Paper ML29

Revolutionizing Clinical Data Transformation: Harnessing GenAI and Open Source for Automated SDTM Conversion

Hao Chen, BeOne Medicines, Shanghai, China
Daniel Huang, BeOne Medicines, New Jersey, US

ABSTRACT

The conversion of clinical trial raw data into SDTM format remains a major bottleneck in drug development, often requiring intensive manual effort and delaying the transition from data to insight. This presentation unveils an innovative framework that utilizes Generative AI (GenAI) and open-source technologies in R and Python to fully automate the SDTM conversion process, delivering significant improvements in efficiency, transparency, and CDISC standards compliance. Our solution leverages GenAI for automated metadata interpretation and code generation, and employs open-source libraries to map raw data structures to SDTM domains. Key features include iterative learning that continuously enhances data mapping accuracy through historical mappings, user feedback, and a modular, open architecture enabling seamless integration with existing EDC platforms and leveraging Databricks advanced data analysis features. We will also discuss technical architecture and share practical insights for deploying GenAI and open-source tools effectively within regulated clinical data workflows.

INTRODUCTION

The landscape of clinical development is currently experiencing a profound data complexity explosion. Modern clinical trials are characterized by a rapid growth in data volume, velocity, and diversity, driven by the integration of novel biomarkers, wearable devices, decentralized trial designs, and real-world data sources. While this wealth of information holds the key to accelerating precision medicine, it simultaneously creates a severe operational bottleneck in data management and statistical programming pipelines.

At the core of this bottleneck is the mandatory requirement to standardize clinical trial data for global regulatory submissions, specifically adhering to the Clinical Data Interchange Standards Consortium (CDISC) standard. Bridging the gap between dynamic, real-world clinical data—which is filled with human factors, exceptions, and variability—and the strictly controlled, standardized formats required by regulatory agencies is exceptionally challenging.

Historically, traditional SDTM transformation has been a profoundly labor-intensive, slow, and error-prone process. Conventional business intelligence (BI) and rule-based automation systems have proven insufficient for modern needs, often resulting in an inherent "decision lag" that prevents real-time, adaptive decision-making. As data scales exponentially, relying purely on human programmers to manually interpret metadata, draft mapping specifications, and write transformation code is no longer sustainable.

We have reached a strategic inflection point where Artificial Intelligence (AI) is the only viable path to simultaneously improve efficiency, quality, and speed at scale. This paper presents an innovative framework that harnesses Generative AI (GenAI) and open-source technologies to fully automate the SDTM conversion process. By shifting from manual, fragmented programming to an AI-native, end-to-end orchestrated workflow, this approach not only drastically reduces turnaround times but also enhances data quality, establishing a single source of truth for downstream clinical insights.

THE EVOLUTION OF SDTM MAPPING

To fully appreciate the impact of a Generative AI-driven architecture, it is essential to understand the historical progression of clinical data transformation. The evolution of SDTM mapping can be categorized into five distinct stages, each representing a leap in automation.

Stage 1: Manual Programmer-Centric

In this foundational stage, workflows consisted of study-specific programs with highly limited reusability, leading to exceptionally long cycle times. Each clinical study demanded its own bespoke set of transformation programs, with little opportunity for code sharing or standardization across projects.

Stage 2: Metadata-Driven

Organizations moved toward central repositories to store mapping rules, allowing specifications to be auto-generated or copied. This represented a significant improvement in reusability, though the underlying logic still required substantial human oversight and manual refinement.

Stage 3: AI/ML-Assisted Mapping (NLP, Classification)

Traditional Machine Learning models were deployed to automatically suggest potential mappings, though extensive manual effort remained. Natural Language Processing (NLP) and classification algorithms helped surface candidate mappings, but the final decision-making rested firmly with human programmers.

