Paper DS-280

From Manual to Automated: An SAS® and R-Based Toolkit for Scalable SDTM Generation

Hardik Sheth, Worldwide Clinical Trials

Eldho Alias, Worldwide Clinical Trials

ABSTRACT

Clinical trials require the submission of SDTM datasets; however, manual programming of SDTM domains is time-consuming and resource intensive. Fully automated solutions often lack flexibility and efficiency when customization is needed to accommodate varying standards or non-standardized data. This paper presents the development of SAS- and R-based toolkits that automate SDTM generation to improve efficiency and reduce manual effort. The toolkits are built on standardized case report forms (CRFs) aligned with CDASH standards, while retaining flexibility to accommodate study-specific requirements. Standard mapping processes are automated, with options for programmers to perform pre- and post-processing as needed.

Implementation of the toolkit reduced SDTM development time by over 30%, while decreasing manual programming effort by over 40% of the traditional total hours spent across multiple studies. The automated approach also improved consistency and data quality, resulting in fewer validation findings and enhanced data visual dashboard compatibilities. The toolkits support a broad range of RAVE data across multiple sponsors, rather than relying on a single sponsor-specific standard. This paper describes the toolkit development methodology and presents measured efficiency gains observed during real-world study implementations.

INTRODUCTION

Standardization of clinical trial data is essential for regulatory submissions and efficient data review, with the Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) serving as the industry standard for organizing clinical trial data. However, implementing SDTM domains across studies can be time-consuming and resource-intensive when performed through traditional manual programming approaches. To address these challenges, this paper presents SAS and R-based SDTM automation toolkit designed to streamline the transformation of raw clinical data into SDTM-compliant domains using a modular and configurable framework. In addition to automating SDTM implementation, the toolkit supports the creation of datasets optimized for dashboards and visualization (See Appendix 2), enabling improved data exploration and study monitoring. The framework also provides flexibility for study-specific custom updates while maintaining standardized processes. Leveraging SAS's data transformation capabilities alongside advanced visualization features, this toolkit improves efficiency, consistency, and usability in SDTM development and clinical data review.

HISTORICAL PROCESS vs AUTOMATED FRAMEWORK

Traditionally the implementation of SDTM domains has relied heavily on study-specific programming, where programmers manually transform raw clinical trial data into SDTM-compliant datasets based on mapping specifications. This process typically involves writing individual SAS programs for each domain, performing manual data transformations, and conducting extensive quality control checks. While this approach allows flexibility for study-specific requirements, it often leads to longer development timelines, limited code reuse, and potential inconsistencies across studies. In addition, preparing datasets for visualization or dashboard reporting is usually handled as a separate downstream process, requiring additional programming and data preparation efforts.

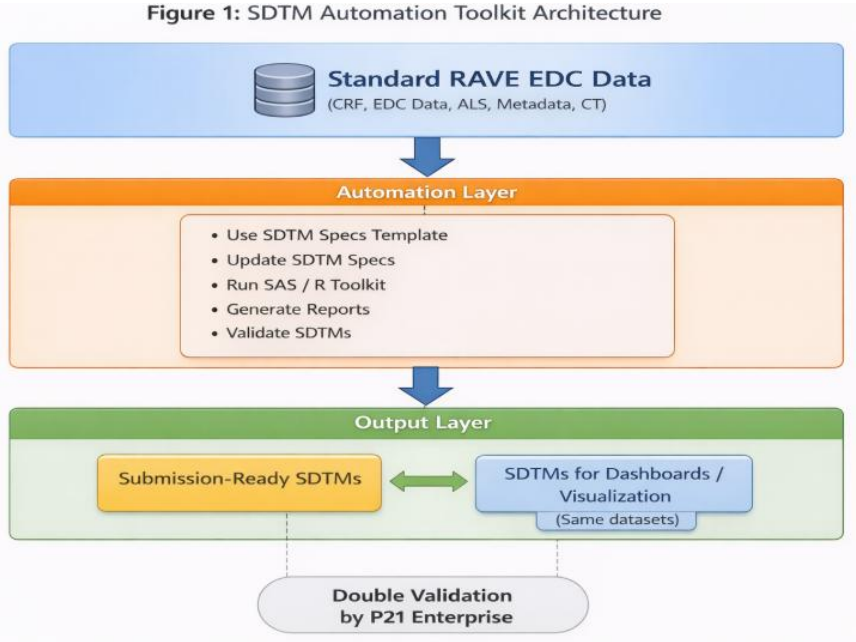
To address these limitations, the proposed SAS and R-based automation toolkit introduces a standardized and configurable framework that streamlines SDTM implementation. The toolkit automates the transformation of raw data into SDTM domains using reusable modules and configurable mapping components, significantly reducing the need for repetitive programming. Furthermore, the framework integrates support for generating datasets optimized for dashboard visualization, enabling faster data exploration and monitoring. At the same time, the toolkit maintains flexibility by allowing controlled study-specific customization without disrupting the standardized workflow. By combining automation, modular design, and visualization support, this framework improves efficiency, enhances consistency across studies, and accelerates the availability of SDTM datasets for downstream analysis and reporting.

