Paper ML14

Enhancing Clinical Data Standardization Through Intelligent Mapping of eCRF Forms and Items to SDTM

Saransh Khandelwal, Lifio, Bangalore, India

Kedar Deshpande, Lifio, Bangalore, India

ABSTRACT

Standardizing clinical trial data is essential for regulatory submission, cross-study analysis, and efficient downstream processing. A key challenge in this process is mapping electronic Case Report Form (eCRF) data to the Study Data Tabulation Model (SDTM), which provides a consistent structure for regulatory review and analysis. Manual mapping is often resource-intensive and error-prone due to variability in eCRF design and terminology across studies.

This paper presents an automated, intelligent approach to streamline two core mapping tasks: eCRF form–to–SDTM domain mapping and eCRF item–to–SDTM variable mapping. The solution integrates semantic similarity detection, syntactic pattern recognition, rule-based refinement, and controlled use of large language models (LLMs) to interpret and align source data with SDTM standards. A multi-stage pipeline progressively improves mapping accuracy by incorporating contextual metadata and domain-specific logic.

By reducing manual effort while improving consistency and transparency, this approach supports faster, scalable, and more reliable clinical data standardization advancing both operational efficiency and regulatory readiness.

1. INTRODUCTION

Clinical trial data must be transformed into standardized formats to support regulatory submissions, cross-study analyses, and long-term data reuse. The Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) serves as the foundational standard for regulatory review by agencies such as the FDA and EMA.

Despite the availability of well-defined SDTM Implementation Guides, a persistent challenge remains mapping study-specific eCRF data to SDTM domains and variables. eCRFs are often designed independently across studies, therapeutic areas, and vendors, resulting in variability in form structures, item naming conventions, and data granularity. As a result, SDTM mapping remains heavily manual, requiring significant domain expertise and repeated quality control cycles.

This paper describes an applied, production-oriented solution that automates both **form-level domain classification** and **item-level variable assignment**, while preserving transparency and allowing human oversight. The goal is not to replace SDTM programmers, but to significantly reduce repetitive manual effort and improve mapping consistency across studies.

2. MAPPING CHALLENGES IN CLINICAL DATA STANDARDIZATION

2.1 Variability in eCRF Design

Electronic Case Report Forms (eCRFs) exhibit substantial variability in structure, organization, and terminology, even when capturing equivalent clinical concepts. The same SDTM-relevant concept may be represented using different form layouts, question phrasing, or structural groupings across studies.

For example, **Exposure (EX)** data may be collected using a dedicated Exposure form in some studies, while in others it may be embedded within a **Dose** or **Treatment Administration** form. Similarly, **Disposition (DS)** information is frequently distributed across multiple forms, such as *End of Trial*, *Informed Consent*, or other study-specific completion forms, rather than being captured within a single consolidated structure. In addition, information related to the **Reproductive System (RP)** may appear within dedicated reproductive health forms in some studies, while in others it is captured as part of broader medical history, screening, or visit-based assessment forms. **Adverse Events (AE)** data may likewise be split between dedicated AE forms and general assessment forms depending on study design.

Such heterogeneity limits the effectiveness of deterministic, rule-based approaches that rely on form names or predefined structural assumptions. Consequently, reliable form-to-domain alignment requires contextual interpretation that considers form intent and content rather than surface-level naming conventions alone. This variability represents a fundamental challenge for scalable and automated form-to-domain mapping in clinical data standardization workflows.

2.2 Complexity of Item-Level Mapping

Item-to-variable mapping introduces additional complexity:

- One item may map to multiple SDTM variables
- Context (domain, form intent, item text, result vs. metadata) affects mapping
- Controlled terminology and CDISC conventions must be respected

2.3 Limitations of Manual Approaches

Manual mapping is:

- Time-consuming and resource-intensive
- Difficult to scale across multiple studies
- Susceptible to inconsistency across programmers and vendors

These challenges motivate the need for an automated yet controlled mapping framework.