Stage 4: Generative-AI Copilots

The advent of Large Language Models (LLMs) introduced the "Copilot Mode." Here, AI acts as an assistant to draft code and explain standards. The operational model is akin to "Mentoring an Intern" (Augmentation): the human programmer provides hand-in-hand guidance to save individual effort. While a meaningful step forward, this mode still fundamentally depends on the programmer's active participation throughout the process.

Stage 5: End-to-End (E2E) Generative AI Orchestration

The current frontier—and the focal point of this paper—is the deployment of fully orchestrated GenAI workflows. This paradigm transcends task-level assistance by autonomously managing the entire data lifecycle, from raw data ingestion to the generation of final SDTM datasets. Operating under a framework of strict human governance and "human-in-the-loop" validation, Stage 5 represents a definitive shift from mere human augmentation to true, scalable, and intelligent automation.

THE POWER OF OPEN-SOURCE: R and the Pharmaverse Ecosystem

The foundation of our automated framework rests heavily on the power, flexibility, and transparency of open-source software, specifically leveraging the R programming language. R has evolved into a mature, production-grade platform for statistical computing, data analysis, and regulatory reporting, supported by an expansive ecosystem (such as tidyverse and ggplot2) that sees broad adoption in the pharma industry.

This large, open-source package ecosystem enables rapid innovation, extensive reuse, and seamless integration with emerging AI technologies. A prime example is the Pharmaverse, a curated network of R packages for clinical reporting workflows. Within our architecture, we heavily utilize **sdtm.oak**, an EDC and Data Standard agnostic transformation engine that automates the conversion of raw clinical data into structured SDTM datasets using standardized mapping algorithms.

To solve the regulatory requirement for human oversight, our framework employs **R Shiny** to build the user interface. Shiny provides a robust framework for developing interactive analytics applications and dashboards, supporting critical Human-in-the-Loop (AI + HI) workflows. This allows subject matter experts to interact with the data in near real-time via web-deployed R logic.

In addition to R, the framework leverages Python for orchestration, AI model integration, and infrastructure-level tasks, along with Databricks for cloud-scale compute, advanced data analysis, and the real-time data lakehouse architecture that underpins the system's storage and retrieval layers.

SATA: THE SDTM AUTOMATED TRANSFORMATION ASSISTANT

To operationalize the convergence of GenAI and open-source computing, We have developed **Sata** (SDTM Automated Transformation Assistant). Sata is an intelligent virtual agent and all-in-one integrated web application designed to autonomously transform raw clinical data into fully compliant SDTM datasets.

Sata eliminates the need for manual mapping specifications; instead, it can ingest metadata directly from raw data sources, Architect Loader Spreadsheet (ALS) files, BlankCRFs, or data descriptions. Operating through three core engines—AI Prediction, a Coding Agent, and Online Review & Update—Sata utilizes standard CDISC rules to generate structured mappings.

The Sata framework is composed of five interconnected modules:

Module 1：Data Ingestion and Profiling Engine:

This module securely connects to EDC systems or raw data repositories. It ingests raw datasets and automatically generates metadata profiles (e.g., data types, missingness, value distributions). This profiling provides the necessary context for the AI to understand the structural realities of the incoming data.

Module 2：AI Metadata Interpreter & Mapper:

Acting as the "brain" of the system, this module feeds the raw metadata and study protocols into the LLM. The AI evaluates the source data against CDISC SDTM standards and generates a comprehensive mapping specification. It identifies the target domain (e.g., AE, DM, VS) and maps source variables to target variables, applying necessary derivations or terminology conversions.

Module 3：Dynamic Code Generator:

Once the mapping specification is approved (or auto-validated based on confidence scores), Sata's code generator translates the logic into executable Python or R scripts. This module uses prompt engineering techniques to ensure the generated code adheres to industry best practices and internal coding standards.

Module 4：Execution and Transformation Pipeline:

Hosted on Databricks, this pipeline executes the AI-generated code against the raw data. It performs the physical data transformation, handling complex joins, transpositions, and formatting to produce the final SDTM datasets.

