Paper DH04 – PHUSE US Connect 2026

Enhancing Clinical Trial Dataset Mapping with AI/ML Techniques: An Approach for CDISC Variable Prediction

Sohan Lal, Geninvo, India
Parveen Kumar, Geninvo, India

ABSTRACT

In clinical trials, the mapping of datasets (ADAM, SDTM) and CDSIC variables to standardized terminologies, such as those defined by CDISC (Clinical Data Interchange Standards Consortium), is essential but often time-consuming and prone to human error. The application of AI/ML techniques offers a promising solution to automate and optimize this mapping process. Machine learning models, including Random Forest and other advanced algorithms, can be trained to predict domain names and map variable names to the correct standards by identifying patterns in historical clinical trial data.
These AI/ML models improve both the speed and accuracy of dataset mapping by analyzing features like variable names, types, and their contextual relationships within clinical trial datasets. By leveraging past clinical trial data, machine learning techniques provide a powerful mechanism for consistent, automated variable mapping, thereby reducing manual intervention and potential errors.
This approach significantly enhances the efficiency of clinical data standardization, ensuring higher consistency in regulatory submissions and contributing to faster, more reliable outcomes in clinical research.

INTRODUCTION

Clinical trials are the cornerstone of modern medical research, providing the structured framework necessary to evaluate new pharmaceuticals, medical devices, and therapeutic interventions. An essential component of any clinical trial is the systematic handling, transformation, and submission of data in formats that meet regulatory expectations set by agencies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA). Among the globally recognized standards for organizing clinical trial data is the Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM), which defines how study data should be structured for regulatory submission.

Despite the structured guidance offered by SDTM, converting raw clinical data into compliant SDTM domains remains a largely manual and resource-intensive process. Source datasets frequently contain inconsistent variable names, unclear labels, heterogeneous structures, and varying data types across studies. This lack of uniformity requires substantial subject-matter expertise and repeated manual review, increasing operational costs, extending development timelines, and raising the risk of inconsistencies or validation issues during regulatory review.
The rapid growth of digital health information and the globalization of clinical research have further emphasized the necessity of standardized data frameworks in biomedical research. Without harmonized formats, integrating or comparing datasets across trials, organizations, or geographic regions becomes complex and error prone. Historically, sponsors submitted customized datasets tailored to individual studies, creating significant challenges for both regulatory reviewers and sponsors. To address these inefficiencies, industry stakeholders established standards organizations to promote harmonization in data collection and reporting practices.

In 1997, pharmaceutical companies, academic institutions, and regulatory representatives formed the Clinical Data Interchange Standards Consortium (CDISC) to develop interoperable standards that support the full lifecycle of clinical research. CDISC's mission is to streamline the acquisition, exchange, submission, and long-term archiving of clinical trial data through standardized structures, thereby accelerating product development and regulatory review. Today, CDISC standards are widely adopted and, in many cases, required by major regulatory agencies such as the FDA, Health Canada, as well as by numerous sponsors worldwide.

This report offers a comprehensive examination of CDISC standards, illustrating their structure, implementation, and impact. It begins with a historical overview explaining the need for data standardization in clinical research. The report then describes the core suite of CDISC standards, including foundational standards such as SDTM and the Analysis Data Model (ADaM), data exchange standards like Define-XML, Therapeutic Area (TA) standards, controlled terminology, and additional components such as CDASH. Practical implementation aspects are discussed, covering the progression from study design and Case Report Form (CRF) development to SDTM mapping and the derivation of analysis datasets. Common software tools, programming environments, and metadata management systems are also highlighted.

Background: SDTM and Variable Mapping Challenges

The Clinical Data Interchange Standards Consortium (CDISC) is a global organization dedicated to developing standards that streamline clinical research data collection and reporting. SDTM is one of the most widely adopted CDISC standards and is essential for regulatory submissions. It organizes clinical data into predefined domains such as:

DM – Demographics
AE – Adverse Events
LB – Laboratory Tests
VS – Vital Signs
EX – Exposure
MH – Medical History

Each domain has defined variable structures, controlled terminology, and ordering sequences. This standardization enables cross-study comparisons and improves regulatory review efficiency.

