



If you want to deliver standardised clinical data sets quickly, talk to the machine

Chris Austin MBA, MSDS and Michael Phillips PhD
ICON plc



EU Connect 13–16 Nov 2022





Introduction

The Importance and Challenge of Clinical Data Standards



CDASH

Standard way to collect data consistently across studies and sponsors so that data collection formats and structures provide clear traceability of submission data into SDTM, delivering more transparency to regulators and others who conduct data review.

SDTM

Standard for organising and formatting data to streamline processes in collection, management, analysis and reporting. Implementing SDTM supports data aggregation and warehousing; fosters mining and reuse; facilitates sharing; helps perform due diligence and other important data review activities; and improves the regulatory review and approval process.

[cdisc.org](https://www.cdisc.org)

Unfortunately, at a study level,
CDASH and SDTM are not always fully
implemented
Many expert human hours are spent
mapping eCRFs to SDTM

CDISC = Clinical Data Interchange Standards Consortium
CDASH = Clinical Data Acquisition Standards Harmonization
SDTM = Study Data Tabulation Model
eCRF = Electronic Case Report Form

Example Mapping for Adverse Events

eCRF

AE description		3
Start date		4
Stop date		5
Outcome	Not Recovered/Not Resolved	6
	Recovered/Resolved	
	Recovered/Resolved with Sequelae	
	Recovering/Resolving	
	Fatal	
	Unknown	
Severity	Grade 1 / Mild	7
	Grade 2 / Moderate	
	Grade 3 / Severe	
	Grade 4 / Life-threatening	
	Grade 5 / Death	
Relationship with study drug	Unrelated	8
	Probably not related / Remote	
	Possibly related	
	Probably related	
	Definitely related	
Fatal		18
Requires Inpatient Hospitalization or prolongation of existing hospitalization		19
Life Threatening		20
Congenital Anomaly		21
Persistent or significant disability/incapacity		22

Field Annotation

3	AETERM	\$200	AE description
4	AESTDAT	dd- MMM yyyy	Start date
5	AEENDAT	dd- MMM yyyy	Stop date
6	AEOUT	1	Outcome
7	AESV	1	Severity
8	AEREL	1	Relationship with study drug
	AEREL3		
18	AESCRIT1	1	Fatal
19	AESCRIT2	1	Requires Inpatient Hospitalization or prolongation of existing hospitalization
20	AESCRIT3	1	Life Threatening
21	AESCRIT4	1	Congenital Anomaly
22	AESCRIT5	1	Persistent or significant disability/incapacity

SDTM Mapping

	Variable Name	Label
3	AETERM	Reported Term for the Adverse Event
4	AESTDTC	Start Date/Time of Adverse Event
5	AEENDTC	End Date/Time of Adverse Event
6	AEOUT	Outcome of Adverse Event
7	AETOXGR	Standard Toxicity Grade
8	AEREL	Causality
18	AESDTH	Results in Death
19	AESHOSP	Requires or Prolongs Hospitalization
20	AESLIFE	Is Life Threatening
21	AESCONG	Congenital Anomaly or Birth Defect
22	AESDISAB	Persist or Signif Disability/Incapacity

Why not train a machine to do this mapping?



Methodology

Scope

Domains

Adverse Events (AE)

Prior and Concomitant Medications (CM)

Demographics (DM)

Vital Signs (VS)

Objective

Train a machine learning model (ML) to map source field names from eCRF to an SDTM standard

Training Materials

109 Completed Source–Target Mapping Specifications

For inclusion in training, we required a minimum of 10 occurrences of the target variable

