PharmaSUG China 2025 – Paper DS- 167

Re-designing SDTM Automation: Maximizing CDASH Process Values and Understanding Code-Specs Relationship

Chunpeng Zhao, BeOne

ABSTRACT

This paper presents a reimagined approach to SDTM automation, offering an ideal blueprint informed by practical lessons from real-world development. It outlines strategies to enhance the value of the CDASH process in both CRF-to-SDTM and eDT-to-SDTM automation workflows. Furthermore, the paper explores the interplay between programming code, specifications, and annotated CRFs (aCRFs), providing insights into how these components collectively drive a more efficient and reliable automation process.

INTRODUCTION

In the context of SDTM automation, four primary types of data are involved: TDM data, CRF data, external data, and derived data. Among these, TDM (Trial Design Model) domains represent digitized protocol data and can be created either before or during protocol development. In fact, protocols themselves can be generated based on these TDM domains.

Leveraging the CDASH-SDTM Combo Model and the T Programming Language (Chunpeng Zhao, 2024), transformation specifications for both CRF data and external data can be defined during the CRF design and Data Transfer Specification (DTS) design phases, respectively. These specifications are treated as a form of programming language—namely, the T Programming Language—which can be directly executed by other programming environments without the need for conversion into traditional code. This approach enables a coding-free transformation process for both CRF and external data into SDTM format.

Derived data, on the other hand, are governed by deviation rules defined in the SDTM Implementation Guide (IG). These rules are standardized and do not require study-specific customization, simplifying their integration into the automation process.

TDM DOMAINS

TDM (Trial Design Model) domains represent a structured form of protocol digitalization. While traditional clinical trial protocols are typically unstructured documents, TDM domains capture the same information in a structured, machine-readable format.

According to the principle of entropy, it is generally easier to convert structured data into unstructured formats than the reverse. This principle underpins the design of some protocol automation tools, where TDM domains—or equivalent structured data—are created first. These tools then automatically generate the protocol document based on the structured TDM input.

This approach implies that TDM domains can either be defined prior to protocol development or extracted directly from protocol automation tools. In both cases, TDM domains serve as a foundational layer for automating and streamlining protocol creation.

CRF DATA SDTM TRANSFORMATION

According to the CDASH process, SDTM transformation rules for CRF data should be defined before or during CRF design. This begins with mapping biomedical concepts to SDTM domains, variables, and controlled terminology (CT). Based on this mapping, the CRF design is then developed, including Form OIDs, Field OIDs, field labels, and field options.

When CRF design specifications and SDTM transformation specifications are created simultaneously—following the CDASH-SDTM Combo Model (Chunpeng Zhao, 2024)—the transformation from CRF-collected data to SDTM-minus data (i.e., SDTM data excluding derived variables) can be executed

immediately once the first EDC record is available. This integrated approach enables a seamless and efficient transformation process, supporting early and automated SDTM data generation.

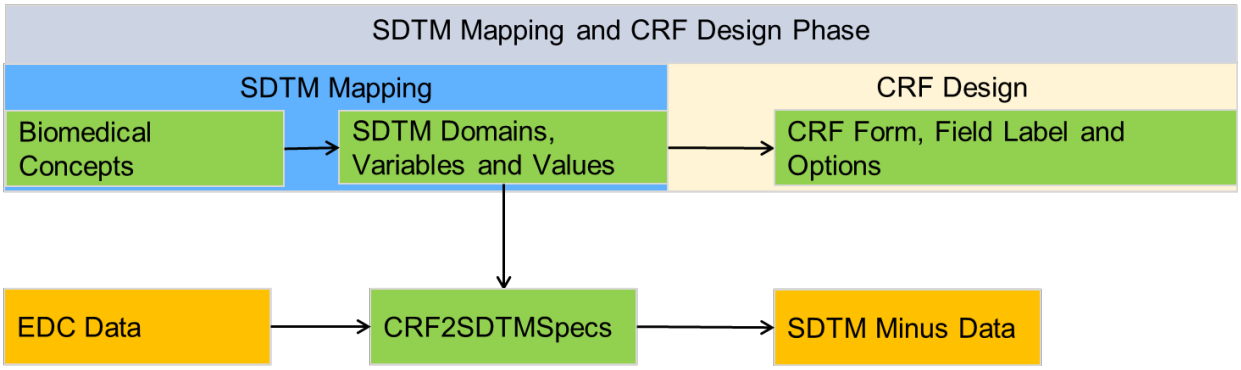

Figure 1. CDISC Process in CRF Data SDTM Transformation

FLOW DATA PROCESSING INSTEAD OF BATCH RUN

Flow-based data processing enables near real-time generation of SDTM-minus data, offering a more dynamic alternative to traditional batch processing. For example, in the RAVE EDC system, two key columns—**RecordID** and **MaxUpdate**—allow for selective extraction of only the records that have been updated. These records are then processed through the CDASH-SDTM Combo Model transformation specifications to produce updated SDTM-minus data, which can replace or remove corresponding entries in the previous version.

For derived SDTM data and ADaM datasets, only subjects with updated records undergo the derivation process. The resulting data then replaces the relevant subject-level records in the existing SDTM and ADaM datasets. This incremental approach improves efficiency, reduces processing time, and supports continuous data integration throughout the study lifecycle.

