ML16

AI-Driven Metadata Integration in Modern TLF Development

Ilan Carmeli, Beaconcure, Boston, USA
Chuck Gelb, Beaconcure, Boston, USA

ABSTRACT

Traditional approaches to metadata development for Tables, Listings, and Figures (TLFs) treat study artifacts (protocol, SAP, specifications, datasets, displays) as separate entities built manually or semi-manually in sequence. This fragmented approach introduces inefficiency, inconsistency, and discontinuity as revisions accumulate throughout a study. Recent advances in Generative AI now enable the extraction, linkage, and validation of information across these sources to create a unified metadata repository. This paper presents the paradigm shift from a document-centric model to a metadata-driven approach where GenAI extracts meaning from TLF, Mock Shell, SAP and more artifacts, assembles them into a unified, traceable framework, and automatically propagates changes across all dependent artifacts. We describe the technical approach, implementation timeline, and compliance considerations for this novel methodology.

INTRODUCTION

Clinical trial TLF development remains a labor-intensive, manual process. Current workflows treat study artifacts (protocols, SAPs, TLF specifications, datasets, display templates) as independent documents that must be manually created, updated, and reconciled. This siloed approach creates operational challenges: information duplication, slow change propagation, fragmented audit trails, and accumulated inconsistencies that surface late in the lifecycle.

When protocol amendments occur or specifications change, manual updates cascade through multiple artifacts. Without automated validation, inconsistencies often appear weeks into TLF generation or during final statistical review. For organizations managing dozens of concurrent studies, this inefficiency multiplies across portfolios, consuming resources and delaying submissions.

Recent advances in Generative AI create an opportunity to reimagine this workflow. Instead of treating artifacts as static documents, AI can extract structured metadata, link it across sources, and create a unified repository that serves as the canonical source of truth. This paper describes that approach and its implementation.

THE CHALLENGE: FRAGMENTED METADATA DEVELOPMENT

Traditional TLF development exhibits several systemic inefficiencies:

Beyond artifact fragmentation, many organizations lack systematic documentation of the metadata definitions and statistical assumptions underlying TLF development. Derivation rules, population definitions, censoring logic, imputation approaches, and display conventions are often embedded in code, informal notes, or individual expertise rather than in governed metadata repositories. As a result, critical rationale for decisions is not consistently captured, versioned, or reviewable.

Over time, this lack of structured documentation leads to reduced reproducibility. When a study is revisited months or years later, whether for an amendment, regulatory question, or post-marketing surveillance the original intent behind specific metadata decisions may be lost. Validation becomes difficult because the basis for outputs cannot be traced to documented rationale. Organizations also become dependent on specific individuals who hold tacit knowledge. When those individuals transition to other roles or leave the organization, institutional knowledge walks out the door, forcing teams to reconstruct decisions from code or archived files a time-consuming and error-prone process.

Without a centralized, metadata-driven framework that enforces harmonized interpretation, discrepancies may only surface during quality control or regulatory review late in the lifecycle when correction is expensive. These inconsistencies contribute to rework, delays, and potential compliance concerns, particularly in complex or adaptive study designs where multiple teams must coordinate implementation of intricate statistical logic. The cost of resolving late-stage interpretation misalignment multiplies across concurrent studies and organizational portfolios.

Statistical Analysis Plans and functional specifications are frequently interpreted differently across programmers, statisticians, and reviewers. Ambiguities in endpoint definitions, analysis populations, visit windows, or derivation logic

can result in divergent implementations. For example, one programmer may interpret a 'last observation' rule to mean the last non-missing value before a visit window closure, while another may interpret it as the last value observed regardless of timing. These subtle differences can compound across multiple variables and analyses.

Inconsistent Interpretation of SAP and Specifications

Lack of Structured Documentation of Metadata and Assumptions

Siloed Information & Manual Linkage

Study artifacts exist in separate documents. Linking them requires manual spreadsheets, emails, and meetings. When metadata changes (e.g., variable redefinition), updates must be manually propagated across all dependent artifacts, creating error-prone and slow processes.

Inefficient Revision Cycles

Protocol amendments trigger cascading manual updates across SAP, specifications, datasets, and displays. Without automated validation, inconsistencies surface late weeks into TLF generation or during statistical review.

Limited Traceability & Scalability Challenges

Traceability between TLF outputs and their metadata sources is often limited to comments or institutional knowledge. When teams rotate or studies are archived, knowledge is lost. Reusing metadata from prior studies requires manual reconstruction. As study complexity grows and organizational portfolios expand, these inefficiencies constrain scalability.

Regulatory & Compliance Risk

Regulatory expectations require transparent, auditable metadata for statistical outputs. When metadata is scattered, assembling coherent audit trails for submissions becomes labor-intensive and incomplete, creating regulatory risk.

THE SOLUTION: AI-DRIVEN UNIFIED METADATA

We propose a paradigm shift from document-centric to metadata-driven TLF development. Using Generative AI, we extract structured metadata from all sources (protocol, SAP, specifications, datasets, displays), link that metadata into a unified repository, and treat that repository as the canonical source of truth. TLF development, validation, and updates flow from this unified model.

Technical Approach: Six-Step Process

1. Extract: GenAI processes protocol, SAP, specifications, datasets, and display templates, extracting variable definitions, analysis rules, table structures, and calculation logic.

2. Integrate: Extracted metadata is linked across sources. Variables are matched, analysis flows traced, and table shells mapped to underlying data, creating a unified metadata graph.

3. Validate: Automated validation ensures consistency and completeness with accuracy gates requiring >99% confidence before presenting to users.

