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. For Race, the model learned that the uppercased version of the term to match the SDTM code list value exactly.

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.

### 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. Referring to 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.