



DH04

Enhancing Clinical Trial Dataset Mapping with AI/ML Techniques:

Techniques: An Approach for CDISC Variable Prediction

Parveen Kumar

March 2026

www.geninvo.com

North America | UK | Europe | India



AGENDA



www.geninvo.com
North America | UK | Europe | India

01

Introduction to SDTM

02

Variable Mapping Challenge

03

Proposed AI/ML-Based Solution

04

Data-Type Aware Modeling

05

Global Label-Based Model

06

Benefits and Future Direction

07

Q&A



INTRODUCTION TO SDTM



www.geninvo.com
North America | UK | Europe | India

SDTM is a global CDISC standard for organizing clinical trial data, designed specifically for regulatory review and cross-study consistency. But SDTM is more than just a format—it's a structured language for clinical data that enables seamless communication across the pharmaceutical industry.

The challenge isn't the lack of standards. It's the lack of intelligent automation to implement them efficiently and accurately.



VARIABLE MAPPING CHALLENGE



www.geninvo.com
North America | UK | Europe | India

The Problem

Manual SDTM mapping is complex, time-consuming, and error-prone, bottlenecks in clinical trial data submission.

Key Challenges

- High variability in raw data naming conventions
- Context-dependent domain assignment
- Heavy reliance on expert knowledge
- Time-consuming manual review processes
- Risk of inconsistency and human error
- Scalability limitations across multiple studies
- Potential regulatory delays due to validation issues



PROPOSED AI/ML-BASED SOLUTION



www.geninvo.com
North America | UK | Europe | India

MULTI-STUDY LEARNING FRAMEWORK

It leverages historical SDTM mappings collected from multiple studies to build intelligent prediction models. Each source field is comprehensively characterized to capture the full context of expert mapping decisions.

Field Name

Variable identifier from
source data

Field Description

Semantic label
providing context

Data Type

Character, numeric, or
datetime classification

Training is built on real-world expert decisions, ensuring the AI learns from proven mapping patterns and domain expertise.

AI APPROACHES USED TO IDENTIFY DATASET DOMAINS



1. Cosine Similarity (Name Matching)

- Compares dataset variable names with known domain patterns.
- Measures how similar the words are.
- Example: Variables like AEDECOD, AESTDTC → likely AE domain.

2. Support Vector Machine (SVM)

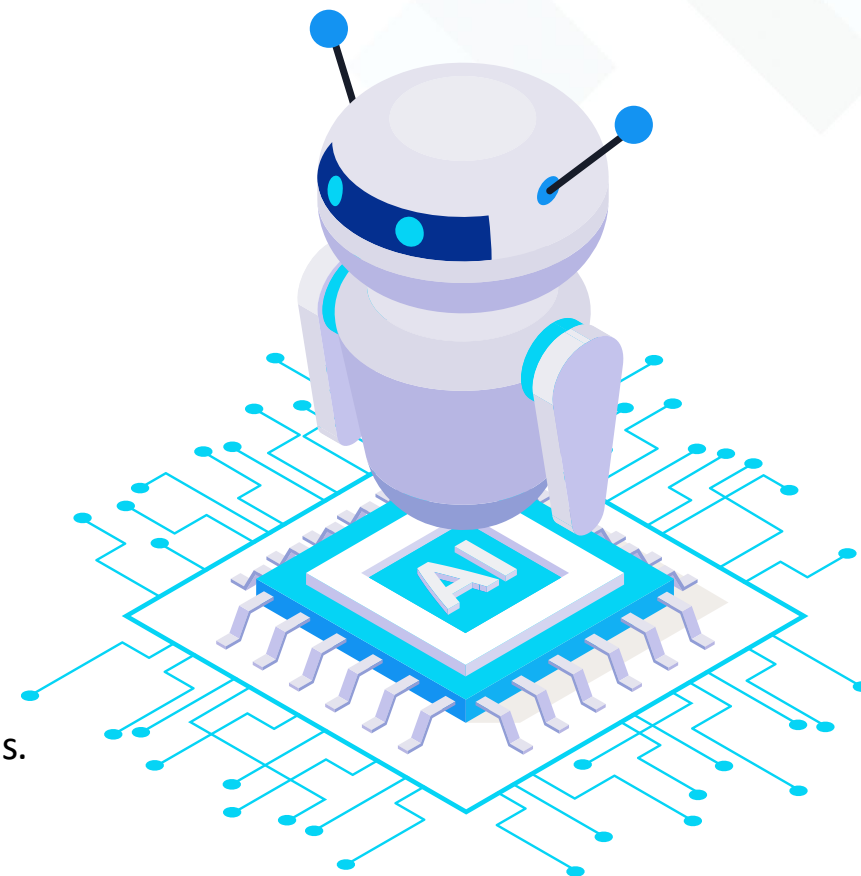
- Learns to separate domains based on patterns in text and data
- Good at detecting small differences between similar domains.

3. Random Forest (Pattern Learning)

- Looks at multiple characteristics of the dataset.
- Uses many small decision trees to detect patterns in variable combinations.