Programming Environment

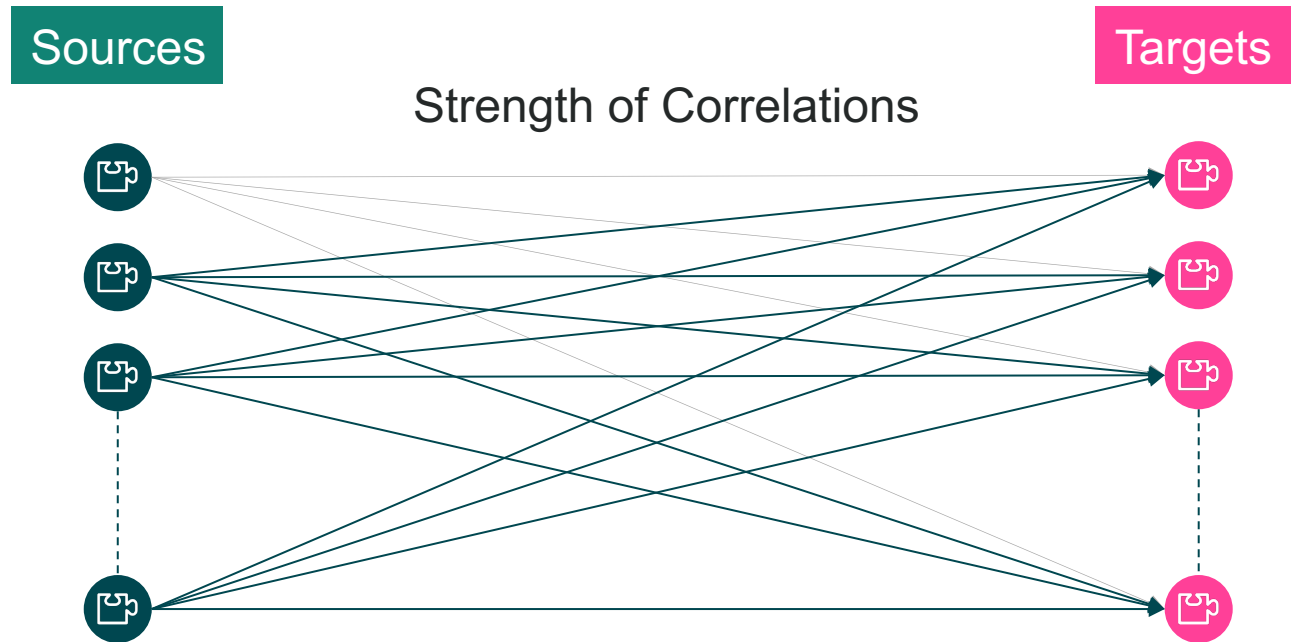
Python (scikit-learn)

Feature Engineering and Modelling

- Source variable names were cleaned of special characters
- Source form and variable name combinations were transformed into 3 features:
 - X.Y
 - X Y plus X.Y
 - Y
- The resulting one-word and two-word pairs were correlated to targets using natural language processing methods
- Four multinomial classification models were used:
 - Random Forest
 - Linear Support Vector
 - Naive Bayes
 - Logistic Regression

Dataset	Variable Name	Source Feature 1	Source Feature 2	Source Feature 3
AE	USUBJID	AE.Subject	AE Subject AE.Subject	Subject
AE	AETERM	AE.AETERM	AE AETERM AE.AETERM	AETERM
AE	AESER	AE.AESER	AE AESER AE.AESER	AESER
AE	AEACN	AE.AEACN4	AE AEACN4 AE.AEACN4	AEACN4
AE	AEREL	AE.AEREL AE.AEREL3	AE AEREL AE.AEREL AE AEREL3 AE.AEREL3	AEREL AEREL3
AE	AEOU	AE.AEOU	AE AEOU AE.AEOU	AEOU
AE	AESCONG	AE.AESCRIT4	AE AESCRIT4 AE.AESCRIT4	AESCRIT4
AE	AESDISAB	AE.AESCRIT5	AE AESCRIT5 AE.AESCRIT5	AESCRIT5
AE	AESDTH	AE.AESCRIT1	AE AESCRIT1 AE.AESCRIT1	AESCRIT1
AE	AESHOSP	AE.AESCRIT2	AE AESCRIT2 AE.AESCRIT2	AESCRIT2
AE	AESLIFE	AE.AESCRIT3	AE AESCRIT3 AE.AESCRIT3	AESCRIT3
AE	AESMIE	AE.AESCRIT6	AE AESCRIT6 AE.AESCRIT6	AESCRIT6
AE	AETOXGR	AE.AESEV	AE AESEV AE.AESEV	AESEV
AE	AESTDTC	AE.AESTDAT	AE AESTDAT AE.AESTDAT	AESTDAT
AE	AEENDTC	AE.AEENDAT	AE AEENDAT AE.AEENDAT	AEENDAT

