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

Chris Austin MBA, MSDS and Michael Phillips PhD

ICON plc

phuse EU Connect 13–16 Nov 2022

Abstract

We trained a model to automatically map source electronic case report form variables to Study Data Tabulation Model standard variables. The application of data collection and data delivery standards to clinical trial data are not always perfectly aligned. We used a combination of machine learning and natural language processing techniques to train a model against four of the core safety domains. Overall, the results were good, with the model accurately mapping 80, 80, 65 and 75 percent of variables for adverse events, concomitant medications, demographics and vital signs, respectively.

Introduction

The Clinical Data Interchange Standards Consortium (CDISC)[1] provides standards for clinical data collection and delivery that are universally adopted across the clinical research industry. According to CDISC, its Clinical Data Acquisition Standards Harmonization (CDASH) provides a 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. The aim of this standard is to deliver more transparency to regulators and others who conduct data review.

CDISC's Study Data Tabulation Model (SDTM) is published as a standard for organising and formatting data to help streamline processes for data collection, management, analysis and reporting. The implementation of 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.

Unfortunately, at a study level, CDASH and SDTM are not always fully implemented and many expert human hours are spent mapping electronic case report forms (eCRFs) to SDTM. Figure 1 shows a typical eCRF, its corresponding field annotation and the SDTM mapping. Nine of the 11 variable names in the eCRF have to be changed, or mapped, for SDTM output.

[1] https://www.cdisc.org/

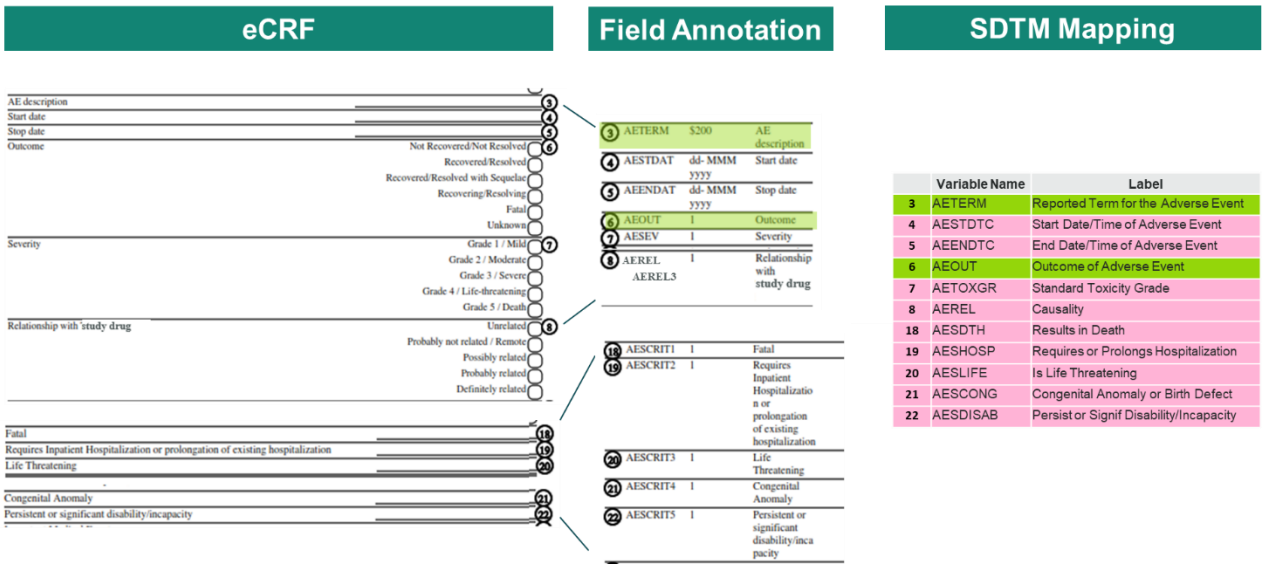

Figure 1: Example SDTM mapping for adverse events

We have completed a proof of concept (POC) to show how machine learning (ML) can be used to partially complete the source–target variable mapping for SDTM outputs. Once trained, the model can be used to do an initial automapping for any new study, saving valuable human time to focus on the more complex aspects of SDTM delivery.

Methodology

Scope

Four SDTM domains were included in the scope of our POC:

- Adverse Events (AE)
- Prior and Concomitant Medications (CM)
- Demographics (DM)
- Vital Signs (VS)

We used 109 source–target mapping specifications completed by SDTM mapping experts to train the ML models. The training data set represented a diverse set of studies, and some of the mappings were highly study specific. Therefore, we set a threshold of 10 occurrences of a target variable for inclusion in training.

Feature engineering and modelling

Source variable names were cleaned of special characters. Source form (domain) 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 (NLP) methods.

