PharmaSUG 2023 - Paper DS-193

Can We Do It Better? Real-Time Validation of SDTM Mapping is Superior to Double Programming

Daniel Rolo, Bremer Louw, Bioforum

ABSTRACT

Accurate, complete, and reliable clinical trial data is paramount to robust decision-making by regulatory authorities. Methodologies used to validate the data in question are resource and time intensive. The clinical industry typically relies on dependable but antiquated methods of validation to ensure that clinical subject data is robust. The most common approach is the use of double programming, predominantly using the SAS® programming language.

This paper introduces a systematic approach to the real-time validation of SDTM mapping of clinical trial data. The approach is module-based comprising of a thorough review of the mapping logic, verification that all source data points have been converted to SDTM (SDTM completeness) and confirming compliance to industry validation rules (SDTM compliance).

Early and frequent validation of SDTM mapping by demonstrating mapping logic correctness, SDTM completeness and compliance, breathes fresh life into the SDTM mapping process and eliminates the resource drain associated with double programming. This paper implores the industry to embrace this evolution in SDTM mapping and paves the way to a quality-by-design approach that empowers professionals to focus on the non-repetitive, decision-making aspects of clinical data handling.

INTRODUCTION

This paper explores a technology-enabled, zero-programming approach to real-time validation of SDTM mapping, addressing three essential components of SDTM data quality: 1) Correctness of mapping logic, 2) SDTM compliance, and 3) SDTM completeness. Exploring these quality components is crucial to the development and execution of a validation methodology that enables a novel approach to the SDTM conversion process.

This paper provides examples of how real-time validation can be achieved without the need for double programming, and attempts to address the burning question: “Can we do it better?”

THE 3 KEY SDTM QUALITY COMPONENTS AND CURRENT VALIDATION PRACTICES

The SDTM framework is the required format for clinical data submissions to regulatory authorities like the FDA [1]. As familiarity with the standard has increased, so has the importance of producing high-quality SDTM datasets. Organizations transforming data to the SDTM standard are required to consider a multitude of resources and guidance documents published by CDISC and affiliated institutions.

Adhering to this diverse set of standards, models, and rules presents challenges, such as ensuring that the chosen standard is compatible with the trial design, using the correct version of the standard, and deciding whether to use company-controlled terminology for certain variables rather than the CDISC-published global set. Over time, organizations have developed common approaches to the creation of SDTM datasets, more specifically, the development of mapping specifications to define how source variables are transformed into SDTM variables and the use of SAS macros to facilitate attribute assignments to SDTM datasets.

To ensure the production of high-quality SDTM datasets, every step in the development process must be rigorously validated. Validation often requires resource-intensive manual reviews, double programming of datasets, and reliance on validation software and tools to “catch” significant issues. To comprehensively account for the quality of SDTM data, SDTM validators need to address the 3 components of SDTM quality, depicted in Figure 1 and described in the following sections.

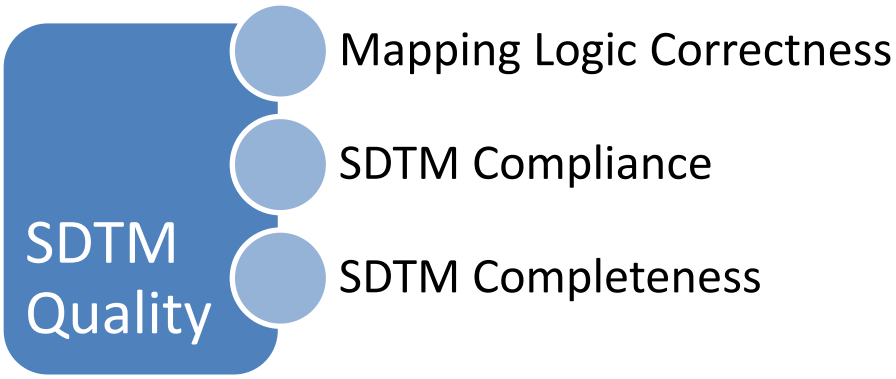


Figure 1: The three key components of SDTM data quality.