Machine Learning Models – Match Probability

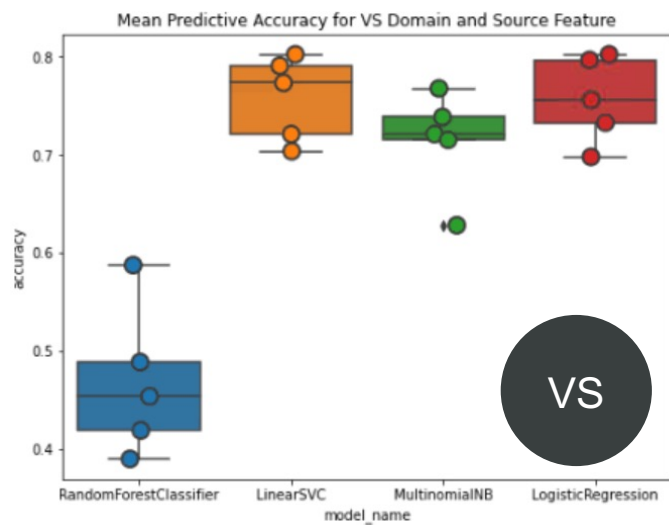
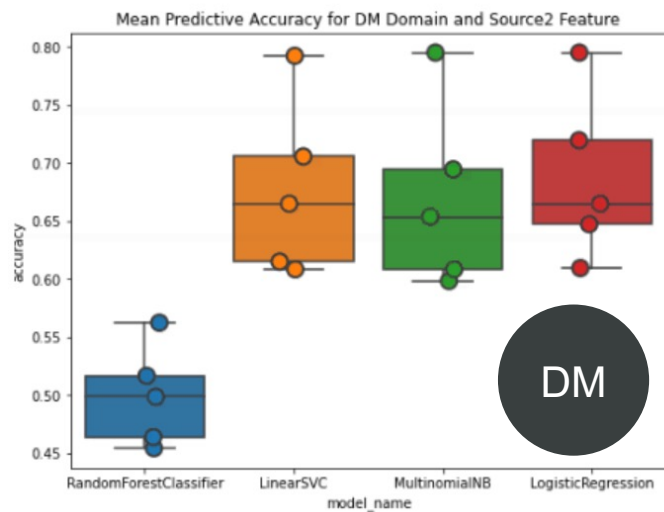
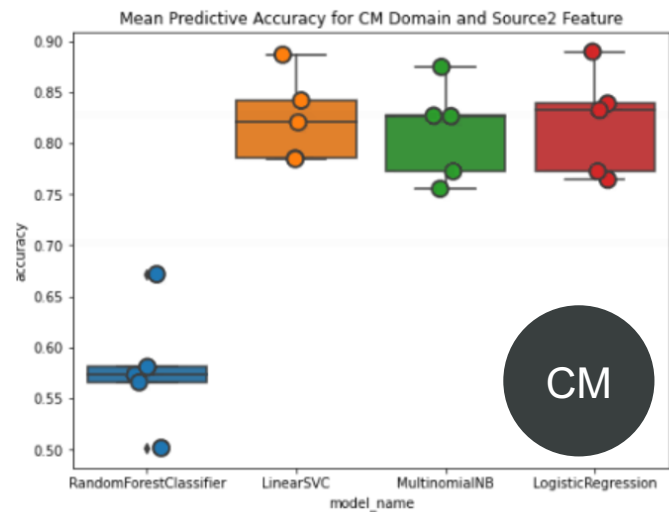
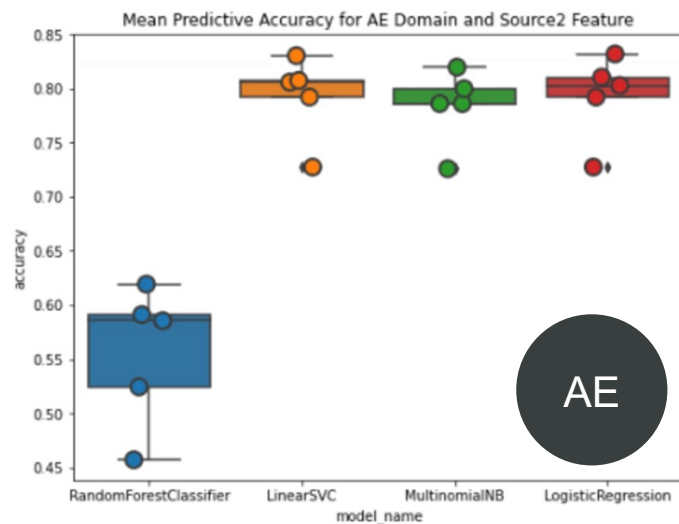