4. Multinomial Naïve Bayes

- Focuses on word frequency in variable names
- Predicts the domain based on how often certain words appear



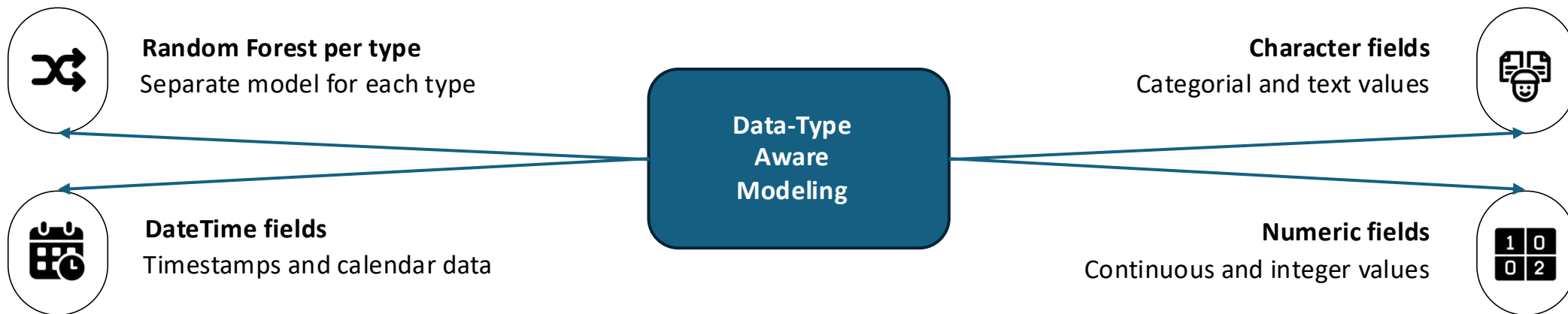
DATA-TYPE AWARE MODELING

Specialized Models by Data Type

Source fields are first categorized into character, numeric, or datetime types. For each SDTM domain, separate Random Forest models are trained per datatype, enabling specialized prediction logic.

- Character_Model
- Numeric_Model
- DateTime_Model

This approach captures structural patterns unique to each data type, improving prediction accuracy.



GLOBAL LABEL-BASED MODEL



www.geninvo.com
North America | UK | Europe | India

All field descriptions are used to train a centralized Label Model that captures descriptive meaning independent of data type. This semantic intelligence handles ambiguous or poorly named fields by understanding the underlying clinical concept.

Real-World Example



Field Name:
VAR01 (uninformative)



Label: "Hemoglobin
Concentration"



Result: Correctly predicted
under Laboratory domain

Label intelligence strengthens contextual understanding, ensuring accurate mapping even when field names are cryptic or inconsistent.

HYBRID PREDICTION STRATEGY



Data Models

Generate datatype-aware predictions



Label Model

Generate semantic predictions



Combine Scores

Aggregate confidence rankings



Top 3 Suggestions

Highest probability auto-selected

This hybrid approach combines structural intelligence with semantic understanding, delivering robust predictions that leverage both naming patterns and contextual meaning

PRACTICAL EXAMPLE



Admin

Metadata

Study Build

Data Transformation

TEI

438

Variable Prediction Data

X

#	Raw Variable	Raw Label	Model Value	Probabilities for Data	Probabilities for Labels	☐	SDTM Variable	Variable Type
1	AETERM	AETERM	AETERM	AEDECOD,0.3899,AETERM,0.3737	AETERM,1	☐	-- Please S...	
2	AECAT	AECAT	AECAT		AECAT,1	☐	-- Please S...	
3	AECAT_STD	AECAT_STD	AESTDTC		AESTDTC,1	☐	-- Please S...	
4	AEVTYPE1	AEVTYPE1	AESEQ	AESEQ,1	AEPRESP,0.15,AEPTCD,0.0931	☐	-- Please S...	
5	AEVTYPE2	AEVTYPE2	AESEQ	AESEQ,1	AEPRESP,0.15,AEPTCD,0.0931	☐	-- Please S...	
6	AEVTYPE3	AEVTYPE3	AESEQ	AESEQ,1	AEPRESP,0.15,AEPTCD,0.0931	☐	-- Please S...	
7	AESTDAT	AESTDAT	AESTDTC	AESTDTC,AEENDTC	AESTDTC,0.6223,AESTDY,0.1405	☐	-- Please S...	
8	AESTDAT_YY	AESTDAT_YY	AESTDY		AESTDY,0.4249,AESEV,0,1	☐	-- Please S...	
9	AESTDAT_MM	AESTDAT_MM	AESTDTC		AESTDTC,0.2776,AE MODIFY,0,1	☐	-- Please S...	
10	AESTDAT_DD	AESTDAT_DD	AESTDTC		AESTDTC,0.2326,AESDTH,0.21	☐	-- Please S...	
11	AESTTIM	AESTTIM	AESDTH	AESDTH,0.1246,AESMIE,0.1212	AESTDTC,0.4928,AEACN,0.0802	☐	-- Please S...	
12	AEENDAT	AEENDAT	AEENDTC	AESTDTC,AEENDTC	AEENDTC,0.4295,AEENTPT,0.1899	☐	-- Please S...	

Req/Exp: 0% Perm: 0%

View Mapping Definition

Reset

Save Progress

Final Save

View/Hide Probabilities

BENEFITS AND FUTURE DIRECTION



www.geninvo.com
North America | UK | Europe | India



Significant reduction in manual SDTM mapping effort accelerates timelines



AI-powered predictions minimize inconsistencies and mapping mistakes



Streamlined data preparation contributes to faster regulatory submissions



Easily handle multiple studies with consistent quality and efficiency



Expert validation ensures accuracy while maintaining automation benefits

QUESTIONS & ANSWERS



www.geninvo.com
North America | UK | Europe | India





Ideas
Delivered
Intuitively

Thank You!

+1-833-GEN-INVO (436-4686) | info@geninvo.com | www.geninvo.com

