), the emphasis on traceability between these stages is paramount.

The process begins with **Structured Protocol Design & Authoring**, which defines the **Protocol** and dictates the design of the **eCRF** (Electronic Case Report Form) for data collection and the **SAP** (Statistical Analysis Plan) for analysis. The SAP also guides the creation of **Mocks** (shells for Tables, Listings, and Figures).

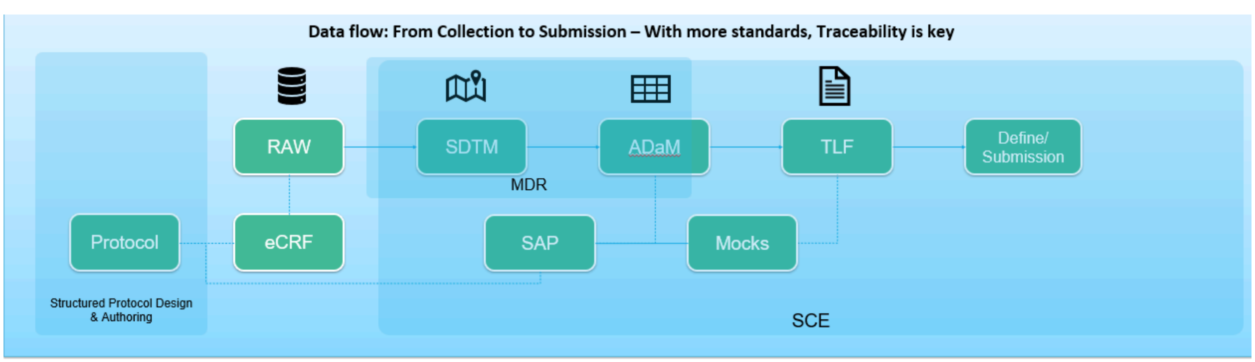
Data collection results in **RAW** data, which must then be transformed into standardized datasets:

- **SDTM (Study Data Tabulation Model):** Structured data for regulatory submission.
- **ADaM (Analysis Data Model):** Analysis-ready data derived from SDTM.

Both SDTM and ADaM, as well as the relationships between them, are governed by the **MDR (Metadata Repository)**. The ADaM data is then used to generate the **TLF (Tables, Listings, and Figures)** outputs.

Finally, the entire package, including the TLF, the standardized datasets, and supporting documentation like **Define.xml**, is prepared for **Define/Submission**. The analytical work from SAP through TLF creation occurs within a **SCE (Statistical Computing Environment)**.

Traceability must link the **RAW** data to the **SDTM** dataset version, the **ADaM** dataset version, the **TLF** output, and ultimately back to the foundational **Protocol** and **eCRF** from which the data originated. This end-to-end view, governed by robust version control and audit trails, is the essence of a modern, compliant traceability framework.



## ARCHITECTURAL BLUEPRINT: DEFINING THE MODERN SOLUTION

The shortcomings of siloed, manual, and code-only versioning approaches necessitate a shift towards a cohesive, centralized environment. A modern, future-ready solution for clinical trial data management must be architected around core principles that support both regulatory compliance and operational efficiency.

The ideal solution should embody the following eight characteristics: **Teamwork, Reproducibility, Elastic Resources, Agility, Governance, Insights, and Automation.**

#### CORE SYSTEM REQUIREMENTS

A modern, compliant platform must be a **single, comprehensive application** that serves as the **Statistical Computing Environment (SCE)**. Its core capabilities must be built-in, not bolted-on:

- **Integrated Controls:** Must feature built-in **version control, access control, compute management, traceability, reproducibility,** and comprehensive **audit trails**. This ensures all compliance requirements are addressed natively within the application.
- **Polyglot Programming:** The modern clinical trial requires flexibility. The solution must support seamless integration and execution of all leading analytical languages—**SAS, R, Python, R Shiny**, and modern **Agentic AI applications**—all within the same GxP-compliant ecosystem.
- **Automated Governance:** Compliance with **21 CFR Part 11, GxP, and EU Annex 11** must be automated. The system itself should enforce policies, manage digital signatures, and maintain auditable logs, minimizing manual compliance effort.
- **End-to-End Traceability:** Complete visibility from data acquisition (RAW/eCRF) through to submission (Define/TLF) is mandatory. This includes tracking **data provenance, blinding/unblinding status, and the chain of custody** for every data point and derived asset.
- **Reproducibility and Transparency:** The system must guarantee that every dataset, program, and output can be **reproduced with a single-click**, supported by a full, verifiable audit trail for compliance verification.
- **Integrated AI and Automation:** Incorporating Agentic AI capabilities can drive significant efficiency gains through features like intelligent **code review, automated dataset validation,** and **report generation assistance**, while adhering to GxP principles.
- **Visualization and Control:** The platform should provide visual tracking of study progress and offer **one-click validation summary reports** to accelerate business delivery milestones such as the topline report, CSR, ISS, and ISE.