The different models used different approaches to categorise these correlations but they all provided a probability score for each match based on what they learned from the training data



Results

Overall Results



Domain	Overall Model Test Accuracy
AE	80%
CM	80%
DM	65%
VS	75%

*Accuracy = percent of variables correctly mapped

Results When Models Applied to Specific Unseen Study

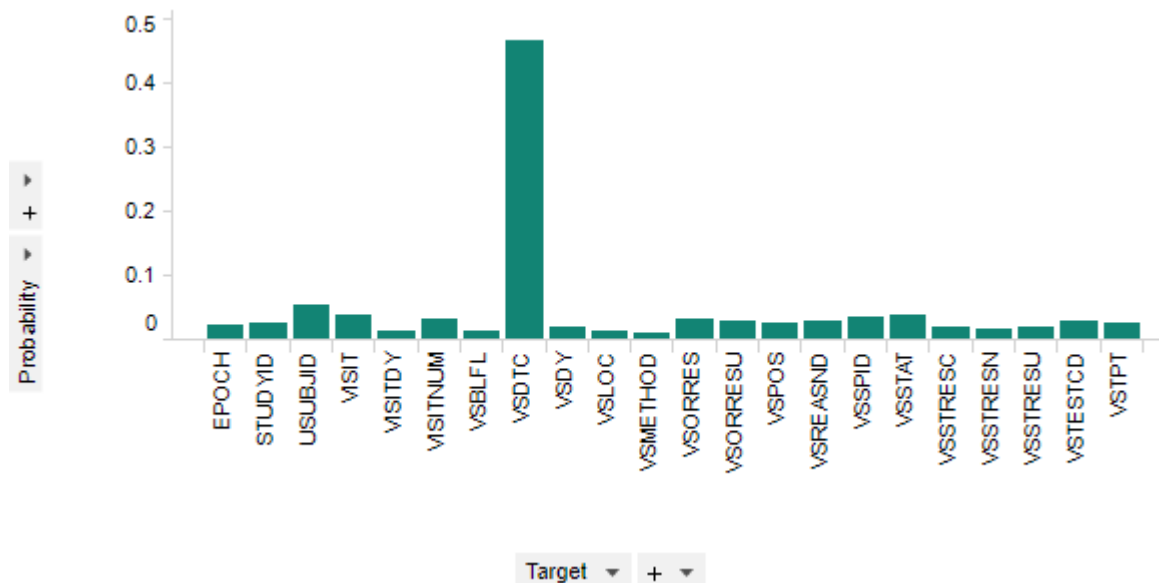
Percent Source to Target Variables Mapped		
Domain	Model Results (General)	Field Test on a Specific (Unseen) Study
AE	80%	30 of 35 mapped (83.3%)
CM	80%	17 of 20 mapped (85%)
DM	65%	15 of 17 mapped (88.2%)
VS	75%	9 out of 10 mapped (90%)

Probability of Variable Match – True Positive Example

VS source =

```
RD_FRMVS.VSDAT  
RD_FRMVS.VSTIM  
RD_FRMVS1.VSDAT1  
RD_FRMVS1.VSTIM1  
RD_FRMVSU.VSDATUNS  
RD_FRMVSU.VSTIMUNS
```