4. Trace: The unified repository maintains bidirectional traceability to source artifacts with complete audit trails.

5. Generate: From the unified metadata repository, AI automatically generates the actual TLF tables themselves. Rather than statisticians and programmers manually creating table outputs from scratch, the tables are generated directly from the unified metadata and underlying datasets, ensuring consistency across outputs. Generated tables retain full traceability to their metadata antecedents and source data, and users can validate or customize them before final deployment.

6. Enable: Updates to one artifact automatically propagate through the graph. Users access the repository via a controlled interface that enforces review gates and maintains audit trails.

Fast-Track Implementation

Implementation follows a rapid deployment model: Week 1-4 covers environment setup and first study pilot load with live testing. Week 2-3 incorporates user feedback and rapid refinements. Week 4+ scales to additional studies with continuous optimization. This approach enables organizations to realize value quickly while iteratively improving the system.

Validation & Regulatory Compliance

The approach incorporates multiple validation layers. An AI extraction layer processes artifacts. An automated validation engine applies consistency checks with 99%+ accuracy gates. Critical metadata requires manual review gates. All decisions generate FDA-compliant audit trail logs with change management and traceability to the source.

Risk mitigation includes: (1) No sensitive data is sent to external APIs, closed-loop processing; (2) 21 CFR Part 11 compliant audit trails; (3) vendor-neutral data formats to prevent lock-in; (4) manual override capabilities for edge cases. This de-risked approach addresses regulatory and security concerns inherent to AI-driven systems.

VALUE PROPOSITION

The unified metadata approach delivers multiple dimensions of value. For efficiency, new study metadata extraction and integration is projected to be 30-40% faster compared to traditional workflows, with study updates and revisions propagating automatically rather than manually. For quality, automated validation eliminates many manual reconciliation steps and catches inconsistencies earlier. For compliance, the complete, automated audit trail satisfies regulatory expectations for metadata transparency.

Key advantages over traditional approaches: (1) No custom coding required, GenAI handles natural language understanding across heterogeneous artifact types; (2) Backward compatible, works alongside existing workflows and tools; (3) Fast implementation pilot to deployment in weeks; (4) Low risk, validation gates and audit trails address regulatory concerns. For organizations managing concurrent studies, scalability improves as the portfolio grows since metadata reuse accelerates with each new study.

ILLUSTRATIVE SCENARIO: PROTOCOL AMENDMENT

Consider a protocol amendment that redefines a key efficacy variable mid-study. In traditional workflow:

Traditional: Amendment approved → SAP manually updated → Specifications updated → Datasets regenerated → Table shells updated → Manual reconciliation and QA across all touchpoints → Multiple iteration cycles → Inconsistencies discovered late.

AI-Driven: Amendment uploaded → AI extracts variable change → Unified metadata repository updated → Dependent artifacts automatically identified and flagged → Impact analysis presented for review → Stakeholders approve → All downstream artifacts auto-propagate changes → Complete audit trail logged → Regulatory-ready. This transformation shifts the workflow from sequential manual steps to automated, traceable propagation.

DISCUSSION

The pharmaceutical and CRO industries face unprecedented pressure. Study protocols continue growing, larger populations, more complex endpoints, more regulatory requirements. Meanwhile, submission timelines remain compressed. Simultaneously, organizations are managing increasingly larger concurrent study portfolios. Under these pressures, manual metadata management has become the limiting factor. Statisticians spend disproportionate time reconciling artifacts instead of focusing on statistical insights. Programmers interpret ambiguous specifications ad hoc instead of working from clear, integrated requirements. Managers lack visibility into true bottlenecks. This inefficiency is not sustainable.

Industry Context & Motivation

This work addresses a critical operational gap in modern clinical trials. As study complexity increases and regulatory timelines compress, the document-centric paradigm becomes unsustainable. AI-driven metadata integration offers a practical, scalable solution compatible with existing tools (SAS, EDC, regulatory platforms) without requiring a major infrastructure overhaul.

The AI-driven metadata approach offers distinct advantages over alternative strategies. First, it requires no custom coding, Generative AI's natural language understanding works across heterogeneous artifact types without specialized training. Second, it is backward compatible with existing tools (SAS, EDC systems, regulatory platforms); it augments workflows rather than replacing infrastructure. Third, implementation is rapid, from pilot to production deployment within weeks, not months. Fourth, it addresses regulatory concerns through validation gates, audit trails, and data governance controls, making it viable for submissions. Fifth, it scales: as an organization's study portfolio grows, per-study costs decrease because metadata from completed studies becomes reusable assets for future work. This creates a compounding efficiency benefit.

Comparative Advantages

Future development opportunities include integration with CDISC standards (SDTM, ADaM) for enhanced interoperability, machine learning models trained on study-specific patterns to improve extraction accuracy over time, and predictive analytics to identify consistency gaps proactively. Multi-language support and expansion to other trial artifacts (adverse event narratives, laboratory descriptions) would further broaden applicability.

CONCLUSION

AI-driven metadata integration represents a paradigm shift in TLF development. By unifying disparate study artifacts into a single, traceable, AI-managed repository, organizations can improve efficiency, consistency, compliance, and scalability. The approach is implementable within weeks, compatible with existing workflows, and addresses both operational and regulatory concerns. We encourage practitioners to evaluate this methodology through pilot studies and assess its impact on TLF development in their organizations.

CONTACT INFORMATION

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

Author Names: Ilan Carmeli
Company: Beaconcure
Address: 867 Boylston Street, 5th Floor #1491, Boston, MA
Email: Ilan@beaconcure.com
Website: www.beaconcure.com
LinkedIn: http://www.linkedin.com/company/beaconcure