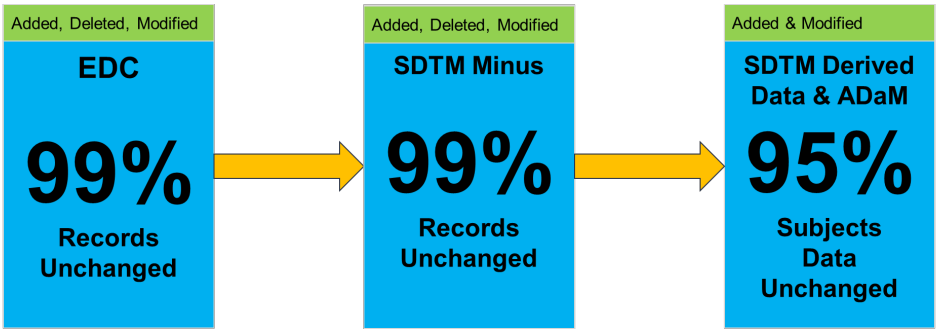

Figure 2. Flow Data Processing focuses solely on updated data

EXTERNAL DATA SDTM TRANSFORMATION

In alignment with the CDASH process, SDTM transformation rules for external data should be defined during the Data Transfer Specification (DTS) design phase. This begins with mapping biomedical concepts to SDTM domains, variables, and controlled terminology (CT). Based on this mapping, DTS variable names and their code lists are then designed.

CDISC-compliant variables are required in the DTS, while additional equivalent variables with alternative wording may be included as optional elements to support flexibility and quality control. For example, as shown in Table 1, variables prefixed with **R_** are optional. These variables are not used for submission or analysis but serve a QC purpose—ensuring that wording changes from raw data to SDTM are correctly implemented.

This structured approach ensures consistency, traceability, and compliance in the transformation of external data into SDTM format.

Table 1 Sample of DTS Variable List

Variable Name	Variable Label
SUBJID	Subject Identifier for the Study
VISIT	SDTM Visit Name
R_VISIT	Raw Visit Name
TPT	SDTM Planned Time Point Name
R_TPT	Raw Planned Time Point Name
REFID	SDTM Specimen ID
DTC	SDTM Date/Time of Specimen Collection
DOMAIN	SDTM Domain Abbreviation
CAT	SDTM Test Panel Name / Category
R_CAT	Raw Test Panel Name / Category
TESTCD	SDTM Short Name of Measurement, Test or Examination
R_TESTCD	Raw Short Name of Measurement, Test or Examination
TEST	SDTM Name of Measurement, Test or Examination
R_TEST	Raw Name of Measurement, Test or Examination
ORRES	SDTM Result or Finding in Original Units
R_ORRES	Raw Result or Finding in Original Units
ORRESU	SDTM Original Units
R_ORRESU	Raw Original Units

SDTM DERIVED DATA

Once TDM domain data, CRF SDTM-minus data, and external SDTM-minus data are available, the derivation of SDTM derived data can proceed using standardized rules defined in the SDTM Implementation Guide (IG). These derivation rules are consistent across studies, enabling a unified approach to generating derived data.

Given this consistency, it is feasible to develop a single, reusable set of code that produces SDTM derived datasets using the three input sources: TDM domain data, CRF SDTM-minus data, and external SDTM-minus data. This approach enhances efficiency, reduces variability, and supports scalable automation across multiple studies.

Figure 3. SDTM Derived Data Full Automation

SPECIFICATIONS AS CODE: LINKING SDTM SPECS, ACRF, AND PROGRAMMING

Building on the CDASH-SDTM Combo Model and the T Programming Language (Chunpeng Zhao, 2024), SDTM specifications can serve as the single source of truth, encompassing all mapping logic found in both the annotated CRF (aCRF) and programming code. When these specifications are designed according to the Combo Model, both aCRF and code can be automatically derived from them.

If the SDTM specifications are machine-readable, they effectively function as T Programming Language scripts. Since SDTM datasets are considered source data for submission, there is no business requirement to manually write programming code—provided the SDTM data can be generated directly from the specifications using the T Programming Language. This makes fully automated, code-free SDTM generation not only possible but practical.

CONCLUSION

By leveraging protocol automation tools or initiating TDM domain creation prior to protocol development, TDM data can be generated automatically. Through the CDASH-SDTM Combo Model and flow-based data processing, CRF data can be transformed into SDTM-minus datasets in real time—without manual coding. Similarly, by defining CDISC-compliant variables during DTS design, external data can be transformed before transfer, streamlining the process.

SDTM derived data follows standardized rules defined in the SDTM Implementation Guide, which remain consistent across studies and are independent of protocol or statistical analysis plans. As a result, the entire SDTM dataset—comprising TDM, CRF, external, and derived data—can be generated through a fully automated, coding-free process, significantly reducing both effort and cost.

REFERENCES

Chunpeng Zhao, 2024. “CDASH-SDTM Combo Model in SDTM Automation” https://www.lexjansen.com/pharmasug-cn/2024/DS/Pharmasug-China-2024-DS10031_Final_Paper.pdf

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

Chunpeng Zhao
BeOne
Chunpeng.zhao@beigene.com