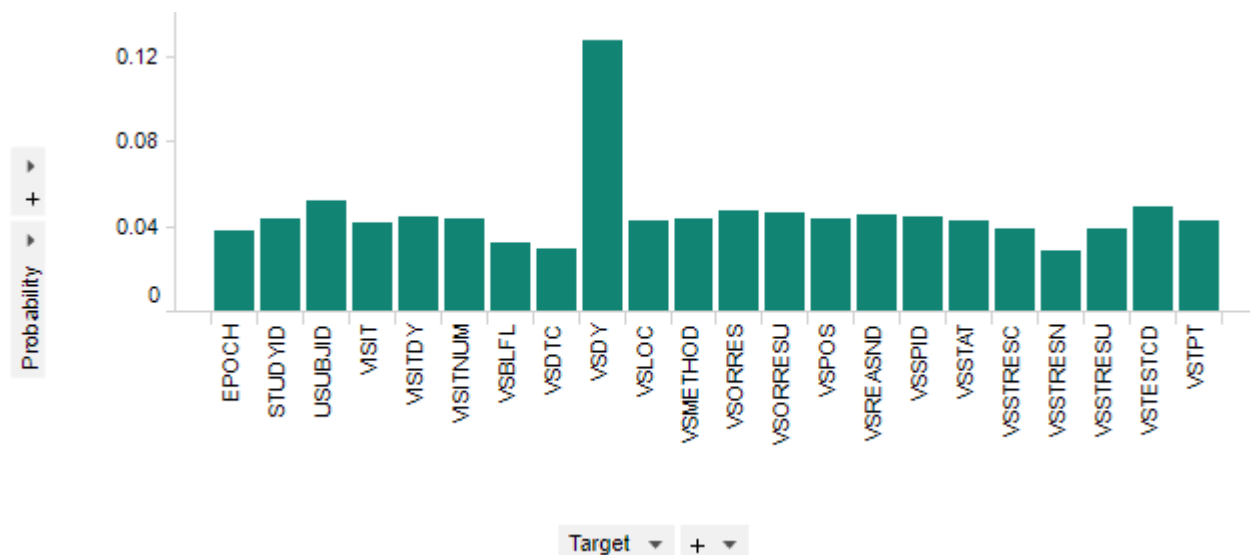
The trained model provides a match probability for each target variable, summing to 1
The clear winner in this case is VSDTC, which also turns out to be correct



Probability of Variable Match – False Positive Example

VS source = DM.RFSTDTC

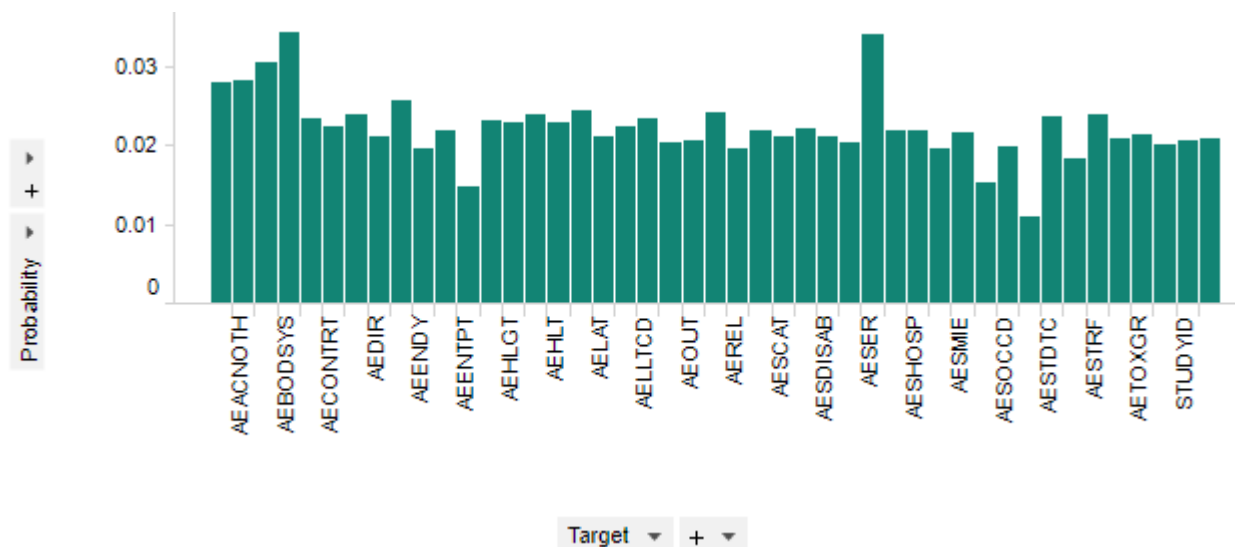
- The winner in this case is VSDY, but the correct answer was VSRFTDTC
- Reason: In the training data, RFSTDTC more often mapped to VSDY, so the model was simply reflecting its training