Challenges in Traditional SDTM Mapping
Despite its benefits, SDTM adoption introduces several operational challenges:

- Inconsistent naming conventions in raw datasets
- Ambiguous or missing variable labels
- Context-dependent domain assignments
- Time-consuming manual reviews
- Dependence on expert availability
- Scalability issues across portfolios
- Increased risk of regulatory validation errors

These challenges motivate the need for intelligent, automated, and scalable solutions.

Literature Review and Related Work

Recent academic and industry research has explored AI-based automation for SDTM mapping and clinical data transformation. Studies on supervised machine learning approaches have demonstrated that predictive models trained on historical mappings can significantly reduce manual effort. Neural network-based embedding techniques, such as Siamese networks and transformer models, have been applied to classify metadata fields into SDTM domains with promising results.

Industry white papers and posters emphasize hybrid approaches that combine automation with human validation to maintain compliance. Broader AI applications in clinical trials—including predictive analytics, protocol optimization, and document automation—demonstrate the growing role of intelligent systems in accelerating research workflows. Collectively, these works support the feasibility of AI-driven SDTM automation while highlighting the importance of regulatory oversight and transparency

Artificial Intelligence and Machine Learning

Artificial Intelligence (AI) and Machine Learning (ML) are revolutionizing clinical trials by providing tools that can analyze large volumes of data, identify hidden patterns, and predict outcomes. Python, with its rich ecosystem of libraries such as TensorFlow, Keras, scikit-learn, and PyTorch, is well-equipped to handle the complexities of AI and ML in clinical research.

The proposed framework introduces a hybrid architecture that merges structural and semantic intelligence. The goal is not to replace experts but to augment their decision-making capabilities by providing ranked mapping suggestions.

Core Principles:

1. Learning from Historical Data
2. Data-Type–Aware Structural Modeling
3. Semantic Contextual Understanding
4. Hybrid Probability Aggregation
5. Human-in-the-Loop Validation

1. SDTM data domain prediction

Python's robust machine learning ecosystem allowed us to implement models for predictive analysis. Python's machine learning libraries allow researchers to build predictive models that can forecast the likelihood of success or failure for a treatment, based on factors such as patient demographics, medical history, and different domains data (AE, LB, VS etc.)

For example, we used the **Random Forest** algorithm, **NLTK**, **re** and **word2Vec**, a powerful ensemble learning method, to predict and automate certain data mappings. Used Python's **scikit-learn** library to develop, train, and evaluate the Random Forest model, which allowed us to make data-driven predictions based on the patterns observed in the dataset.

To predict the best possible **SDTM variable mappings**, we have developed four specialized machine learning models:

- **Character Model**
- **Label Model**
- **Datetime Model**
- **Numeric Model**

Each of these models is trained on historical clinical trial data, and work by analysing the data and providing the most probable mappings based on statistical probabilities. This approach significantly reduces manual intervention and enhances the consistency and accuracy of the mappings.

The models predict the appropriate method based on user selections and variable patterns. For example, if a variable contains "DAT," "DT," or "DTC," we associate it with a date-related method. Once the final dataset is created, we apply SDTM terminology to ensure compliance with standard definitions. Additionally, we can arrange variables in the correct order as per the sequence defined in the Implementation Guide (IG). We also evaluate that the dataset includes all required and expected variables to meet SDTM standards. Our predictive models—Character Model, Label Model, Datetime Model, and Numeric Model—analyze the data and determine the best possible SDTM variable mappings based on probability. These models help automate the mapping process, improving accuracy and efficiency.

Once the model completes the mapping predictions, **Python** handles storage tasks, ensuring that SDTM terminology is consistently applied throughout the dataset. Also include logic to flag discrepancies in the mapped values, particularly when a value does not conform to SDTM standards. In such cases, a list of discrepancies is generated, allowing users to review and select the correct mappings, improving overall accuracy.