MAPPING LOGIC CORRECTNESS

Mapping logic correctness refers to the validity of an instruction being executed to transform a source variable into a target SDTM variable. The correctness of the mapping logic directly influences the quality of resulting SDTM data. An example of scrutinizing mapping logic is assessing how EPOCH values are assigned in subject-level domains. EPOCH assignment decisions are multi-faceted and rely on several assumptions including the design of the trial, the definition of trial elements, and determining tiebreakers for observations that occur on the "border" of two elements.

Establishing if the mapping logic in SAS programs is correct and appropriate for the study, can require multiple rounds of reviews or programmatic validation and time-consuming manual tracking of data values, however, even this approach does not fully qualify the data point as being of high quality! It is also difficult to keep programs and specifications aligned through multiple revision cycles.

To ensure that mapping logic correctly applies to all existing data, we must consider the actual data values when scrutinizing a particular mapping rule. Data issues and anomalies identified while analyzing data may require that a value be traced back through SDTM transformation to the source data. When using double programming validation, the ability to trace data typically requires the tracing of programmed code which can be unnecessarily time consuming.

SDTM COMPLIANCE

Mapping logic determines how a data point is represented in the SDTM data; however, the representation must also adhere to CDISC and regulatory standards, with niche considerations depending on the therapeutic area and circumstances surrounding the trial (e.g., the modified recommendations for representation of subject visit data in studies disrupted by the COVID-19 pandemic [2]). Commonly, SDTM compliance checks are performed using validation tools such as Pinnacle 21 Validator or programs that compare SDTM data to a series of rules to assess compliance.

The available compliance tools require the creation of the SDTM dataset to be complete before executing compliance checks on the dataset, and only then corrections can be made. This results in a repetitive, back-and-forth cycle until all issues have been resolved.

SDTM COMPLETENESS

Given the substantial impact that missing data can have on clinical trial results [3], data completeness (accountability and integrity) is the cornerstone of a validation strategy. Typical data sources for contemporary clinical trials range from the Case Report Form (CRF) data to external sources such as protocol deviations from clinical trial management systems, laboratory biomarker assays, and patient-reported data from wearables such as health tracking devices.

Numerous data reviews, reconciliation programming, double-programming, and spot-checks are performed during the lifecycle of the study. However, these attempts to account for all the data are often not sufficient to ensure data completeness. For example, a mapping rule in an SDTM specification

document may have a minor flaw that omits a specific visit for a lab assessment. If the double-programming process follows the insufficient instruction, the faulty mapping logic will result in two versions of incomplete data that do not detect the original omission.

Verifying the mapping logic and dataset compliance alone cannot determine that a source record or data point is accounted for in the SDTM data. This inadequacy highlights the need for a solution that verifies data completeness. Some solutions using existing programmatic practices have been posited in recent years to address data completeness and traceability, for example, the definition of data traceability variables in the ADaM implementation guide [4] or the use of a tagging approach to individually tag variables/data points during the SDTM development process [5]. “Can we do better?” Yes, later in this paper we will explore how we can extend the latter tagging approach to include real-time data completeness tracking.

REAL-TIME SDTM VALIDATION SOLUTIONS

We have explained that validation of mapping logic correctness, SDTM compliance, and completeness are fundamental to high-quality SDTM data. However, current validation approaches are resource- and time-intensive and inefficient at producing high-quality SDTM data. In addition, validation of SDTM data is often postponed until SDTM development is complete. This delay in validation results in a continual iterative development process. Clearly, we require a solution to validate SDTM data in real-time while addressing the factors that contribute to high-quality SDTM data.