COMPARISON OF TRADITIONAL AND AUTOMATED APPROACHES

Feature	Traditional SDTM Process	Automation Toolkit
Programming Approach	Study-specific manual programming	Automated and modular framework
Code Reusability	Limited	High reuse across studies
Implementation Time	Longer due to repetitive coding	Reduced through automation
Consistency Across Studies	Variable	Standardized structure
Visualization Support	Separate downstream process	Integrated dashboard-ready datasets
Study Customization	Manual code modification	Controlled configurable updates

ARCHITECTURE AND WORKFLOW

Figure 1: SDTM Automation Toolkit Architecture



The toolkit architecture consists of three key layers that enhance the efficiency and robustness of the SDTM mapping process.

- Data Layer
- Automation Layer
- Output Layer

DATA LAYER

The Data Layer represents the foundational input for the SDTM automation framework and consists of standardized RAVE Electronic Data Capture (EDC) data, along with Non-EDC vendor data sources. This includes raw clinical trial data captured through Case Report Forms (CRFs), along with supporting artifacts such as Annotation Listing Specifications (ALS), study metadata, and controlled terminology. These components collectively define the structure, semantics, and regulatory context required for SDTM transformation.

A critical step within this layer is the comprehensive review and harmonization of all inputs to ensure alignment with study protocols and CDISC standards. This includes verifying CRF annotations against ALS mappings, ensuring metadata completeness (e.g., variable attributes, codelists, origins), and validating controlled terminology usage. Any inconsistencies or gaps identified at this stage are addressed prior to downstream processing, thereby reducing rework and improving the reliability of automated transformations. By enforcing standardized and high-quality inputs, the Data Layer establishes a robust foundation for consistent and scalable SDTM implementation.

AUTOMATION LAYER

The Automation Layer is the core engine of the framework, where SDTM implementation is executed using parallel SAS and R-based toolkits that follow a consistent, metadata-driven approach. This layer leverages SDTM specification templates, mapping metadata, and configurable business rules to automate the transformation of raw EDC data into SDTM-compliant domains.

The processing workflow includes domain creation, variable derivations, controlled terminology assignments, and dataset structuring in accordance with CDISC SDTM standards. The framework supports study-specific customization through pre-processing (e.g., data standardization, data cleaning, intermediate transformations) and post-processing (e.g., domain-specific adjustments, sponsor-defined variables), allowing flexibility without compromising standardization.

Within this layer, the toolkit generates comprehensive error-tracking and warning reports (See Appendix 1) that capture issues such as controlled terminology discrepancies, variables exceeding length limits (e.g., greater than 200 characters), user-defined warnings, and variables containing only missing values. These reports enable users to proactively identify and address potential data issues, facilitating more efficient decision-making. As a result, the framework reduces manual programming effort, enhances reproducibility, and ensures consistency across studies while maintaining high-quality outputs.