3. SCOPE AND OBJECTIVES

The solution focuses on two complementary automation tasks:

1. **eCRF Form (+Item) → SDTM Domain Mapping**
2. **eCRF Item → SDTM Variable Mapping (Given Domain Mapping)**

Design objectives include:

- High mapping accuracy with explainable logic
- Reduction in manual effort
- Scalability across studies and therapeutic areas
- Alignment with CDISC and regulatory expectations

4. OVERALL SOLUTION ARCHITECTURE

The solution uses a **hybrid architecture** combining:

- Semantic similarity using sentence embeddings
- Syntactic pattern matching
- Rule-based CDISC logic
- Limited, controlled use of LLMs for contextual interpretation

A multi-stage pipeline allows mappings to be refined progressively rather than relying on a single prediction step. Outputs from each stage are aggregated to determine the most reliable final mapping.

5. FORM(ITEM)-TO-SDTM DOMAIN MAPPING

Form-to-domain mapping is implemented using a staged approach:

1. **Initial Semantic Matching**
   Initial domain assignment is performed using semantic similarity computed between eCRF metadata and SDTM domain definitions. On the eCRF side, weighted embeddings [1] are generated from **form labels and item descriptions**, while on the SDTM side embeddings are derived from **domain labels and domain descriptions**.

   The resulting vector representations are compared using cosine similarity, and similarity scores are evaluated against **calibrated confidence thresholds**. High-confidence matches exceeding the upper threshold are assigned directly, while lower-confidence cases are forwarded to subsequent refinement stages. This weighted, dual-context

representation enables more accurate semantic alignment by capturing both form intent and SDTM domain semantics.

2. **Rule-Based Refinement**
   Following initial semantic assignment, domain-specific exclusion rules are applied to eliminate inapplicable SDTM domains, such as derived or trial design domains. For records with ambiguous semantic matches, additional contextual constraints are used to refine domain selection.

   For **Findings domains** - including **Laboratory Tests (LB)** and **Electrocardiogram (EG)** - syntactic rules based on naming patterns and variable characteristics are applied to strengthen domain discrimination. Similarly, **Disposition (DS)** mappings leverage domain-specific syntactic cues [for e.g. End of Trial, Informed Consent] derived from form intent and item structure to improve classification accuracy.

   This rule-based refinement layer complements semantic similarity scoring by enforcing SDTM conventions and reducing misclassification in structurally similar or context-sensitive domains.

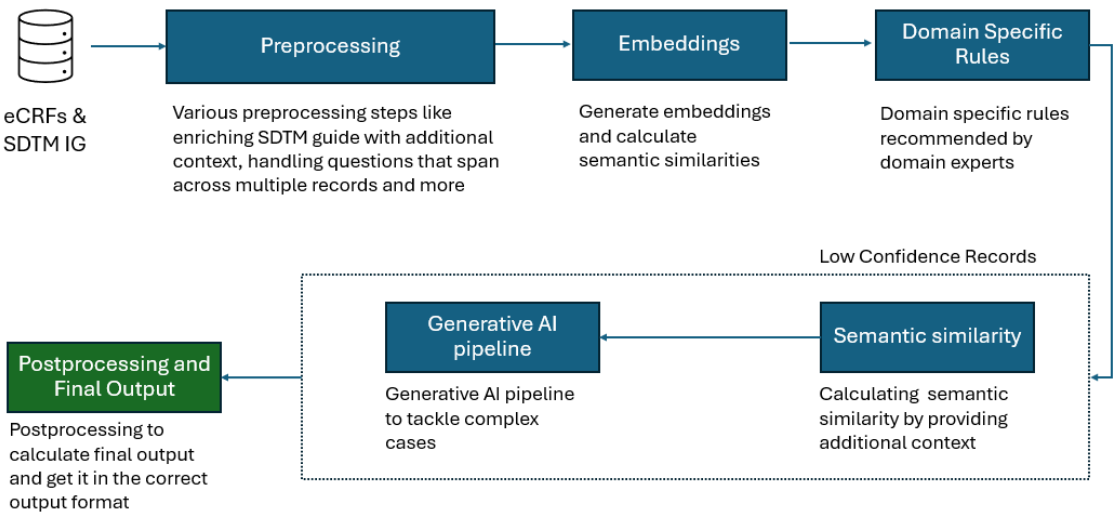

Fig.1 Methodology-Form to Domain Mapping

3. **Contextual Re-Evaluation**
   For low-confidence cases, additional metadata and expert-provided contextual descriptions for SDTM Domains are introduced to improve semantic understanding.