MAPPING LOGIC CORRECTNESS

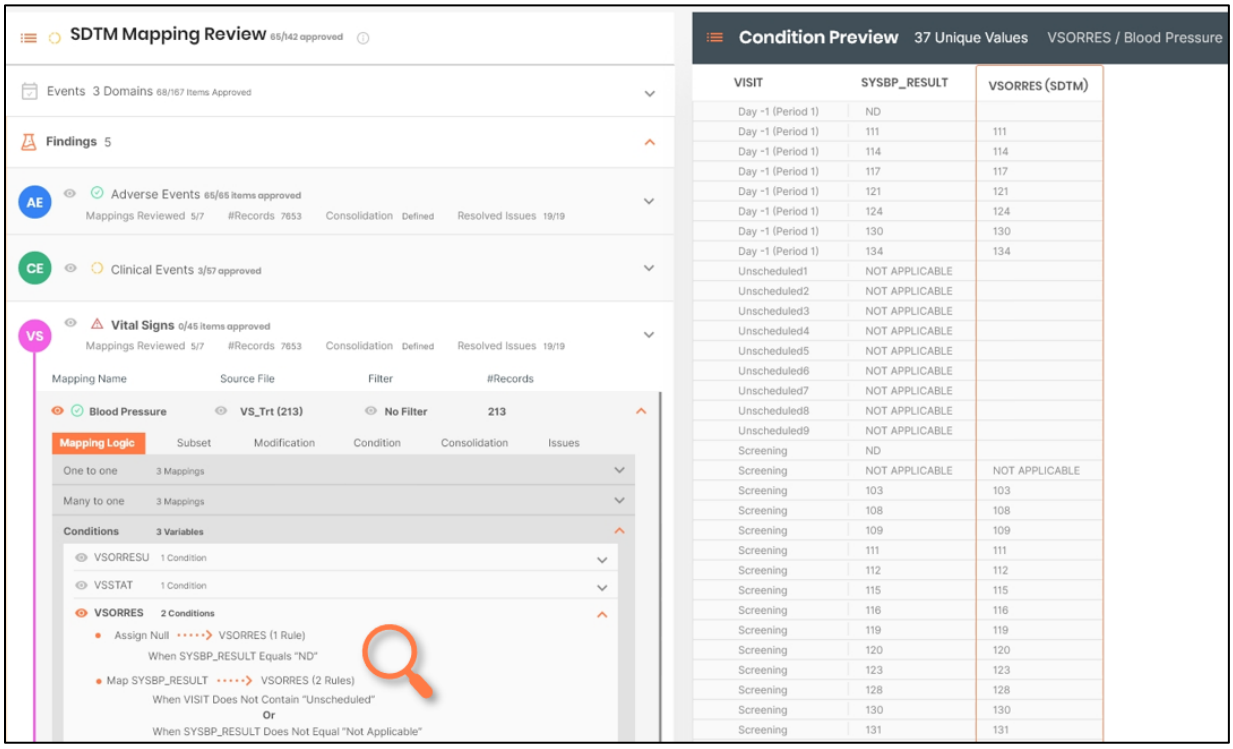
Machine-learning recommendation layer

Introducing machine-learning to support the initial development of mapping instructions will increase the accuracy of mapping logic in real-time. The proposed solution will ingest source data and use machine-learning methodology to recommend mapping logic that the SDTM developer can accept or modify in real-time. The SDTM developer will interact with the system via a graphical user interface to map data based on the system’s recommendations, as depicted in Display 1: A machine-learning powered interface providing mapping recommendations. This approach does not require you to create programming scripts to execute mapping instructions as mentioned earlier.

Display 1: A machine-learning powered interface providing mapping recommendations.

Review tool

Supplementing the initial development of mapping instructions with a tool that facilitates SDTM mapping review will increase the accuracy of mapping logic by allowing the reviewer to access the real time data and ensure that the mapping logic is correct and covers all received data rather than only expected data. The proposed review tool, depicted in Display 2, enables a user to perform and document mapping logic reviews in a user-friendly manner. Holistic data review enables the user to assess anomalies such as source data issues identified by aggregated data and outliers. The advantage of this tool is that a user can review with ease at any time in the SDTM development cycle.