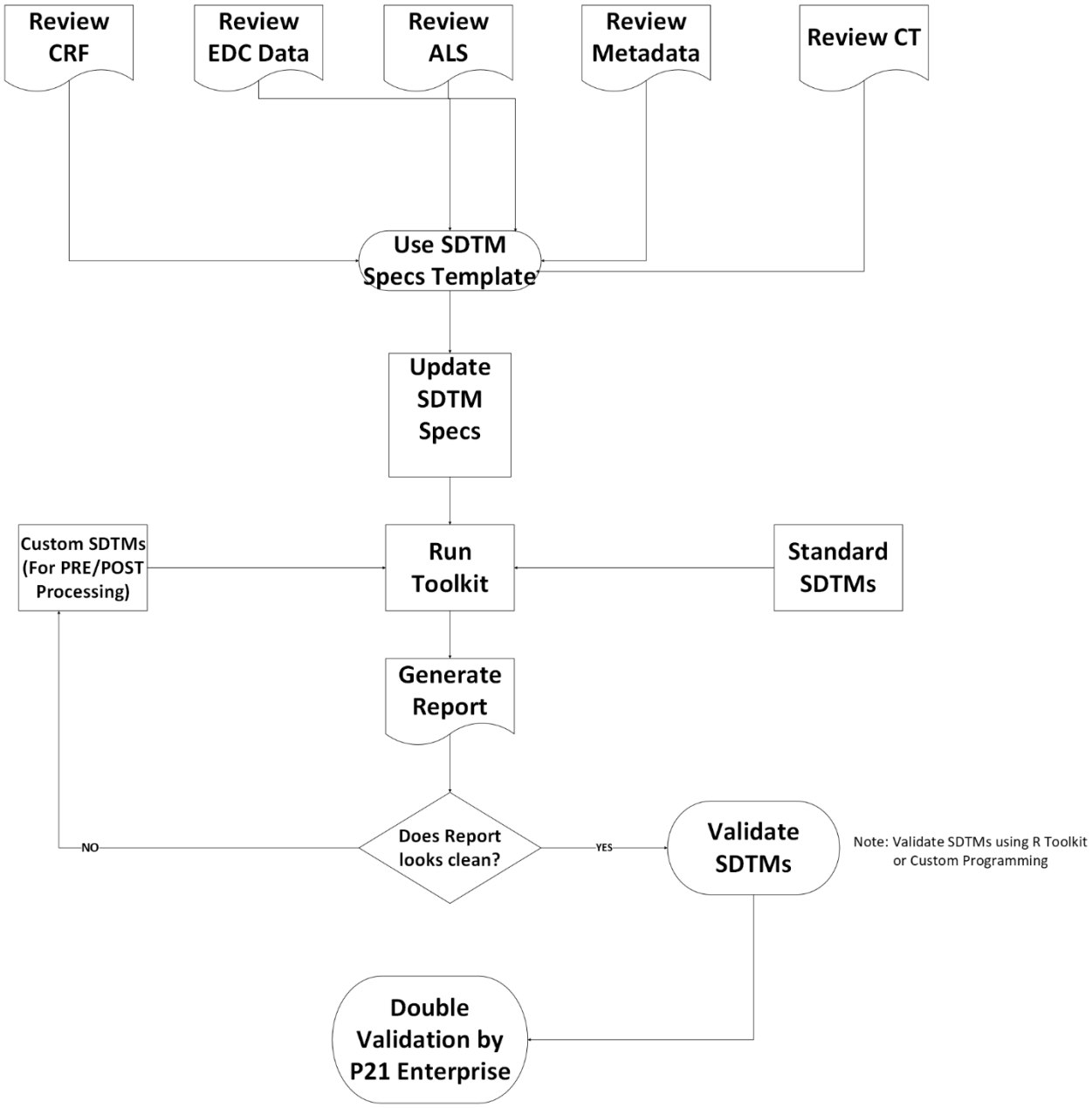
OUTPUT LAYER

The Output Layer delivers the final SDTM datasets generated by the framework. It includes both submission-ready SDTMs for regulatory purposes and SDTMs used for dashboard visualization (See Appendix 2), which are the same datasets optimized for interactive data exploration and monitoring. We need to perform validation by Pinnacle 21 to assure the SDTMs are compliant with regulatory purposes. This layer ensures that standardized data is readily available for both regulatory submission and analytical review.