Module 5：Validation and Continuous Learning Loop (Human-in-the-Loop):

Crucial for regulatory compliance, Sata includes an automated validation suite (similar to Pinnacle 21) that checks the output against CDISC rules. Any discrepancies or low-confidence AI mappings are flagged for human review. When a human programmer corrects a mapping or modifies the code, Sata captures this feedback, updating its vector databases and few-shot prompting examples to improve future accuracy.

SYSTEM ARCHITECTURE

The automated SDTM framework is built upon the Databricks Lakehouse Platform Foundation, utilizing an "AI Acceleration Turbine" to process data. This robust infrastructure is comprised of several interconnected layers:

Data & Governance Layer: Utilizes Unity Catalog to manage data pipelines securely.

Compute & Scheduling Layer: Provides serverless compute and job scheduling.

Model & Validator Layer: Houses the core AI models and business rules.

AI Knowledge Base Layer: Leverages Vector Search and a comprehensive CDISC Library.

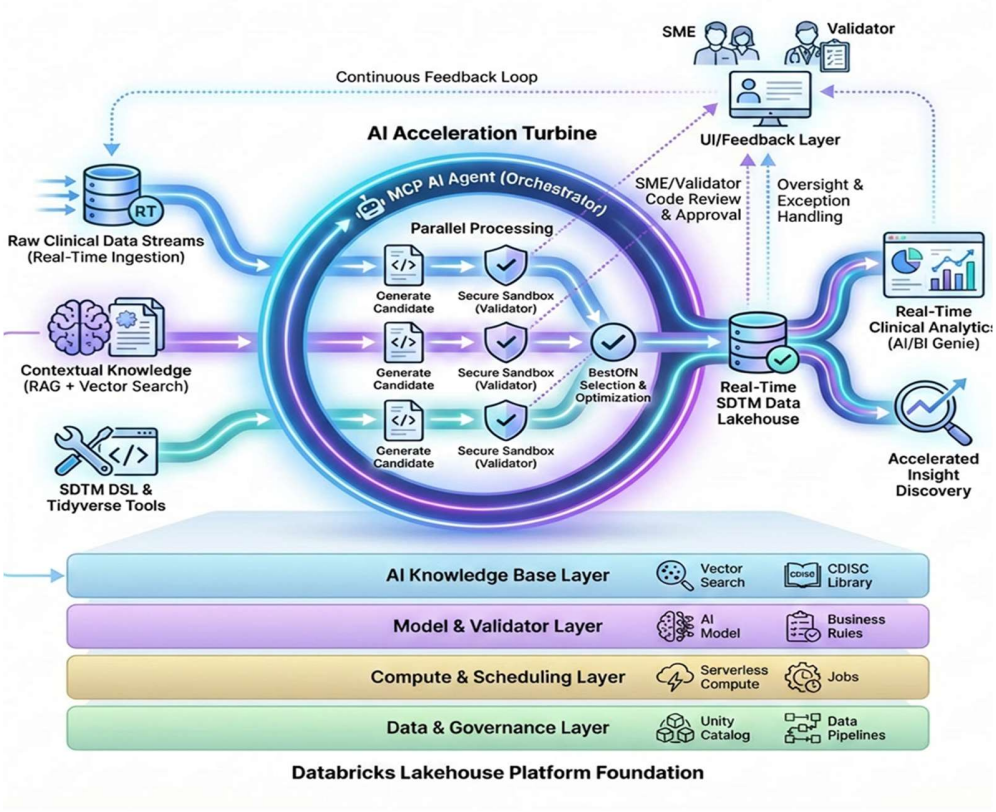

Figure 1. System Architecture

When raw clinical data streams are ingested in real-time, they are enriched with contextual knowledge (via RAG and Vector Search) and processed using SDTM DSL and Tidyverse tools. An MCP AI Orchestrator Agent drives the data into a Parallel Processing loop where multiple candidate mappings are generated. Each candidate is rigorously tested within a Secure Sandbox (Validator), and a "BestOfN Selection & Optimization" algorithm chooses the superior logic. The output is then pushed to a Real-Time SDTM Data Lakehouse to support accelerated insight discovery and clinical analytics.

