Paper SD07

SDTM automation by using metadata to drive R functions for data transformation

Krishna Gopal, Moderna Inc, Cambridge MA, USA
Jaya Inampudi, Saama Inc, Campbell CA, USA

ABSTRACT

This paper outlines an approach leveraging metadata and R programming to streamline the raw to SDTM conversion process.

By defining dataset specifications, variable mappings, and transformation rules in a centralized metadata repository, the automation framework can dynamically generate the study SDTM Specification and the necessary R code to execute the source-to-SDTM data transformation. This metadata-driven approach ensures consistency, reduces manual effort, and facilitates updates as SDTM standards evolve. Importantly, the generated SDTM specification is user-editable, allowing subject matter experts to review, customize, and maintain the transformation logic as needed. Having digitized the specification, we can automate the generation of R code to transform the source data to SDTM data.

This abstract demonstrates how a metadata-centric design, combined with the power of R programming and user-editable code, can optimize the SDTM automation workflow, leading to improved efficiency, quality, and regulatory compliance in clinical data management.

INTRODUCTION

The pharmaceutical and clinical research industries require efficient and standardized data handling for regulatory submission and analysis. The Study Data Tabulation Model (SDTM) provides a structured framework for organizing clinical trial data in a consistent and regulatory-compliant manner. However, converting raw clinical data to SDTM format is often labor-intensive, prone to errors, and subject to evolving regulatory standards. This paper explores a metadata-driven approach to SDTM automation using R programming, reducing manual effort while ensuring consistency and compliance.

BACKGROUND

SDTM AND ITS IMPORTANCE

SDTM, developed by the Clinical Data Interchange Standards Consortium (CDISC), standardizes clinical trial data for regulatory submission. The model specifies domain datasets, variable naming conventions, controlled terminology, and data structure guidelines. Adherence to SDTM ensures data traceability, quality, and regulatory acceptance.

CHALLENGES IN RAW TO SDTM CONVERSION

Traditional SDTM conversion varies significantly across organizations. Some companies still rely on fully manual processes involving dataset creation, variable mapping, transformation rule implementation, and quality checks, making the process highly time-consuming and error prone. Others have implemented elaborate sets of standards and semi-automated workflows to improve efficiency, but these methods can still require substantial manual intervention and maintenance. Variability in raw data formats and evolving CDISC standards further complicate the process. Automating the conversion through metadata-driven methodologies can significantly enhance efficiency and data quality by reducing inconsistencies and minimizing manual effort while allowing flexibility for evolving standards.

A CASE FOR METADATA DRIVEN SDTM AUTOMATION FRAMEWORK

METADATA REPOSITORY

A centralized metadata repository stores dataset specifications, variable mappings, and transformation rules. This repository serves as the foundation for automating SDTM conversion, ensuring standardization across multiple studies. The centralized repository enables the selection, use, maintenance, versioning of CDISC SDTM implementation standards, and their implementation guides. In addition to using Global regulatory standards, the metadata repository enables the use of SDTM mapping based on historical content from other similar reference studies, enabling re-use of prior mapping based on cosine similarity between mapping parameters.

AUTOMATED GENERATION OF SDTM SPECIFICATIONS

Using the metadata repository, and historical mappings in global standards the automation framework creates the required raw data to SDTM Mappings and additionally generates Pinnacle 21 compatible SDTM specifications. Additionally, the generated specifications contain the calls to R-functions that show how each raw data variable is mapped to the destination SDTM variable and enables a review of unmapped/unmappable entire and/or required SDTM variables for which a source has not been identified based on the available raw-data. The specification mapping process is dynamic, and the system understands which domains are relevant based on historical precedent. This user-editable specification enables subject matter experts to review the mappings and refine transformation logic while maintaining compliance with SDTM guidelines. The set of transformation functions is a user-extensible library for mapping of variables and enables complex/custom derivations.

CREATING AND EXECUTING R-CODE BASED ON SPECIFICATIONS

The SDTM specification is digitized to facilitate automatic R code generation. The automation framework interprets the metadata and dynamically creates R scripts that:

- Identify the corresponding raw dataset(s)
- Apply transformation logic to one destination (SDTM) domain- variable at a time
- Stack the set of transformation logic for the domain, one function call per variable.
- Output a standalone R code with calls to the R-function library, enabling users to edit R code if necessary to fine-tune the programming aspects.
- Generate final CDISC SDTM-compliant datasets

These steps enable a user to not only create SDTM specifications in an automated manner, but also, generate a monolithic R code consisting of function calls for each corresponding variables, enabling users to debug individual process steps and optimize the output data.