Display 2: An example of a review tool to assess the accuracy of SDTM mapping logic.

SDTM COMPLIANCE

Real-time validation engine

Real-time SDTM compliance validation can be achieved with a tool that integrates compliance validation of transformed SDTM data as mapping instructions are executed on data. This solution can be configured to execute industry compliance rules, regulatory business rules, or custom rules as required by the user. The real-time approach differs from current validation practices in two ways:

- Real-time validation feedback enables the mapper to view and adjust instructions as SDTM data is created.
- The ability to develop and execute custom validation rules, utilizing the CDISC CORE framework, to enhance SDTM compliance validation during the mapping process.

SDTM COMPLETENESS

A SAS-based tagging approach

The significance of data accountability is undeniable; but assessing data completeness still presents obstacles. As stated previously, the current validation practices of manual reviews and double

programming support the validation of the transformed data but do not comprehensively account for all sources of data.

A programmatic solution for checking data completeness is to use a series of SAS macros that perform the SDTM mapping transformations while outputting a copy of the source dataset with tags indicating which data points have been transformed to SDTM. This methodology splits the SDTM programming and mapping process into a series of “building blocks” in which each variable mapping is handled by a macro statement that first executes the mapping instruction, and then marks the source record as having been mapped. Figure 2, Figure 3 and Figure 4 illustrate how the macros would transform data to SDTM standards and tag the source data for the applicable use case.

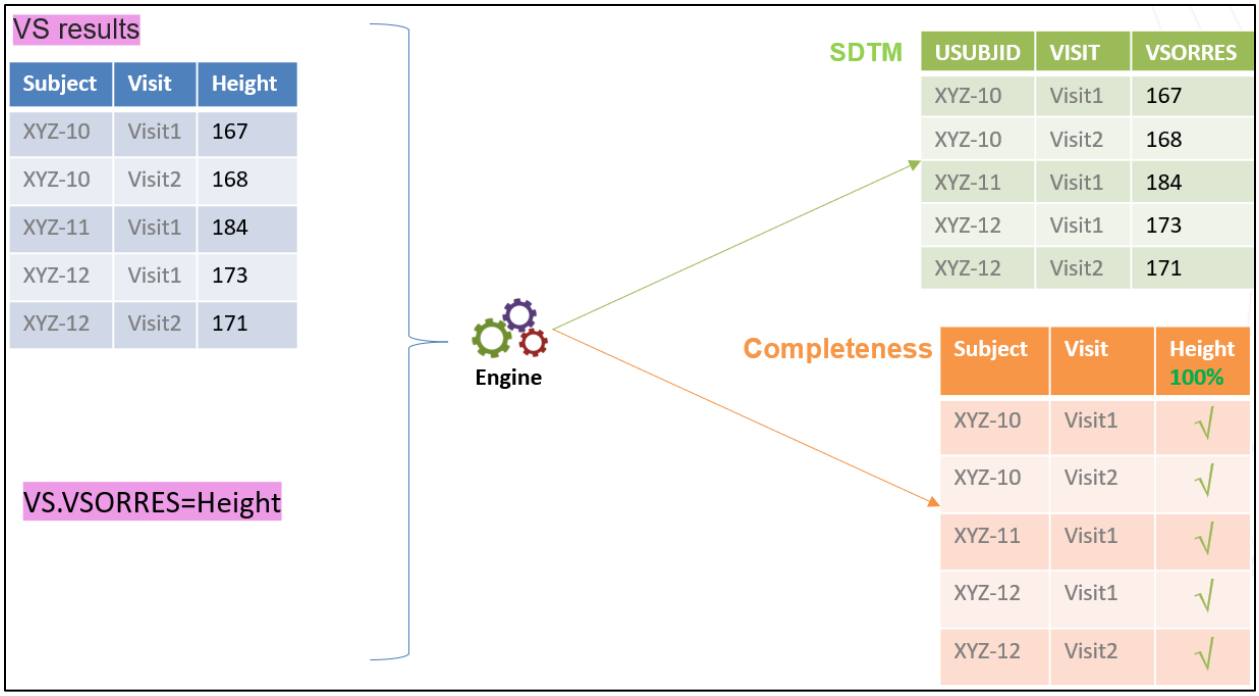


Figure 2: All values are mapped to SDTM and tagged in a completeness dataset.