We used the python scikit-learn library to develop and train four multinomial classification models:

- Random Forest
- Linear Support Vector
- Naive Bayes
- Logistic Regression

The different models used different approaches to categorise the NLP correlations, but they all provided a probability score for each match based on what they learned from the training data, as illustrated in Figure 2.

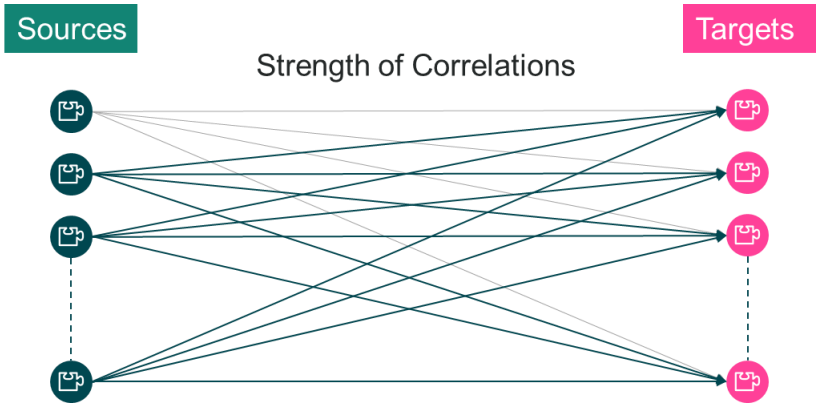

Figure 2: Match probability and strength of correlations

The data were partitioned 75/25 for training and testing, respectively.

Results

The overall model accuracy for each of the domains is described in Table 1. We use the term accuracy to mean percentage of variables correctly mapped.

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

Table 1: Overall model accuracy

Three of the models gave similar results; the fourth, random forest, was less successful. Figure 3 shows the results for each model by domain.