Methods for Predicting Clinical Domain Names

To automate the classification of clinical datasets into appropriate CDISC domains, we implemented a hybrid machine learning framework combining similarity-based and supervised learning approaches. The objective was to predict the correct domain label (e.g., AE, DM, LB, VS) based on the structure and content of dataset column names and values.

Feature Engineering

The prediction process began with structured feature extraction from raw datasets. For each dataset, we derived features from:

- Column names (e.g., AETERM, USUBJID, LBTESTCD)
- Column labels and metadata
- Data type patterns (numeric, character, date)
- Value distributions (e.g., presence of lab test codes, adverse event terms)
- Frequency of controlled terminology matches

Text-based features were vectorized using TF-IDF representations to capture semantic similarity between dataset variables and known domain templates. These features formed the input for downstream models.

1. **Cosine Similarity Approach**

Cosine similarity was used as a baseline similarity-matching method. We constructed vector representations of each dataset based on variable names and value tokens. These vectors were compared against pre-defined domain reference vectors derived from standard SDTM domain specifications.
The cosine similarity score measures the angular similarity between two vectors:

$$\text{Similarity} = \frac{A \cdot B}{||\ A\ ||\ \ ||\ B\ ||}$$

The domain with the highest similarity score was selected as the predicted class.
How it helped:
Effective for capturing lexical overlaps (e.g., variables starting with AE aligning with AE domain).
Robust when naming conventions closely follow CDISC patterns.
Fast and interpretable baseline method.
However, it relies heavily on textual similarity and may underperform when variable naming is inconsistent.

2. **Random Forest**

Random Forest, an ensemble tree-based classifier, was trained on engineered features including term frequencies, presence of key identifiers (e.g., USUBJID, VISIT), and statistical summaries of column values.

Random Forest builds multiple decision trees and aggregates predictions via majority voting. It handles nonlinear feature interactions and high-dimensional data effectively.

How it helped:

- Captured complex patterns beyond simple name similarity.
- Identified domain-specific combinations of variables (e.g., AETERM + AESEV strongly indicating AE domain).
- Reduced overfitting via ensemble averaging.
- Provided feature importance metrics for interpretability.

This model performed well when datasets contained mixed structured and semi-structured features.

3. **Support Vector Machine (SVM)**

SVM was used as a high-dimensional classifier capable of separating domains using optimal decision boundaries. Using TF-IDF text features and numerical metadata features, the SVM model learned hyperplanes that maximized class separation.

With appropriate kernels (e.g., linear or RBF), SVM effectively handled sparse textual features common in clinical metadata.

How it helped:

- Strong performance on high-dimensional sparse text features.
- Effective when domain distinctions depended on subtle textual differences.
- Reduced misclassification between closely related domains.

SVM proved particularly useful when domain differentiation required boundary-based discrimination rather than rule-based separation.