Probability of Variable Match – Inconclusive Example

AE source = RD_FRMAE.AEPRIOR

There is no clear winner in this case; the highest match is AEBODSYS, but this is not a clear winner
The correct match was AESTRTPT, which didn't even feature in the training dataset
Reason: Insufficient data in the training dataset; the test study was too unusual



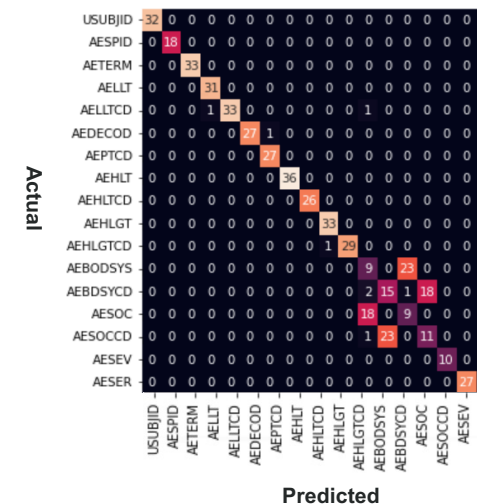
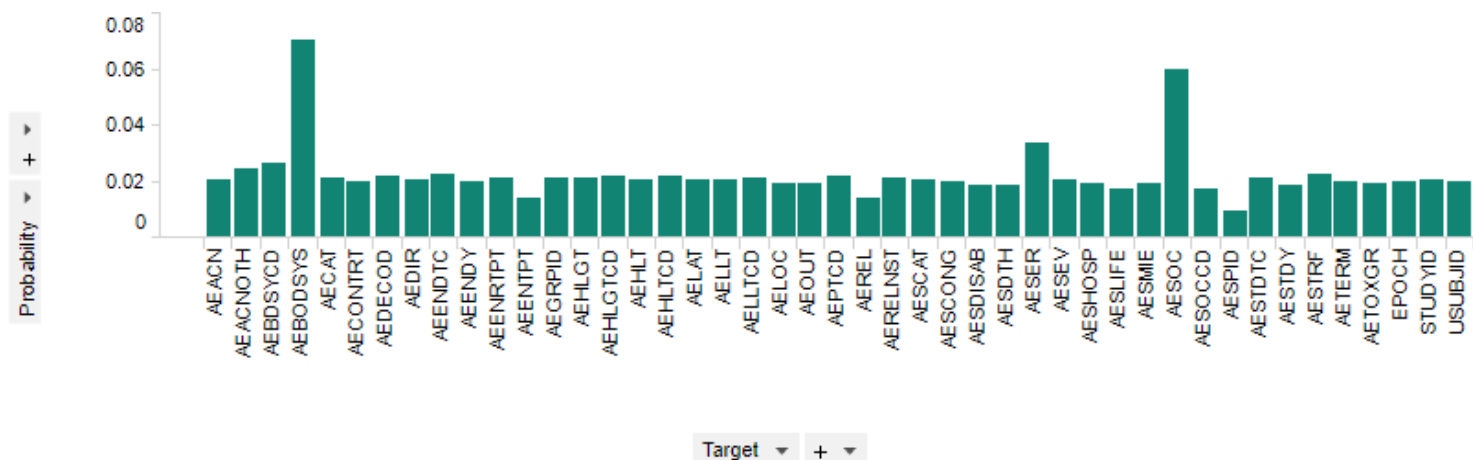
Probability of Variable Match – Many-to-Many Example

AE source = RD_FRMAE.AESOC

This source variable maps to more than one target

Therefore, the training data set has insufficient data to distinguish between the two mappings

Target	Description	Source	Additional Logic
AEBODSYS	Body System or Organ Class	RD_FRMAE.AESOC	Sponsor-Specific Controlled Terminology
AESOC	Primary System Organ Class	RD_FRMAE.AESOC	





Conclusions

Conclusions



ML can deliver pretty decent results for source target mapping for SDTM

If we enrich the training data, we can overcome many of the limitations observed so far

If we develop a clear confidence measure, those who conduct SDTM mapping can easily incorporate into existing processes to accelerate timelines and improve accuracy and precision

What next?



Extend the training data sets to include more studies and more information, such as mapping logic



Combine NLP and ML to handle value mapping and controlled terminology requirements



Replace visual analysis of probability matching with a statistical algorithm that provides an easy-to-process confidence measure for end users

Thank You!

iconplc.com