KEY FEATURES AND CORE BENEFITS

The Sata framework is delivered as an all-in-one integrated web application, providing a unified interface for data upload, mapping review, code inspection, execution monitoring, and output validation. This design eliminates the need for users to navigate between multiple tools and environments, reducing friction and cognitive overhead.

The efficiency gains are dramatic. The time required to produce initial SDTM datasets can be reduced from hundreds of programmer-hours to merely a few hours during the automated prediction and code generation phases. Furthermore, by orchestrating the entire end-to-end process—from raw data upload to final SDTM output, compliance checking, and iterative refinement, the system can deliver complete, finalized results in just a few days, compared to the weeks or months demanded by conventional approaches.

The framework supports broad data inputs, accommodating both corporate-standard and industry-standard raw data formats. This flexibility is essential in an industry where different EDC platforms, CROs, and partner organizations produce data in varying formats and conventions.

The core benefits extend across the entire clinical data value chain. Faster SDTM delivery and AI-augmented cloud-scale compute reduce the time from data collection to analysis-ready datasets. Less manual work and fewer errors translate to higher data quality and lower rework costs. Earlier availability of standardized data enables better and faster decisions about trial conduct, safety signals, and efficacy trends. Accelerated downstream workflows—

including analysis dataset creation, statistical analysis, and regulatory submission preparation—contribute directly to faster time-to-market for new therapies. And by establishing a single source of truth for SDTM transformation logic, the system enhances consistency, reproducibility, and audit readiness across the organization.

THE END-TO-END MULTI-AGENT WORKFLOW

Sata avoids relying on a single monolithic model by instead orchestrating a highly specialized, multi-agent end-to-end workflow. This workflow autonomously processes data from raw upload through AI mapping prediction, code generation, execution, debugging, SDTM checking, and final output generation.

Phase 1: Information Extraction

The **Parser Agent** ingests ALS files to create JSON inventories, while the **Librarian Agent** retrieves CDISC standards and the **Rules Curator** formats company-specific mapping rules. Together, these agents establish the complete contextual foundation required for downstream mapping.

Phase 2: Predictive Mapping

The **Domain Assignment Agent** and **Mapping Agent** propose logical connections between raw data fields and SDTM target variables. These are iteratively verified by a **Critic Agent**, culminating in human validation to produce a final Mapping Result JSON. The iterative critic loop ensures that mapping proposals are internally consistent before they reach human reviewers.

Phase 3: Execution & Testing

The **Code Generator** translates the mappings into executable R code, leveraging the sdtm.oak framework. A **Code Parser** analyzes the oak codebase to allow the **Test Case Generator** to build automated tests. The R code is run by an **Executor in Sandbox** and monitored by a **Code Tester** for auto-debugging.

Phase 4: Finalization

An **SDTM Executor** produces the final structured SDTM tables, subject to a final round of human validation. The output datasets are accompanied by full provenance metadata, including the mapping JSON, generated code, test results, and validation logs.

The following illustrates the Mapping Prediction and Code Executor agents:

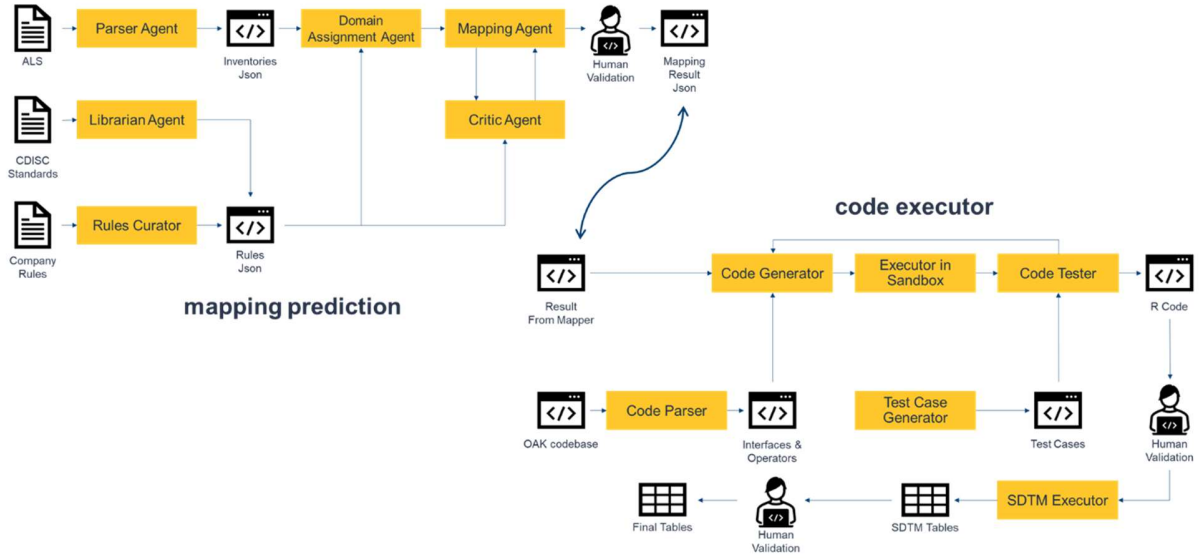

Figure 2. Mapping Prediction and Code Executor agents

PREDICTION METHODS AND PERFORMANCE METRICS

To balance computational efficiency with regulatory accuracy, the Sata predictive engine cascades through four distinct prediction methodologies.

Prediction Methodology Cascade

Method 1 – Exact Match (100% Confidence): Direct alignment of labels and definitions requiring no deep thinking, offering "Good" traceability. This method is applied first and resolves straightforward, unambiguous field mappings.

Method 2 – Fuzzy Match (>95% Confidence): High-confidence similarity algorithms that resolve minor discrepancies without deep reasoning. Fuzzy matching handles variations in naming conventions, abbreviations, and minor label differences across studies.

Method 3 – Fuzzy + Reasoning (70–95% Confidence): Generates batches of candidates with AI-generated rationale and deterministic workflow checks. This method is invoked when simple string matching is insufficient and contextual understanding of the variable's purpose is required.

Method 4 – Deep Reasoning (Search): Concept-level inference that requires understanding the CDISC Implementation Guide and historical mappings. This method naturally requires post-processing and strict workflow controls, and is reserved for the most complex or novel mapping scenarios.

Empirical Performance Metrics

Evaluations across four distinct clinical studies (Studies A, B, C, and D) demonstrate remarkable accuracy. The following table summarizes the key metrics:

Metric	Study A	Study B	Study C	Study D	Average
Rule Accuracy	96.84	93.45	93.64	94.53	93.94
Field Accuracy	93.83	89.13	86.75	90.71	89.29
Rule Accuracy (ex-SUPP)	97.95	92.88	93.93	96.52	94.93
Field Accuracy (ex-SUPP)	95.98	89.56	87.80	94.13	91.36

Table 1. Prediction Accuracy Across Four Clinical Studies

Because SUPP domains are highly study-specific, removing them from the evaluation highlights the system's core capabilities even further. Rule Accuracy (ex-SUPP) reached 94.93% in average, and Field Accuracy (ex-SUPP) hit 91.36%, demonstrating that the system handles standard SDTM domains with good precision.

HUMAN-IN-THE-LOOP (AI + HI) GOVERNANCE

The deployment of AI in the highly regulated biopharmaceutical industry demands absolute traceability. Therefore, Sata is wrapped in a strict Human-in-the-Loop (HITL) governance model.

