Paper ML20

SDTM Specification Automation Using an ML Approach to Provide Mappings

Benzi Alex Mathew, Saama Technologies, Campbell CA, USA
Jaya Simha Inampudi, Saama Technologies, Campbell CA, USA

ABSTRACT

We present an AI-powered clinical data management tool for automating the creation of SDTM specifications. By leveraging past mappings and knowledge within a therapeutic area, along with a deep understanding of CDISC standards, users can dynamically generate customized SDTM specifications for new studies.

The AI engine analyzes historical data transformations, mapping patterns, and CDISC guidelines to propose a tailored set of SDTM datasets, variables, and conversion rules. This automation eliminates the tedium of manual effort typically required to design SDTM specifications from scratch, saving time and ensuring consistency across studies.

Furthermore, the generated specifications are user-editable, enabling subject matter experts to refine the AI-proposed logic as needed. This collaborative approach combines the efficiency of AI with the flexibility for human oversight, optimizing the SDTM specification development process.

This abstract demonstrates how artificial intelligence can augment clinical data management, streamlining regulatory submission requirements through intelligent automation.

1. INTRODUCTION

The Study Data Tabulation Model (SDTM) is a critical standard for organizing and submitting clinical trial data to regulatory agencies. Traditional SDTM specification creation is labor-intensive, prone to inconsistencies, and relies heavily on manual mapping, which increases the risk of errors. With the increasing volume and complexity of clinical data, there is a pressing need for automated tools that not only generate accurate SDTM specifications but also incorporate historical knowledge and industry standards.

In this paper, we introduce an AI-powered clinical data management tool designed to automate the creation of SDTM specifications. The tool leverages past mappings, a deep understanding of CDISC standards, and advanced AI techniques to generate initial SDTM mappings, which can then be refined by subject matter experts. Our approach is structured into three distinct phases—Input, Agentic Flow, and Output—each contributing to a layered and robust automation process.

This approach significantly reduces specification drafting time while maintaining high CDISC compliance, aligning with FDA goals and addressing challenges in clinical data management.

2. BACKGROUND AND RELATED WORK

2.1 SDTM AND CDISC STANDARDS

SDTM, as defined by the Clinical Data Interchange Standards Consortium (CDISC), provides a structured format for clinical trial data. The primary purpose of SDTM is to facilitate regulatory review and ensure consistency across submissions. However, creating SDTM-compliant datasets requires a deep understanding of CDISC guidelines and often involves repetitive, manual mapping tasks that are both time-consuming and error-prone.

2.2 TRADITIONAL METHODS FOR SDTM SPECIFICATION DEVELOPMENT

Traditionally, SDTM specification development has been a manual process, requiring programming and data management professionals to interpret source data and SDTM Implementation Guides (IGs) to create mappings. This manual approach is resource-intensive and lacks scalability when dealing with large datasets or multiple studies. While some organizations have developed tools to automate SDTM specs generation using standardized metadata, these tools are ineffective when study-specific data deviates from global standards, necessitating manual intervention.

2.3 EMERGENCE OF AI IN CLINICAL DATA STANDARDIZATION

Recent years have seen increasing interest in applying artificial intelligence to clinical data management. Previous approaches have primarily focused on rule-based automation and simple machine learning models for pattern recognition in data mapping. However, these solutions often struggle with the complexity and variability of clinical data structures.

2.4 RAG TECHNOLOGY IN HEALTHCARE

Retrieval-Augmented Generation represents a significant advancement in AI technology, combining the power of large language models with precise information retrieval mechanisms. While RAG has shown promise in various healthcare applications, its application to SDTM automation represents a novel approach to clinical data standardization.

3. METHODOLOGY AND SYSTEM ARCHITECTURE

Our solution employs a three-phase process flow: Input, Agentic Flow, and Output.

3.1 INPUT PHASE

The system uses three primary data sources to begin the mapping process:

- **Source Data/Metadata**: This includes the raw clinical trial data and its associated metadata, which needs to be converted to SDTM format.
- **SDTM IG (Implementation Guide) + SDTM Model (xls) + Historical Specs:** The standard SDTM documentation (in PDF and XLS formats) is used in addition to historical mapping specifications from past studies. The historical context gives examples of successful transformations, which improves the system's understanding of domain-specific requirements and informs the mapping process.
- **Global Function Definitions**: These predefined functions and transformation rules ensure that data is consistently processed according to CDISC standards and guide the mapping process.