DETAILED PROCESS FLOW OF SDTM GENERATION USING TOOLKITS

The figure illustrates the end-to-end SDTM automation workflow, beginning with Standard RAVE EDC data and review of key study documentation. The process progresses through specification-driven mapping, toolkit execution, and automated report generation. Built-in validation steps ensure data quality and compliance throughout the workflow. The framework supports both standard and non-standard data structures through configurable pre- and post-processing. This automated approach enhances consistency, traceability, and efficiency across studies, ultimately producing SDTMs suitable for both regulatory submission and downstream visualization.

Process Chart for Toolkit



KEY ADVANTAGES OF TOOLKIT ARCHITECTURE

The proposed SDTM automation framework provides several advantages over traditional SDTM implementation approaches, particularly in terms of efficiency, standardization, and scalability.

1. **Metadata-Driven Standardization**
   The framework leverages a metadata-driven approach using SDTM specification templates and configurable rules, ensuring consistent domain generation across studies. This reduces variability introduced by manual programming and promotes alignment with CDISC standards.

2. **Dual-Technology Flexibility (SAS and R)**
   By implementing parallel toolkits in both SAS and R, the framework provides flexibility for organizations to adopt either technology based on project needs, resource availability, or long-term strategy. Despite using different programming environments, both toolkits follow identical workflows and produce consistent outputs.

3. **Reduced Manual Effort and Increased Efficiency**
   Automation of domain creation, derivations, and validation significantly reduces manual programming effort and development time. Reusable components and standardized processes enable faster study start-up and delivery timelines.

4. **Integrated Validation and Error Tracking (See Appendix 1)**
   The framework incorporates built-in validation checks and error-tracking reports, allowing early identification of data issues. This proactive approach improves data quality and reduces the need for extensive downstream rework.

5. **Support for Study-Specific Customization**
   Through pre- and post-processing mechanisms, the framework allows flexible handling of study-specific requirements while maintaining a standardized core structure. This ensures adaptability without compromising consistency.

6. **Unified Output for Submission and Visualization**
   The framework generates a single set of SDTM datasets that serves both regulatory submission and dashboard visualization (See Appendix 2) purposes. This eliminates duplication of effort, ensures consistency between outputs, and supports real-time data review and decision-making.

7. **Scalability and Reusability Across Studies**
   The modular design of the toolkits enables easy reuse across multiple studies, therapeutic areas, and sponsors. This scalability makes the framework suitable for large-scale clinical development programs.

ASSUMPTIONS AND RULES

To support a standardized and automated SDTM implementation framework, the following assumptions and rules are defined to ensure consistency, scalability, and robustness across studies:

- Standardized Case Report Forms (CRFs) are utilized wherever feasible to promote uniform data collection.
- Toolkit is built on standardized RAVE EDC Database.

- A standardized SDTM specification template serves as the foundational framework for all implementations.
- Comprehensive error handling and validation mechanisms are incorporated to ensure reliable and reproducible execution.
- The toolkits are designed to support mapping for both standard and non-standard CRF and EDC data structures.
- Pre-processing and post-processing steps are implemented to address non-standard CRF designs and sponsor-specific standardization requirements.

WHY TWO (SAS- AND R-) TOOLKITS?