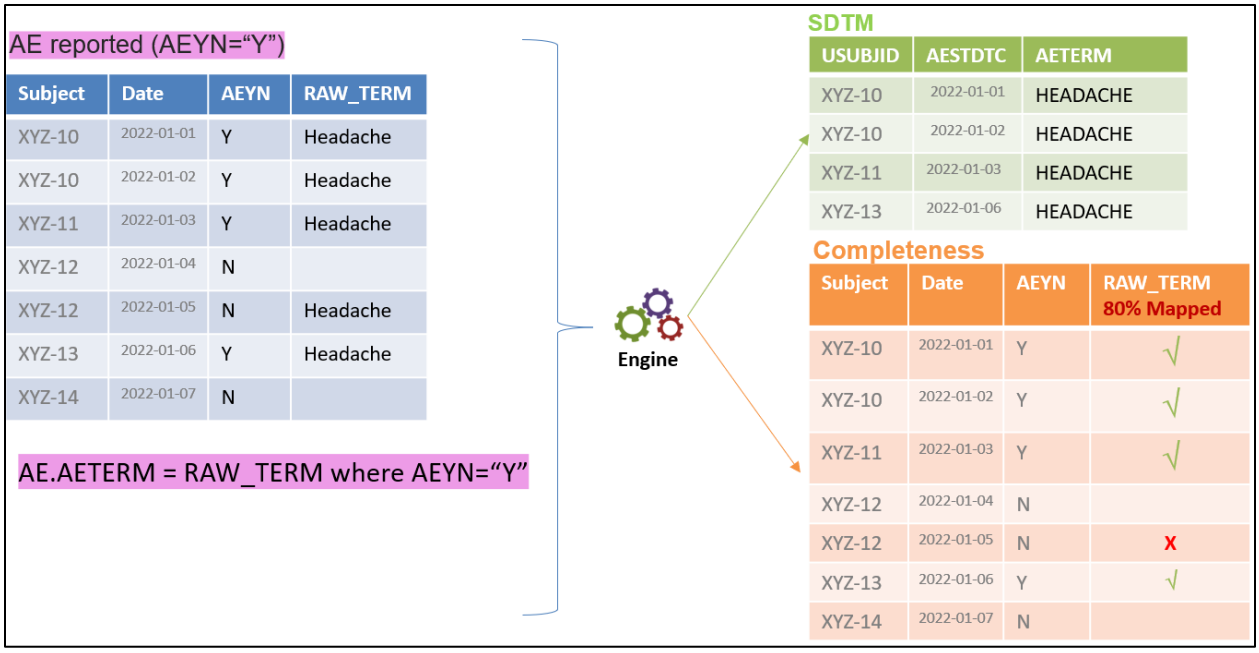


Figure 3: A subset of values are mapped to SDTM and tagged in a completeness dataset.