4. **Multinomial Naïve Bayes**

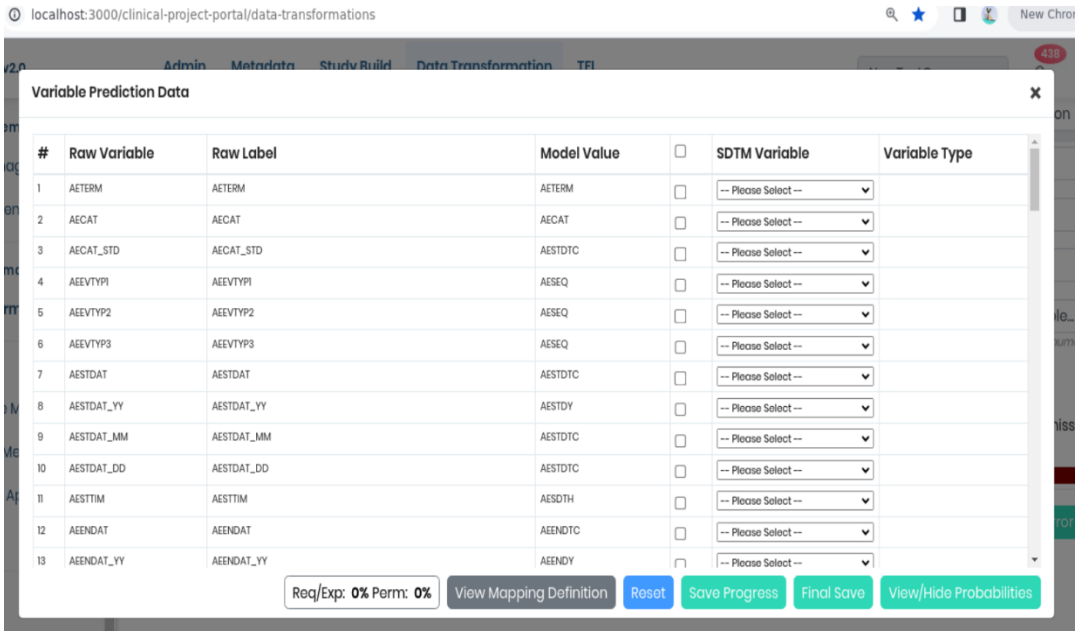
Multinomial Naïve Bayes (MNB) was applied primarily to text-based features derived from variable names and value tokens. This probabilistic classifier assumes conditional independence between features and estimates class probabilities based on word frequencies.

How it helped:

- Efficient for text-heavy metadata.
- Worked well when domain-specific terminology strongly characterized the dataset.
- Computationally lightweight and scalable.

Although simpler than tree-based models, MNB provided competitive baseline performance and strong interpretability.

Variable Prediction and SDTM Mapping Interface



The screenshot illustrates the **Variable Prediction Data** interface that supports semi-automated mapping of raw clinical dataset variables to their corresponding SDTM variables.

Overview of the Interface

The table presents a structured comparison between raw dataset metadata and machine learning model predictions. Each row represents a raw variable from the source dataset and contains the following key fields:

- **Raw Variable** – The original variable name from the raw dataset (e.g., AETERM, AESTDAT).
- **Raw Label** – The descriptive label associated with the variable.
- **Model Value** – The predicted SDTM variable generated by the domain classification model.
- **Selection Checkbox** – Allows the user to confirm or exclude model predictions.
- **SDTM Variable** – A dropdown menu enabling manual confirmation or correction of the predicted SDTM variable.
- **Variable Type** – Indicates classification or metadata related to the variable.

This interface integrates automated domain prediction with human validation, enabling a controlled and auditable mapping workflow.