4. **Targeted LLM Classification**
   A controlled generative AI step is used for domains with known ambiguity (e.g., Concomitant Medications), ensuring outputs remain constrained and auditable. We

used Qwen 2.5 Model [3] for this step - to specifically identify records belonging to the Concomitant Medication (CM) SDTM domain.

5. **Final Aggregation**
   Predictions from all stages are evaluated together to determine the final SDTM domain assignment.

6. HUMAN-IN-THE-LOOP

To mitigate error propagation from incorrect domain assignments, a human-in-the-loop validation step is incorporated between form-to-domain and item-to-variable mapping. Domain predictions generated by the automated classifier are reviewed by clinical data experts, who can confirm or correct assigned SDTM domains prior to downstream processing. This intervention ensures that variable-level mapping operates on validated domain constraints, significantly reducing cascading errors caused by incorrect domain classification. The approach balances automation with expert oversight, preserving efficiency while maintaining high mapping accuracy in regulated environments.

7. ITEM-TO-SDTM VARIABLE MAPPING

This section describes the methodology used to automatically map individual eCRF items to their corresponding SDTM variables within a pre-assigned domain. The approach follows a staged pipeline that first enriches and standardizes source metadata, then applies a hybrid semantic and syntactic matching strategy, and finally enforces CDISC-driven business rules.

7.1 Preprocessing and Context Enrichment

Preprocessing is applied to both SDTM metadata and eCRF item definitions to standardize inputs and enrich semantic context prior to item-level mapping. SDTM Implementation Guide content is first flattened to extract dataset- and variable-level information, followed by removal of system and identifier variables (e.g., DOMAIN, STUDYID, SUBJID, EPOCH) that are not relevant for semantic comparison. Variable labels are enhanced by appending domain context for generic terms such as *category*, *location*, or *time*, improving semantic specificity. Special handling is applied for result variables (e.g., ORRES) in domains such as LB, VS, and EG by substituting domain-specific result descriptions.

On the eCRF side, nested form structures are flattened to the item level and filtered to retain only valid records. Each eCRF item carries a **pre-assigned predicted_sdtm_domain**, obtained from the preceding form-to-domain classification stage, which constrains the candidate variable search space. Question text is cleaned to remove formatting artifacts and enriched with contextual cues, including result-related prefixes and explicit differentiation between original and standardized units. eCRF items are then merged with SDTM domain metadata

corresponding to the predicted domain and undergo the same contextual label enhancement as SDTM variables. This ensures semantic alignment between source items and candidate SDTM variables while preserving domain-level constraints established upstream.

This enrichment improves downstream semantic matching.