The system is designed to purposefully surface uncertainty and isolate potential errors. Through the R Shiny UI, programming subject matter experts (SMEs) can conduct rapid expert reviews. The interface highlights low-confidence mappings, flags inconsistencies with historical patterns, and provides side-by-side comparisons of AI-proposed versus reference mappings.

Crucially, this interaction creates a continuous feedback loop. Every SME correction is fed back into the AI Knowledge Base, allowing the models to continuously improve and adapt to specific data nuances over time, while strictly maintaining regulatory compliance. This learning mechanism ensures that the system becomes increasingly accurate with each study processed, embodying a virtuous cycle of human expertise and machine intelligence.

CONCLUSION

The SDTM Automated Transformation Assistant represents a meaningful advance in the application of Generative AI to clinical data management. By combining LLM-powered mapping prediction, automated code generation in R, multi-agent orchestration, and human-in-the-loop governance, the framework addresses the fundamental limitations of manual and rule-based approaches to SDTM conversion.

Generative AI is reshaping the biopharma industry, and clinical trials represent a prime transformation area where AI can deliver faster cycle times and smarter execution. The Sata framework demonstrates that this promise can be realized practically, with measurable accuracy across diverse studies and a workflow design that respects the compliance and quality requirements of regulated environments.

The combination of a curated knowledge library with LLM reasoning capabilities enables an end-to-end AI system in which AI handles mapping suggestions, code drafting, reasoning, testing, and iteration, while standard and deterministic methods ensure quality control and traceability. The HITL workflow ensures that this automation augments rather than replaces human expertise, surfacing uncertainty, isolating errors, and enabling rapid expert review. As reviewer corrections flow back into the system, the learning loop closes—models improve continuously while compliance is maintained.

Looking ahead, we anticipate that this architectural pattern—AI agents governed by human oversight, built on open-source foundations, and continuously refined through operational feedback—will become the standard approach for clinical data transformation across the industry. The open-source nature of the underlying technologies, including R, the pharmaverse ecosystem, and sdtm.oak, ensures that the innovations demonstrated here are accessible to the broader clinical data science community and can be adapted, extended, and improved collaboratively.

REFERENCES

CDISC. *SDTM Implementation Guide* https://www.cdisc.org

POSIT. Easy web apps for data science without the compromises. https://shiny.posit.co

Databricks. Bring AI to your data to help you bring AI to the world. https://www.databricks.com
Pharmaverse. *sdtm.oak: EDC and Data Standard Agnostic SDTM Transformation Engine*. https://pharmaverse.github.io/sdtm.oak
Hao Chen. Leveraging R for Real-Time Data Analysis and Reporting in the AI+HI Paradigm. https://pharmarug.github.io/pharmarug-2024/program.html
Anthropic. Claude Opus 4.6 (February 2026 version) [Large language model]. https://www.anthropic.com
Google DeepMind. Gemini 3.1 Pro (February 2026 version) [Large language model]. https://deepmind.google
OpenAI. GPT-5.2 (December 2025 version) [Large language model]. https://openai.com

ACKNOWLEDGMENTS
The authors would like to thank the BeOne Medicines Scientific Programming teams for their invaluable contributions to the development of the Sata framework.

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

Author Name: Hao Chen
Company: BeOne Medicines Co., Ltd. (formerly BeiGene Co., Ltd.)
Address: 16th floor of CPIC XINTIANDI TOWER 2, 288 Xizang Road (S), Shanghai, China
Mobile: (+86) 186 0212 2927
Email: hao1.chen@beonemed.com
Website: https://www.linkedin.com/in/hao-harry-chen-98709128

Author Name: Daniel Huang
Company: BeOne Medicines USA, Inc. (formerly BeiGene USA, Inc.)
Address: New Providence, NJ
Mobile: (434) 882-1408
Email: daniel.huang@beonemed.com
Website:https://www.linkedin.com/in/jian-hua-daniel-huang-9973183b