3.2 AGENTIC FLOW PHASE

The Agentic Flow Phase is the core of the system, where intelligent processing takes place to generate an initial SDTM mapping. This is achieved through several integrated components:

- The **HyDE (Hypothetical Document Embedding) Query Transformation** uses Claude 3.5 Sonnet and curated context from the input phase to produce a hypothetical SDTM mapping, which serves as an educated draft based on the system's understanding of SDTM principles.
- Simultaneously, the **Milvus Vector DB** processes the SDTM documentation using the ncbi/MedCPT-Query-Encoder to create mathematical representations (embeddings). This makes the documentation searchable and comparable in a high-dimensional space.
- A **two-stage context retrieval process** is then employed:
    a. **Context Retrieval** uses an ensemble approach (BM25 and similarity matching) to find relevant sections of the SDTM guidance, and
    b. **Context Re-rank** uses a cross-encoder (marco-MiniLM-L-6-v2) to refine and prioritize the retrieved context.
- The final stages integrate and finalize the mapping:
    a. **RAG Prompt** combines source table data with refined context and retrieved functional specifications to form the basis for the final mapping, and
    b. **LLM Invocation** uses a two-stage approach with
        i. Maker model to generate the initial SDTM mapping based on the integrated context, and
        ii. Checker model to validate the mapping against the SDTM context and CDISC standards.

3.3 OUTPUT PHASE

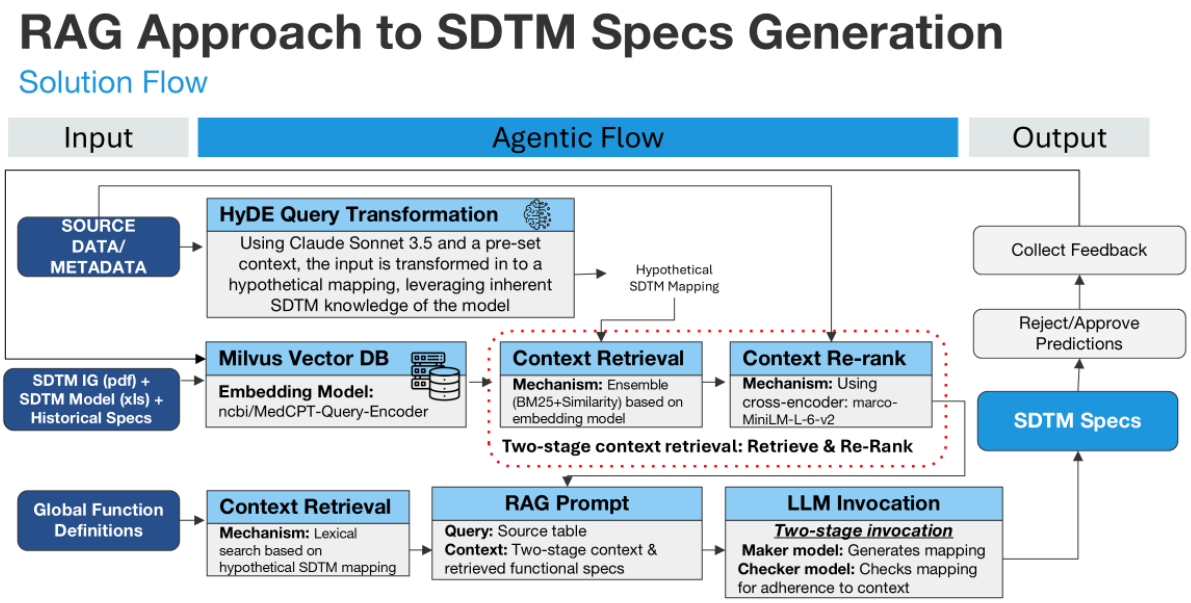
During the Output Phase, the system generates the final SDTM specifications:

- **SDTM Specification Generation:** The validated mapping is transformed into a comprehensive set of SDTM specifications.

- **Validation and Feedback Cycle:** The specifications are then validated by subject matter experts, who can either approve or reject the mappings. The feedback provided is incorporated into the system and used to update the historical specifications database and enhance the precision of future mappings. The system's effectiveness lies in its multi-layered approach to ensuring accuracy. Each component builds upon the previous one, with multiple validation steps and the incorporation of historical knowledge.

The two-stage approach to both context retrieval and LLM invocation helps maintain high standards of accuracy while

adhering to SDTM requirements.



3.4 KEY FEATURES

1) **PROCESS ALIGNEMNT**
    a) **Human-Centric Design:** The solution mirrors manual SDTM specification workflows (e.g., source analysis → mapping → validation) to minimize the learning curve. Outputs are presented in familiar spreadsheet/PDF formats commonly used by clinical programmers.
    b) **Regulatory Compliance by Default:** CDISC/SDTM rules are embedded directly into the workflow, like how human experts reference implementation guides.
2) **TECHNICAL FOUNDATION**
    a) **State-of-the-Art LLM Backbone:** The solution utilizes Claude Sonnet 3.5 for its intrinsic SDTM knowledge from training data, context-aware reasoning about study-specific requirements, and natural language interpretation of metadata.
    b) **Hybrid AI Architecture:** The solution combines Generative AI (hypothesis creation via HyDE), Retrieval Systems (precision context via RAG), and Rule-Based Constraints (global function specs as guardrails).
3) **KNOWLEDGE INTEGRATION**
    a) **HyDE (Hypothetical Document Embeddings):** HyDE generates "first-draft" mappings by asking the LLM: "What would a perfect SDTM spec look like for this source data?" and using this hypothetical template to guide context retrieval.
    b) **Contextual Layering:** The LLM knowledge is augmented with study-specific context (SDTM IG versions, therapeutic area nuances), historical wisdom (proven mappings from past studies), and global rules (institutional best practices, e.g., "Always map DATE variables to ISO 8601").
4) **PRECISION ENGINEERING**
    a) **Retrieve-and-Re-Rank Pipeline:** The pipeline broadly identifies relevant SDTM guidance sections using BM25 + Vector Search and reorders results by contextual relevance to the specific mapping task using Cross-Encoder (marco-MiniLM).
    b) **Agentic Workflow:** The workflow intelligently chains components: HyDE Draft → Context Retrieval → RAG-Augmented Generation → Checker Validation. Each layer refines outputs while maintaining auditability.