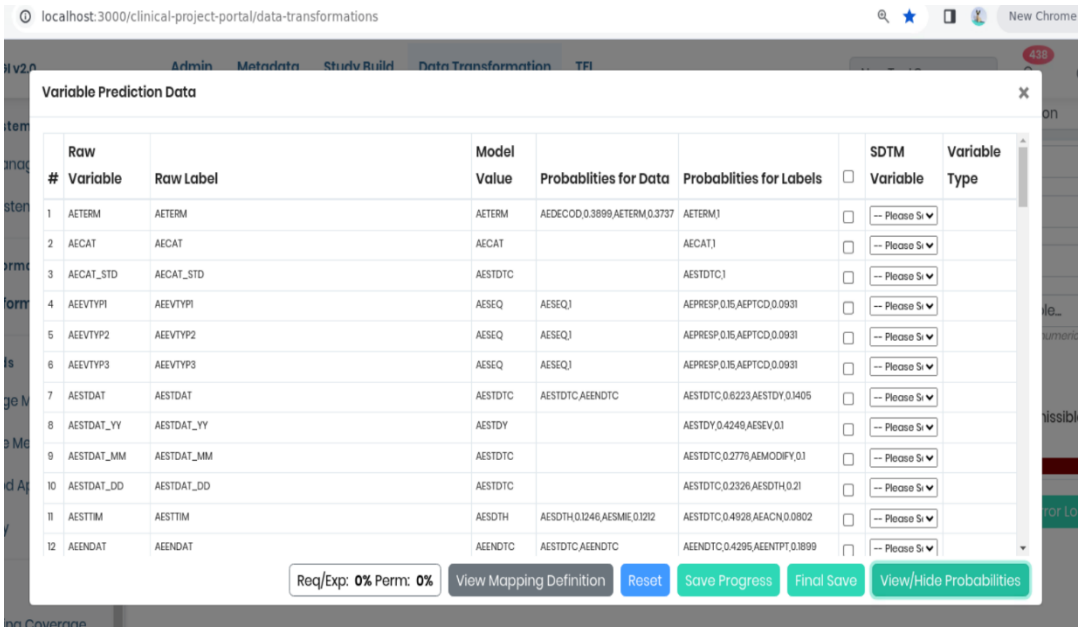
Role of Machine Learning Predictions

The Model Value column reflects predictions generated using machine learning techniques

- Variable naming patterns (e.g., AE-prefixed variables mapping to the AE domain),
- Text similarity between raw labels and SDTM terminology,
- Structural features (e.g., presence of date components like _YY, _MM, _DD),
- Statistical and categorical value patterns.

For example:

- **AETERM** is predicted as AETERM in SDTM,
- Date-related components such as **AESTDAT_YY/MM/DD** are mapped toward ISO 8601 SDTM datetime structures (AESTDTC),
- Event severity variables (e.g., AEVTYP) align with expected SDTM AE domain attributes.



Human-in-the-Loop Validation

Although model predictions provide an initial mapping suggestion, final control remains with the user. The dropdown under **SDTM Variable** allows users to:

- Accept the predicted mapping,
- Override the suggestion,
- Select an alternative SDTM variable if needed.

This hybrid workflow ensures:

- Improved efficiency through automation,
- Reduced manual mapping effort,
- Regulatory compliance through user validation,
- Traceability of mapping decisions.

Role of Probability Scores

Unlike a simple prediction output, this interface displays probability distributions across potential SDTM variables, reflecting model confidence levels.

Two probability dimensions are shown:

1. **Probabilities for Data**
   - Derived from analysis of column values, patterns, and structural characteristics.
   - For example, date components or ISO-formatted values may increase the probability of mapping to --DTC variables**.**
2. **Probabilities for Labels**
   - Based on textual similarity between raw labels and SDTM terminology**.**
   - Uses techniques such as TF-IDF vectorization and cosine similarity to evaluate semantic alignment.

For instance:

- A variable such as AESTDAT may show high probability for AESTDTC, indicating date conversion mapping.
- Variables like AEVTYP1/2/3 may display strong alignment with AESEQ or related AE domain attributes based on naming patterns and value structure.

Displaying probability scores improves transparency and trust in automated predictions

REFERENCES

- **CDISC Standards: How They Work with SDTM & ADaM Examples | IntuitionLabs**
- **https://web.pointcrosslifesciences.com/sdtm-automation/**
- **SDTM Automation : Streamlining Clinical Trial Data**

ACKNOWLEDGMENTS

We would like to express our heartfelt gratitude to Geninvo Technology Pvt Ltd and PHUSE for providing the opportunity to contribute to this paper, as well as for their continuous support and encouragement throughout the research process. We also extend our sincere appreciation to all those who contributed to the completion of this paper, with special thanks to our colleagues for their valuable feedback and support.

CONTACT INFORMATION

Your comments and questions are valued and encouraged

Contact the author at: **sohan.l225@geninvo.com**
Author Name: Sohan Lal
Company: Geninvo technologies pvt. ltd
Address: 1408 East Empire Street 61701
Bloomington (Illinois), United States.
Work Phone: +1 309-807-5006
Email: sohan.l225@geninvo.com
Website: GENINVO is the go-to partner for life science industry.

Contact the co-author at: **parveen.k12@geninvo.com**
Author Name: Parveen Kumar
Company: Geninvo technologies pvt. ltd
Address: 1408 East Empire Street 61701
Bloomington (Illinois), United States.
Work Phone: +1 309-807-5006
Email: parveen.k12@geninvo.com
Website: GENINVO is the go-to partner for life science industry.