# Automating SDTM for Clinical Trial Data Submissions Using Artificial Intelligence and Machine Learning

**Melanie R. Ciotti, MBA[1], Dale W. Usner, PhD[1], Richard B. Abelson, PhD[1]**
[1]Statistics & Data Corporation, Tempe, AZ, USA
mciotti@sdcclinical.com

## Purpose

SDTM, or Study Data Tabulation Model, is one of the required standards for data submission to the US Food & Drug Administration (FDA). Manually mapping data fields from a clinical trial database to their corresponding SDTM fields has increased statistical programming resources on a clinical trial by approximately 20%. Automating this process through artificial intelligence (AI) aims to reduce the increased time and resources required to format high-quality, CDISC-compliant data for FDA submission.

## Methods

An AI model was developed to automate SDTM mapping via the following 3-step process.

**1. Predict the SDTM Variable:** A machine learning (ML) model was trained using twelve (12) training datasets to predict the corresponding SDTM domain and SDTM variable based on the observed data outcomes. For example, when raw data consistently indicated "M" or "F," the model learned that this corresponded to the SDTM domain DM (demographics) and SDTM variable SEX. **Figure 1a** shows an example of observed data for RACE as a single selection adjacent to the ML-predicted SDTM domain (DM) and variable (RACE).

A Similarity model was employed to measure how closely the clinical database, or EDC (Electronic Data Capture), variable name matches the SDTM variable name. Name-matching via Similarity was used both in conjunction with ML and independently in cases where ML was ineffective, such as with non-descript binary data (e.g., TRUE/FALSE and YES/NO). For example, when a clinical database is built to allow for multiple selections on the variable RACE, each race value could be captured as a TRUE or FALSE outcome. In **Figure 1b**, the AI uses a Similarity model to accurately predict the SDTM domain (DM) and variable (RACE) based on the EDC variable names (RACE_AMERICAN, RACE_ASIAN, etc.).

Together, ML and Similarity produced a baseline prediction of both the SDTM domain and the SDTM variable associated with the observed data outcomes.

**2. Validate and Derive Fields:** Once the SDTM domain and variable are predicted, the model checks relevant reference documentation – CDISC SDTM Implementation Guide and CDISC SDTM Controlled Terminology – to validate and derive fields based on current submission guidelines.

Another Similarity model is employed to validate that the observed values match what is expected in the reference material, including proper formatting. Referencing the CDISC code list values for the SDTM variable RACE shown in **Figure 2**, the model recognizes that the observed values match the code list values, but the code list values are formatted in all capital letters ("AMERICAN INDIAN OR ALASKA NATIVE," "ASIAN," etc.). The AI model maps to the uppercased version of the term to match the SDTM code list value exactly.

If an observed value is not present in a non-extensible code list, the value should be mapped to a supplemental domain. Because RACE is a non-extensible code list in CDISC SDTM standards, the model automatically maps the free text value in "Other, Specify" to a supplemental domain when the patient selects "Other" as their Race (see **Figure 3**).

Additional fields are derived as necessary, such as deriving the patient's age from their date of birth (DM.BRTHDTC) and informed consent date (DM.RFICDTC) if age is not collected independently.

**3. Create SDTM Datasets and aCRF:** In the final step, the model automatically generates SDTM study datasets and an SDTM annotated Case Report Form (aCRF) indicating the domain and variable name for each field. To do this, the program scans the clinical database aCRF to find the EDC variable names; matches those to the SDTM variable names it predicted, validated, and derived in the previous steps; and annotates the proper SDTM variable name adjacent to the corresponding variable. **Figure 3** shows the AI-generated aCRF with EDC variable names in blue text and SDTM domain and variable names in green boxes with red text.

**Figure 1a: Observed Data and SDTM Predictions for RACE (Single Selection)**

| Observed Data RACE | SDTM DM.RACE |
|---|---|
| White | WHITE |
| Asian | ASIAN |
| White | WHITE |
| White | WHITE |
| White | WHITE |
| American Indian or Alaska Native | AMERICAN INDIAN OR ALASKA NATIVE |
| White | WHITE |
| White | WHITE |
| Black or African American | BLACK OR AFRICAN AMERICAN |

**Figure 1b: Observed Data and SDTM Predictions for RACE (Multiple Selections)**

| Observed Data | | | | | | SDTM DM.RACE |
|---|---|---|---|---|---|---|
| RACE_ AMERICAN | RACE_ ASIAN | RACE_ BLACK | RACE_ HAWAIIAN | RACE_ WHITE | RACE_ OTHER | |
| FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | WHITE |
| FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | ASIAN |
| FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | WHITE |
| FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | WHITE |
| FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | WHITE |
| TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | AMERICAN INDIAN OR ALASKA NATIVE |
| FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | WHITE |
| FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | WHITE |
| FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | BLACK OR AFRICAN AMERICAN |

**Figure 2: Sample CDISC Code List for Variable RACE**

| Code | Codelist Cod | Codelist Name | CDISC Submission Value |
|---|---|---|---|
| C74457 | | Race | RACE |
| C41259 | C74457 | Race | AMERICAN INDIAN OR ALASKA NATIVE |
| C41260 | C74457 | Race | ASIAN |
| C16352 | C74457 | Race | BLACK OR AFRICAN AMERICAN |
| C41219 | C74457 | Race | NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER |
| C41261 | C74457 | Race | WHITE |

## Results

The AI model predicted each SDTM domain and variable with at least 80% probability of match. The predicted outcome from the model was always accurate when the reported probability of match exceeded 90%. Based on these results, variables where the probability of match do not exceed 90% will require manual review.

One key success factor was selecting high-quality training datasets to develop the initial functionality. Training datasets should be largely homogenous when starting out, preferably developed by a single company using the same EDC system across the selected studies and mapping to a consistent, current version of SDTM.

Another key success factor was the AI's ability to identify relationships beyond the collected measure. As shown in **Figure 1b**, many variables are indistinguishable based on the observed measure alone. Referencing the EDC variable name and other indicators is paramount in training the AI to correctly distinguish between multiple variables with TRUE/FALSE and YES/NO outcomes.

## Conclusions

Artificial intelligence and machine learning can be used to automate the SDTM dataset and SDTM aCRF development process, thereby decreasing the time and resources required to create high-quality, CDISC-compliant data packages for regulatory submission.

This AI model can be initially applied as a validation technique where statistical programmers manually program the production (or primary) datasets and the AI generates the validation datasets for comparison. In the future, if an AI model is to develop the production datasets that will be included in the final regulatory submission, additional considerations should be made for specifications (spec) creation, dataset validation, risk mitigation, and accountability. Additionally, processes would need to be implemented for re-training and up-versioning based on the software lifecycle. The AI's ability to self-report performance, adapt to changes in the regulatory guidance materials, and provide traceability and documentation throughout its process will be key to its long-term success.

At present, this AI model is undergoing continued refinement with additional training datasets. SDC is developing AI and ML models to automate key processes in clinical trial data collection, management, analysis, and reporting.

**Figure 3: SDTM Annotated Case Report Forms Generated by Artificial Intelligence (RACE as Single Selection on Left; RACE as Multiple Selections on Right)**