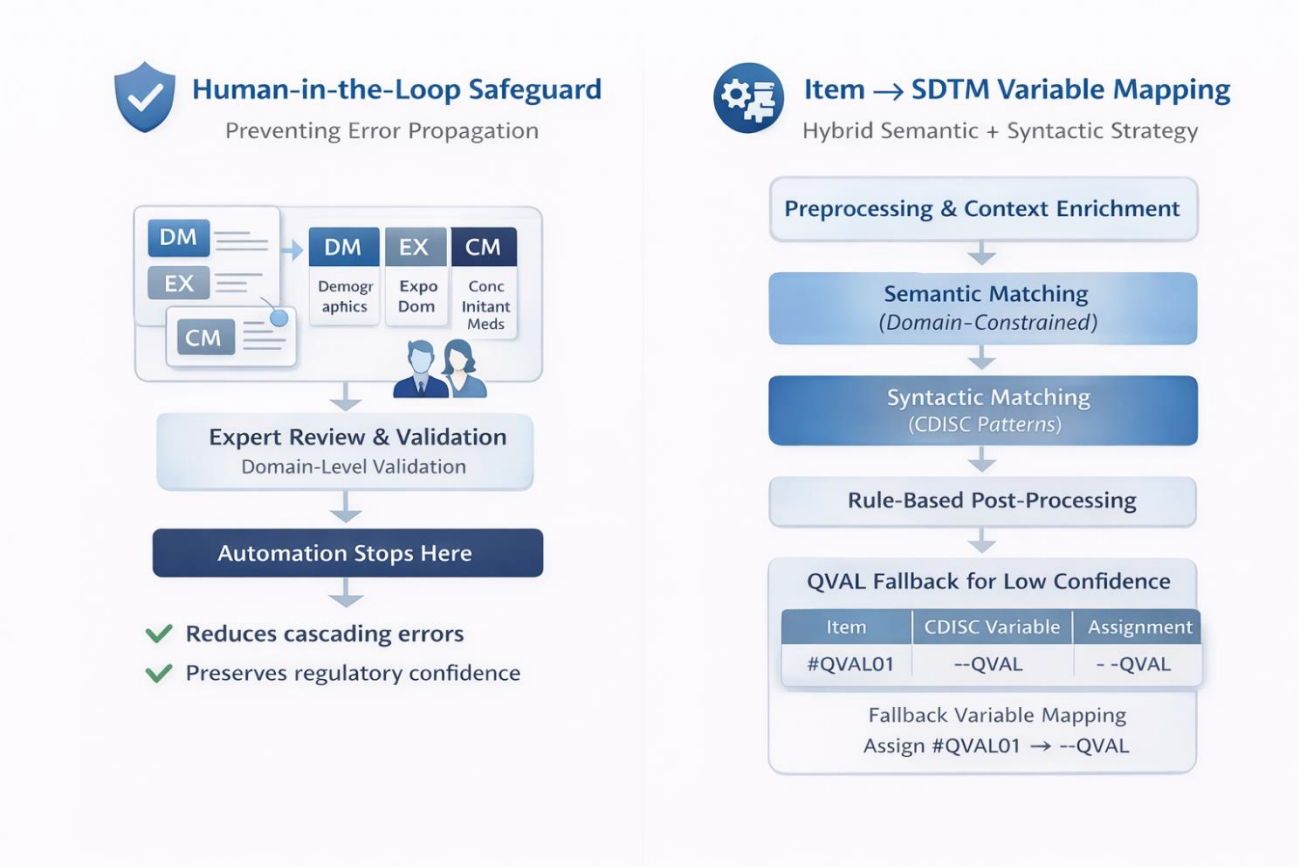

Fig. 2 Methodology - Item to variable Mapping

7.2 Hybrid Semantic and Syntactic Matching

Item-to-variable mapping is implemented using a hybrid framework that combines semantic similarity modelling with deterministic syntactic rule evaluation. The two approaches are executed independently and subsequently reconciled through a defined prioritization strategy.

Semantic matching is performed first and serves as the primary mechanism for capturing conceptual similarity between eCRF items and SDTM variables. For each eCRF item, candidate SDTM variables are restricted to the item's pre-assigned predicted_sdtm_domain, obtained from the upstream form-to-domain classification step. Dual sentence embedding [1] [2] models are used in an ensemble configuration to encode enriched eCRF question text as well as SDTM variable labels and associated CDISC notes. Domain-specific weighting schemes are applied to combine embeddings, with increased emphasis on variable labels for findings domains such as Laboratory Tests (LB), Vital Signs (VS), and ECG (EG). Cosine similarity scores are computed between item and variable embeddings, and calibrated thresholds are used to determine

mapping outcomes, including single-variable assignments, multi-variable candidates, or fallback to supplemental qualifiers (QVAL) for low-confidence matches.

Syntactic matching is executed in parallel to identify high-precision mappings based on CDISC naming conventions and structural patterns. This includes detection of form-level prefixes, normalization of item identifiers, suffix-based inference for date and time variables (e.g., STDAT/STTIM → STDTC), and direct variable name matches. Additional deterministic rules handle special cases such as "not done" indicators and performance flags.

When a syntactic match is identified, it is explicitly prioritized over semantic predictions, as such matches represent deterministic alignments with SDTM conventions and exhibit lower ambiguity. Finally, outputs from both approaches are consolidated and refined through domain-specific post-processing rules to generate the final SDTM variable assignments.

7.3 Rule-Based Post-Processing

Following semantic and syntactic matching, a rule-based post-processing layer is applied to enforce CDISC compliance and refine final item-to-variable assignments. This step encodes domain-driven business rules that capture mandatory co-occurrence and inference relationships defined in the SDTM standard. Co-occurrence rules ensure complementary variables are assigned together, such as pairing DOSE with DOSTXT when either is detected. Inference rules derive required variables from semantically related ones, including automatic addition of TESTCD when TEST is predicted and insertion of TEST and TESTCD when ORRES is identified. Additional rules infer STAT from REASND and enforce numeric–character result pairings (STRESN ↔ STRESC) where applicable. For low-confidence semantic matches, supplemental qualifiers (QVAL) are retained to preserve data without forcing incorrect specificity. By applying these deterministic refinements, the post-processing layer improves regulatory consistency, reduces downstream rework, and ensures the practical usability of automated mappings.