## BUSINESS IMPACT: RETURN ON INVESTMENT (ROI)

Clinical Trial Analysis and Reporting Key Performance Indicators (KPIs) must focus on business deliverables and critical activities. Implementing a comprehensive traceability framework is not merely a regulatory obligation; it is a profound driver of efficiency and cost savings. Customers who adopt a modern, integrated approach should strive to achieve significant and measurable returns on investment.

#### OPERATIONAL EFFICIENCY GAINS: ACCELERATING TIMELINES AND REDUCING MANUAL LABOR

Implementing an automated, integrated Statistical Computing Environment (SCE) directly boosts operational agility. It can substantially reduce cycle times and programmer efforts by embedding compliance as an automated byproduct of the statistical workflow.

Below are example metrics of potential improvements:

| Metric | Target Improvement |
|---|---|
| **Delivery of Top-Line TFL Results** | One (1) week to less than 1 day |
| **Delivery of CSR (Clinical Study Report)** | One (1) month to less than 1 week |
| **Answering Regulatory Health Queries Such as ‘who, what and when’….** | Reduced by up to 100+ hours |

| Metric | Target Improvement |
|---|---|
| **Code Review and Debugging Effort Per Study** | Reduced by up to 200+ hours |

The aggregated efficiency from automating change tracking, managing dependencies, and accelerating reviews translates into tangible financial benefits, with significant annual savings for an organization.

Beyond speed and cost savings, the framework dramatically reduces regulatory risk and rework. By providing full, verifiable audit trails and automated version control for both code and data, the solution guarantees the results are reproducible. By enforcing standards and automatically logging lineage and version, the risk of non-compliance findings during audits is substantially mitigated. The integrated controls ensure that compliance with GxP, 21 CFR Part 11, and EU Annex 11 is enforced natively, reducing the reliance on error-prone manual steps.

#### OTHER EXAMPLE KPIs

The success of the framework is also measured by key indicators that reflect the health and predictability of the clinical data process. These metrics highlight the benefits of integrating quality checks and governance directly into the workflow:

- SDTM/ADaM and TLF Cycle Time: Measures the days from when SDTM is finalized to ADaM completion, and the days from ADaM completion to the final TLF package delivery.
- TLF Iteration Burden per Study: Tracks the total number of iteration Quality Control (QC) loops across all Tables, Listings, and Figures (TLFs) in a study.
- QC Pass Rate: Represents the percentage of SDTM/ADaM datasets and TLFs that pass QC on the first attempt.

Achieving these KPI improvements is essential for accelerating business delivery milestones like the topline report, Clinical Study Report (CSR), Integrated Summary of Safety (ISS), and Integrated Summary of Efficacy (ISE).

## CONCLUSION

Following or implementing a comprehensive framework can provide a definitive blueprint for establishing a verifiable, reproducible, and regulator-ready clinical data environment. By systematically integrating version control, access governance, audit trails, and end-to-end lineage into a single, comprehensive application, sponsors can move beyond compliance as a manual burden and transform it into an automated, competitive advantage.

This proactive design approach is essential for the modern, data-intensive clinical trial landscape. It eliminates "black box" data steps, supports real-time data review, and ensures inspection readiness. Ultimately, this transformation accelerates the delivery of safe and effective medical treatments to patients, confirming that regulatory compliance should be a byproduct of your work—not the work itself.

## REFERENCES

International Council for Harmonisation. ICH Harmonised Guideline: General Considerations for Clinical Studies (E8).

U.S. Food and Drug Administration. Title 21, Code of Federal Regulations, Part 11: Electronic Records; Electronic Signatures. Accessed November 16, 2025.

U.S. Food and Drug Administration. Guidance for Industry: Part 11, Electronic Records; Electronic Signatures — Scope and Application. September 2003.

U.S. Food and Drug Administration. Guidance for Industry: Computerized Systems Used in Clinical Investigations. May 2007.

European Commission. EudraLex: The Rules Governing Medicinal Products in the European Union, Volume 4, Good Manufacturing Practice, Chapter 4: Computerised Systems (Annex 11).

European Medicines Agency (EMA). Guideline on good clinical practice compliance and inspection. October 2017.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Author Name: Bill Qubeck
Company: Sycamore Informatics
Address: Waltham, MA, USA
Email: bill.qubeck@sycamoreinformatics.com
Website: www.sycamoreinformatics.com

Co-author Name: Shilpa Sood
Company: Sycamore Informatics
Address: Noida, India
Email: Shilpa.sood@sycamoreinformatics.com
Website: www.sycamoreinformatics.com