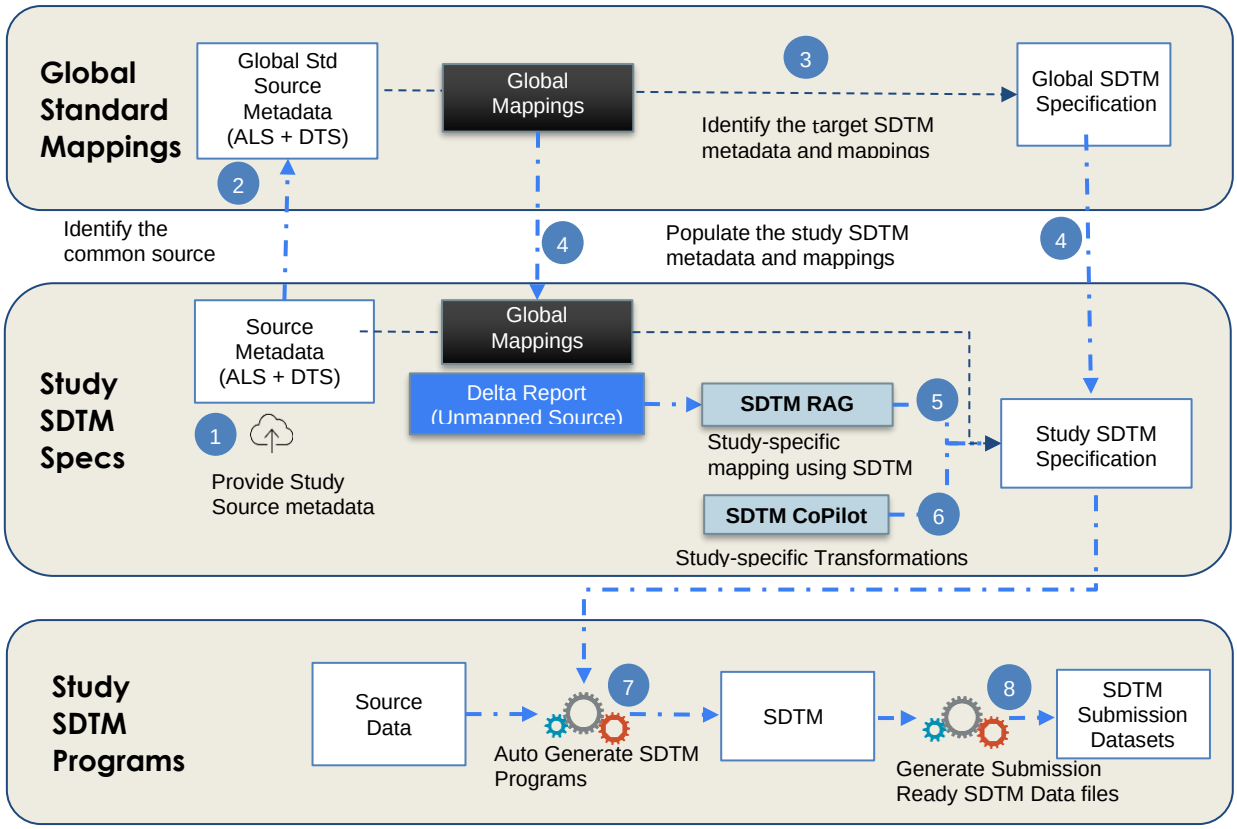

Figure 1. Data and information flow diagram for automated generation of SDTM data, Specifications and Programs

CONCLUSION

Metadata-driven SDTM automation using R programming presents a transformative approach to clinical data management. By centralizing dataset specifications and transformation rules, this approach enhances consistency, reduces manual effort, and ensures compliance with evolving CDISC standards. The flexibility of user-editable SDTM specifications and R code further empowers subject matter experts to customize transformation logic while maintaining automation efficiency. With continued advancements in metadata management, machine learning, and cloud-based deployment, the future of SDTM automation holds promising opportunities for improved regulatory compliance and data quality in clinical research.

RECOMMENDED READING

SDTMIG: https://www.cdisc.org/standards/foundational/sdtmig/sdtm-and-sdtmig-conformance-rules-v2-0
SDTM: https://www.cdisc.org/standards/foundational/sdtm/sdtm-v2-0

CONTACT INFORMATION

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

Author Name: Krishna Gopal
Company: Moderna, Inc.
Address: 523 Binny St, Cambridge MA
Email: kgopal@modernatx.com

Author Name: Jaya Inampudi
Company: Saama, Inc.
Address: 900 E Hamilton Ave Suite 430, Campbell, CA 95008
Email: jaya.inampudi@saama.com