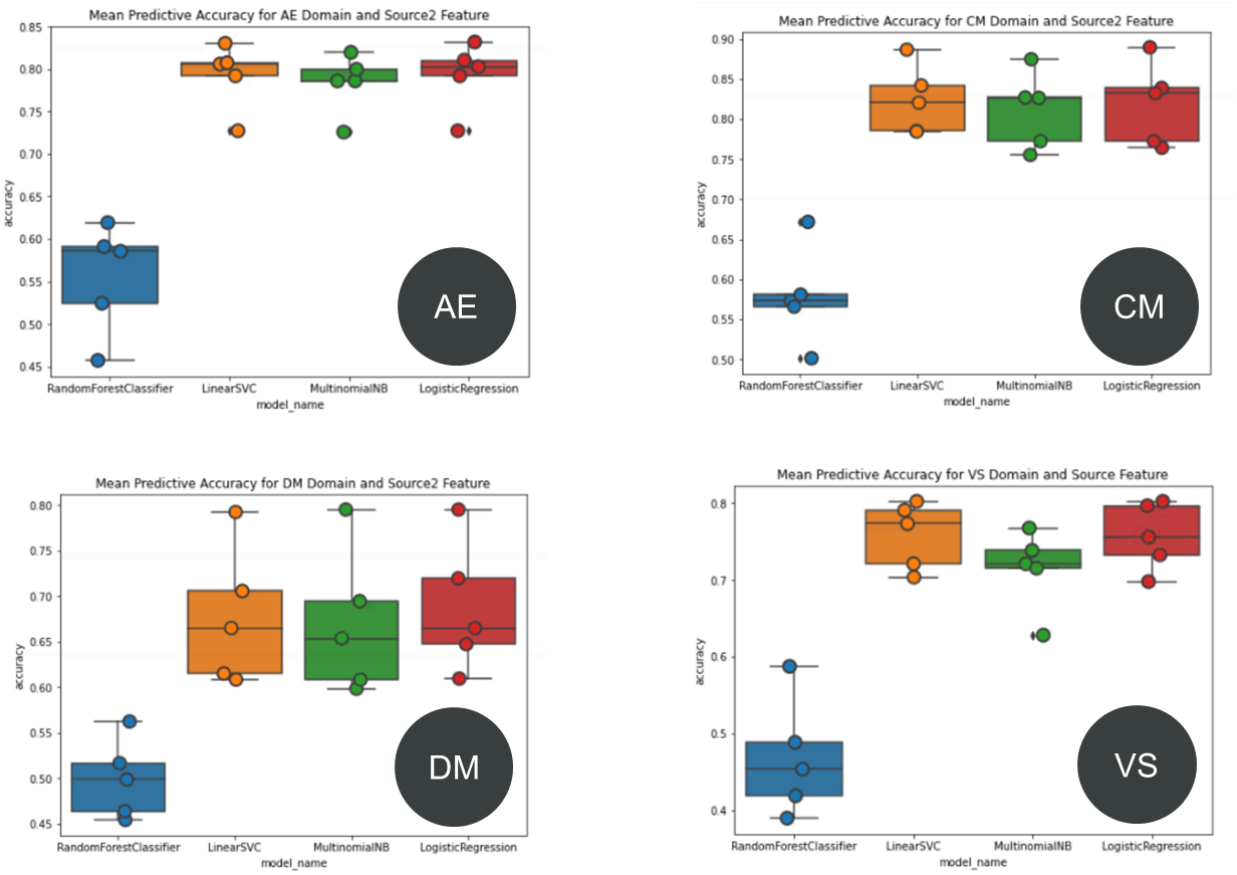

Figure 3: Comparison of results for each model and domain (note: "source2" refers to the feature X.Y)

Finally, we looked at a single study not used for training and used the trained model to map the data from this study. This exercise enabled us to evaluate the model's performance at a more detailed level. Overall, the results were better than the model test output (as shown in Table 2), but we noted some important characteristics, which we discuss further below.

Percent Source to Target Variables Mapped		
Domain	**Model Results (General)**	**Field Test on a Specific (Unseen) Study**
AE	80%	**83.3% (3**0 of 35 mapped correctly)
CM	80%	**85% (**17 of 20 mapped correctly)
DM	65%	**88.2%** (15 of 17 mapped correctly**)**
VS	75%	**90%** (9 out of 10 mapped correctly**)**

Table 2: Model performance for an individual (unseen study) compared with general test results

Our confidence in a particular prediction—whether we called it correct or incorrect—was based on the relative probability scores reported for each variable match to the source variable(s). The trained model provides a match probability for each target variable, summing to 1.

In many cases, the result was clearcut, as illustrated in Figure 4.

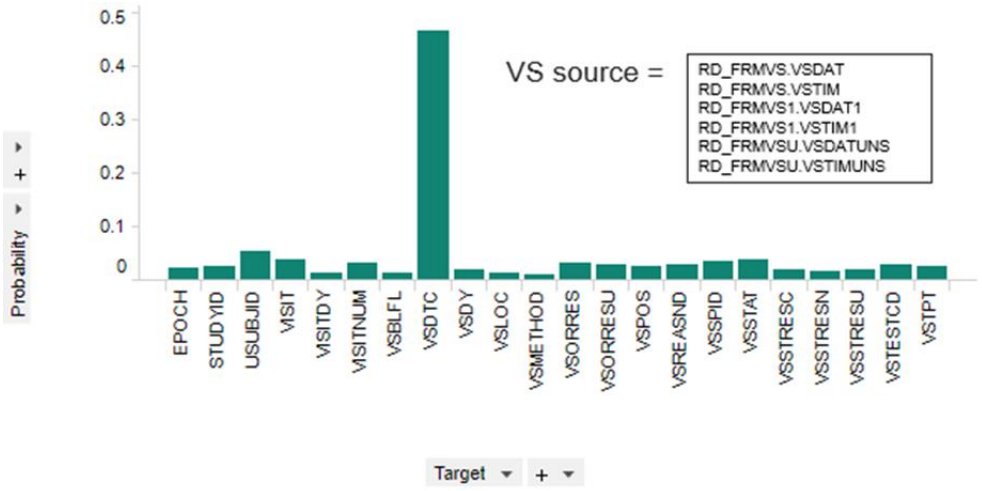

Figure 4: Clear and correct mapping of 6 VS source variables to the SDTM variable VSDTC

In some cases, we identified a false positive, ie, an apparent strong indication of a match but one that is incorrect.

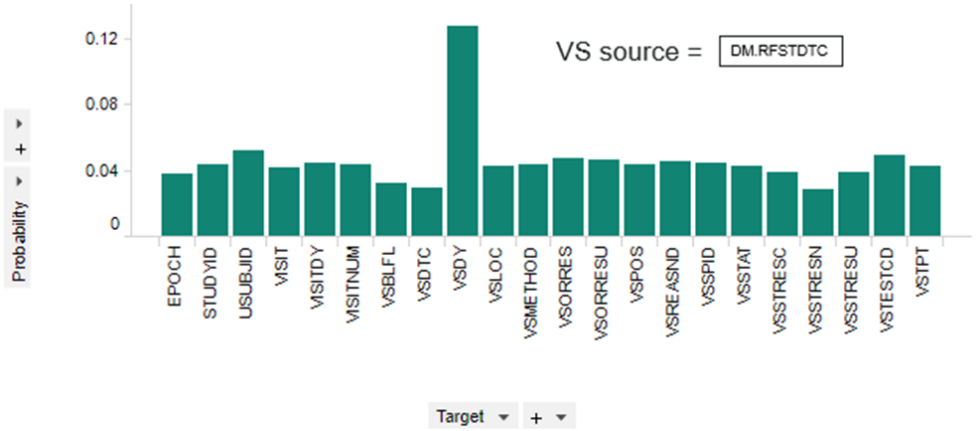

Figure 5: Clear but incorrect mapping of a VS source variable to the SDTM variable VSDY