8. EVALUATION AND RESULTS

The solution was evaluated using 10 real-world clinical studies annotated by domain experts, covering both form-level and item-level mappings. Evaluation was conducted under multiple scenarios to assess the impact of domain assignment accuracy on downstream variable mapping performance.

When expert-annotated SDTM domains were provided as input to the item-level mapping pipeline, the approach achieved approximately 90% accuracy in identifying correct SDTM variable assignments. When predicted domains generated by the upstream form-to-domain classification algorithm were used instead, the accuracy of valuable item-level mappings was approximately 75%. This reduction reflects error propagation, as an incorrect domain

assignment constrains the candidate variable search space and causes all subsequent variable mappings for that item to be incorrect.

Independently, the standalone form-to-domain mapping algorithm achieved an accuracy of approximately 80% across the evaluated forms. In addition to accuracy gains, the end-to-end solution automated over 80% of routine item-level mappings, resulting in a 60–70% reduction in manual mapping effort and improved consistency across studies and programmers. Importantly, the framework produces explainable outputs, allowing users to trace mapping decisions and understand the rationale behind suggested assignments.

9. OPERATIONAL IMPACT

The automated mapping framework significantly accelerates study build and SDTM conversion timelines by reducing the reliance on manual, repetitive mapping activities. By improving the accuracy and consistency of initial mappings, the solution minimizes downstream rework during validation and regulatory review cycles. Its scalable design supports consistent standardization across multiple studies, therapeutic areas, and form designs, thereby enhancing overall regulatory readiness. Rather than replacing human expertise, the framework enables SDTM programmers to focus on exception handling, complex mappings, and quality oversight, resulting in more efficient and reliable clinical data standardization workflows.

10. FUTURE ENHANCEMENTS

Future work will focus on expanding annotated validation datasets across a broader range of SDTM domains to improve robustness and generalizability. User feedback and manual corrections will be leveraged to enable continuous refinement of semantic thresholds and rule logic. The framework will be extended to support ADaM and other downstream data standards, enabling end-to-end clinical data standardization. Additional evaluation metrics and improved confidence calibration will be introduced to better quantify mapping reliability and guide human review. These enhancements aim to further strengthen accuracy, scalability, and regulatory alignment.

11. CONCLUSION

This paper presents a practical, production-ready framework for automating eCRF-to-SDTM mapping at both form and item levels. By integrating semantic similarity modelling, deterministic syntactic rules, and controlled application of AI techniques, the approach significantly reduces manual mapping effort while preserving transparency and alignment with CDISC and regulatory expectations.

The proposed hybrid framework demonstrates that intelligent automation, when combined with domain knowledge and human oversight, can scale clinical data standardization without

compromising accuracy or interpretability. This work represents a meaningful step toward more efficient, reliable, and scalable clinical data transformation processes in modern clinical trials.

REFERENCES

1. Nussbaum, Z., Morris, J.X., Duderstadt, B. and Mulyar, A., 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.
2. Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D. and Nie, J.Y., 2024, July. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval* (pp. 641-649).
3. Yang, A. et al. (2024) *Qwen2.5 Technical Report*, arXiv:2412.15115

ACKNOWLEDGEMENTS

RECOMMENDED READINGS

https://huggingface.co/nomic-ai/nomic-embed-text-v1

https://huggingface.co/BAAI/bge-base-en-v1.5

https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-GPTQ-Int4

https://www.cdisc.org/standards/foundational/sdtm

https://www.cdisc.org/standards/foundational/sdtmig/sdtmig-v3-4

https://www.cdisc.org/standards/foundational/adam

CONTACT INFORMATION

Kedar Deshpande
Lifio
Bangalore, India

Phone: 9845242861
Email: kedar.deshpande@lifio.ai

Saransh Khandelwal
Lifio
Bangalore, India
Email: saransh.khandelwal@lifio.ai