5) **VALIDATION AUTOMATION**
   a) **Two-Stage LLM Validation:** A Maker Model proposes mappings, and a Checker Model acts as an "AI reviewer" using retrieved SDTM guidelines and predefined functions (e.g., "Check that all required DOMAINS exist").
   b) **Feedback Loops:** Human approvals/rejections train the system to improve future context prioritization and hypothesis generation.
6) **EVOLUTIONARY DEVELOPMENT**
   a) **Iterative Optimization:** Components were refined through A/B testing of retrieval strategies, precision/recall measurements for context ranking, and human feedback on mapping plausibility.
   b) **Adaptive Knowledge Base:** The historical specs database grows richer with each study, enabling better pattern recognition for rare variables and therapeutic-area-specific tuning.

4. EVALUATION STRATEGY

4.1 VALIDATION APPROACH

The tool and its output were evaluated by the following approach:

- Subject matter experts reviewed the comparison between recommended mappings and actual mappings from historical study information.
- Error patterns, edge cases, and complex mapping scenarios were analyzed.
- The tool was fine-tuned based on the observations.

4.2 RESULTS

The model was tested on a phase 2 study for different types of data (events, interventions, findings, special purpose). The model's accuracy in predicting correct variables was evaluated across nine different source forms. Accuracy ranged from 82.4% for the "vits" source form to 100% for the "demog", "drugdisp", and "exposure" forms. Overall, the model accurately predicted 83 out of 118 total variables, for an overall accuracy rate of 89.8%.

Source Form	# Correct Variables Predicted	# Total number of Variables	% Accuracy
aesae	17	20	85.0%
cmeds	11	12	91.7%
demog	8	8	100.0%
drugdisp	5	5	100.0%
ecg	12	13	92.3%
eq5d3l	18	18	100.0%
exposure	14	14	100.0%
phexam	10	12	83.3%
vits	14	17	82.4%
Grand Total	83	118	92.7%

5. DISCUSSION

5.1 POTENTIAL IMPACT

The proposed system could significantly reduce the time and effort needed for SDTM implementation while maintaining or even improving quality.

Key benefits include:

- Increased consistency across studies

- Reduced manual effort in specification creation
- Faster study setup and implementation
- Automated validation against SDTM IG
- Better leverage of historical mapping decisions

5.2 CHALLENGES AND MITIGATION

- **Handling Novel/Unusual Data Structures**: Flag low-confidence outputs for SME review.
- **Validation of AI-Generated Specifications**: Provide confidence score and justification for mapping, for SME review.
- **Implementation of Sponsor-Specific Rules**: Allow editable global function definitions.

5.3 FUTURE WORK

Large language models (LLMs) are rapidly evolving and have become significantly more efficient since the start of this project. We intend to continuously improve this tool by incorporating state-of-the-art LLM and encoder models. Additionally, we will refine it for specific therapeutic areas by including CDISC Therapeutic Area User Guides (TAUGs) alongside the SDTM Implementation Guide (IG).

6. CONCLUSION

This AI-driven solution automates the creation of SDTM specifications, which are essential for submitting clinical trial data to regulatory bodies. The tool streamlines the clinical data submission process, using a three-phase process: Input (gathering source data, SDTM documentation, and historical specs), Agentic Flow (intelligent processing and mapping), and Output (generating and validating the final SDTM specifications).

By leveraging historical data, CDISC standards, and advanced AI techniques, the tool ensures high compliance and consistency while significantly reducing the time and effort required for SDTM specification development. Human oversight and feedback are incorporated into the multi-layered approach for continuous improvement. Overall, this AI-driven solution offers benefits such as increased consistency, reduced manual effort, and faster study setup.

CONTACT INFORMATION

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

Author Name: Benzi Mathew
Company: Saama Technologies
Address: 900 E Hamilton Ave Suite 430, Campbell, CA 95008
Email: benzi.mathew@saama.com

Author Name: Jaya Simha Inampudi
Company: Saama Technologies
Address: 900 E Hamilton Ave Suite 430, Campbell, CA 95008
Email: jaya.inampudi@saama.com