The explanation for the false positive illustrated in Figure 5 is simply that, in training, VSDY more often mapped to DM.RFSTDTC, so the model was simply reflecting its training. In this case, the study chosen to evaluate the model was very different to the studies used to train the model.

Sometimes, there was no clear match and therefore no conclusive mapping, reflecting insufficient data in the training set.

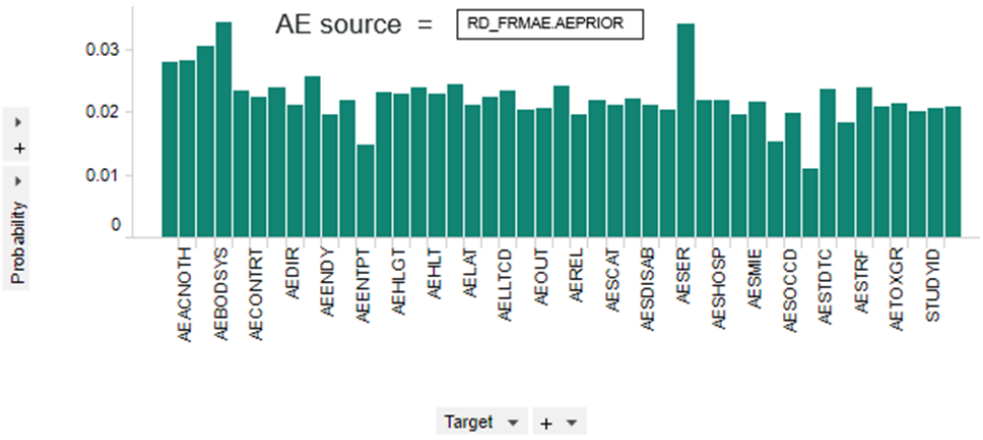

Figure 6: No clear mapping mapping for an AE source variable

The highest match in Figure 6 is AEBODSYS, but this is not a clear winner. The correct match was AESTRTPT, which didn't even feature in the training dataset.

Finally, sometimes, there were competing matches because the model was essentially finding a many-to-many mapping.

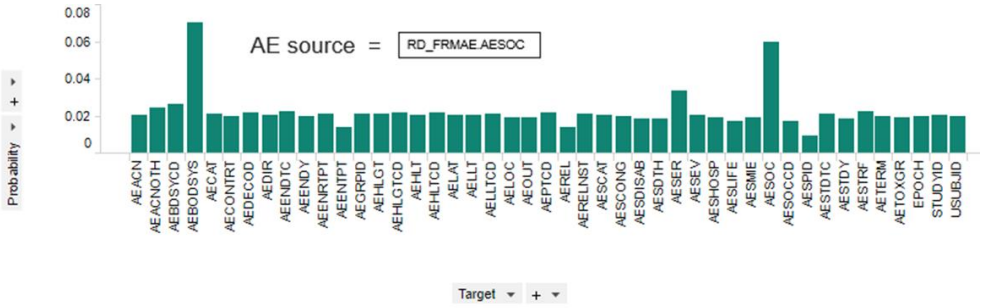

Figure 7: Example of many-to-many mapping for an AE source variable

In the case illustrated in Figure 7, the source variable RD_FRMAE.AESOC was mapped to two targets using some additional logic not used in model training (see Table 3).

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	

Table 3: Mapping for RD_FRMAE.AESOC

Conclusions

ML can deliver pretty decent results for automapping source eCRF variables to target SDTM variables. We encountered some cases where mapping was inconclusive or incorrect, but there are ways to improve accuracy and confidence by enriching the training data and incorporating more of the deeper logic needed to complete an SDTM mapping exercise.

In the future, we would like to extend the training data sets to include more studies and more information, such as mapping logic. We would like to combine NLP and ML to handle value mapping and controlled terminology requirements. To fully integrate automapping with existing mapping processes, we plan to replace the visual analysis of probability matching with a statistical algorithm that provides an easy-to-process confidence measure for end users. This algorithm will allow those who conduct SDTM mapping to make quick decisions about the automapping and complete the parts of the mapping that the ML cannot, accelerating SDTM timelines and improving accuracy and precision in clinical data delivery.