The use of both SAS and R toolkits for SDTM automation provides complementary advantages that enhance flexibility, scalability, and adoption across organizations. SAS remains the industry standard for regulatory submissions, with well-established validation processes and broad acceptance by regulatory agencies. In contrast, R offers greater flexibility, open-source capabilities, and strong support for advanced analytics and visualization. By implementing parallel toolkits with identical workflows and outputs, organizations can leverage SAS for compliance and submission needs while utilizing R for innovation, rapid development, and cost efficiency. This dual-toolkit approach also enables cross-validation of outputs, improves reproducibility, and supports diverse project requirements and user preferences across teams.

FUTURE ENHANCEMENT AND SCALABILITY

This paper presents the initial version of the SDTM automation toolkit, which streamlines and standardizes the SDTM implementation process. While the current framework demonstrates significant improvements in efficiency and consistency, several enhancements are planned to further strengthen its capabilities. Ongoing development efforts focus on incorporating the latest industry and data standards to ensure continued alignment with evolving regulatory expectations. In addition, the toolkit is being enhanced to improve robustness and reduce processing time.

Future enhancements include expansion to support additional databases commonly used by data management during development, thereby increasing the toolkit's applicability across diverse data sources. The toolkit is also a strong candidate for the integration of artificial intelligence, which can further optimize processes such as metadata comparison, mapping validation, and anomaly detection.

Future iterations will also include the integration of metrics and analytics to support data-driven decision-making and provide insights into standardization practices across studies.

CONCLUSION

The SDTM Automation Toolkit provides an efficient and standardized approach to clinical data standardization, significantly reducing manual programming effort while ensuring high-quality, compliant datasets. By automating repetitive tasks and supporting consistent application of industry standards, the Toolkit enhances productivity and reliability in clinical programming workflows. Planned enhancements, including integration with evolving data standards, additional decision-support metrics, and improved scalability, will further strengthen its value and applicability. Overall, the Toolkit represents a practical step toward fully automated, high-quality SDTM data preparation and lays the groundwork for broader adoption in future studies.

CONTACT INFORMATION

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

Hardik Sheth
Worldwide Clinical Trials (Legacy Catalyst Clinical Research)
Email: hardik.sheth@worldwide.com

Eldho Alias
Worldwide Clinical Trials (Legacy Catalyst Clinical Research)
Email: eldho.alias@worldwide.com

APPENDIX 1. ERROR – TRACKING REPORT

This report is designed to monitor and document potential issues in clinical trial datasets, helping ensure data quality and compliance with SDTM standards. The report includes the following tabs:

1. **CT Change** - Tracks changes made to the Controlled Terminology (CT) values to ensure consistent coding across datasets.
2. **RAW Vars with length > 200** - Identifies raw variables exceeding 200 characters, which may require splitting the variables into SUPP-- domain.
3. **Variable with Null value** - Flags SDTM variables containing NULL or missing values that may require review for completeness or imputation.
4. **User Defined Warning** - Captures custom warnings set by the user or system, such as duplicate records, missing key variables, or format mismatches.
    - Example warnings include duplicate records in DS and EX, new 'Not Applicable' values in MH, or missing QNAM variables in SUPPDD.

DOMAIN	WAR_COMMENT
DS	Duplicate records found. Update key variables
EX	Duplicate records found. Update key variables
MH	New value 'N/A: Not Applicable' found in MHTOXGR. Update the format $mhtoxgr.
SUPPDD	Variable corresponding to the QNAM: DTHAENO does not exist in the source dataset.
SUPPDD	Variable corresponding to the QNAM: DTHOTH does not exist in the source dataset.

CT Change | RAW Vars with length gt 200 | Variable with Null value | User Defined Warning

APPENDIX 2. DASHBOARD OUTPUT FROM TOOLKIT-GENERATED SDTMS

Below figure presents a sample dashboard generated using SDTMs produced by the automation toolkit. The same standardized SDTM datasets used for submission are leveraged for visualization, ensuring consistency between analysis and reporting. The dashboard demonstrates how key clinical insights—such as subject disposition, adverse events, and exposure summaries—can be dynamically visualized, enabling faster and more informed decision-making.