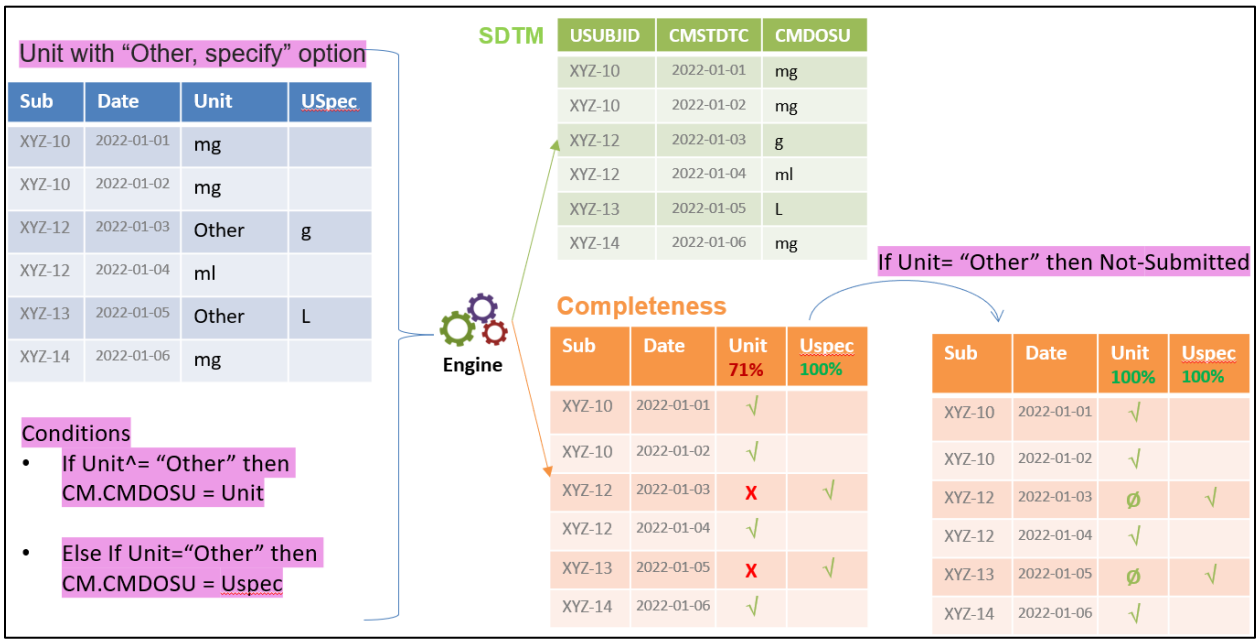


Figure 4: Values are conditionally mapped to SDTM and associated completeness tagging.

Source data tagging makes it possible to generate a report of data completeness that can be used to check data accountability and handle any discrepancies appropriately. The tagging approach replaces the traditional use of SAS data steps and procedures used to transform the data by producing and tagging variables simultaneously. The macros for implementing this solution need to be created and maintained by skilled programmers, but users thereof do not require advanced programming skills. Output 1 illustrates the macro calls required to execute these tagging scenarios described in Figure 2, Figure 3 and Figure 4.

```
/* Scenario 1*/
%all_values(source_lib=source, raw_data=Vital, raw_variable_name=Height, SDTM_domain=VS, SDTM_variable=VSORRES, keys=USUBJID VISIT);
%calculate_completness(raw_data=Vital, raw_variable_name=Height);

/* Scenario 2*/
%subset_values(source_lib=source, raw_data=Adverse, raw_variable_name=RAW_TERM, statement=AEYN eq "Y", SDTM_domain=AE, SDTM_variable=AETERM, keys=USUBJID AESTDTC);
%calculate_completness(raw_data=Adverse, raw_variable_name=RAW_TERM);

/* Scenario 3*/
%conditional_mapping(source_lib=source, raw_data=ConMed, conditions=[Unit | if Unit ne "Other"][Uspec | if Unit eq "Other"], SDTM_domain=CM, SDTM_variable=CMDOSU, keys=USUBJID CMSTDTC);
%calculate_completness(raw_data=ConMed, raw_variable_name=Unit);
%calculate_completness(raw_data=ConMed, raw_variable_name=Uspec);
%add_not_submitted(raw_data=ConMed, raw_variable_name=Unit, statement=Unit eq "Other");
%calculate_completness(raw_data=ConMed, raw_variable_name=Unit);
```

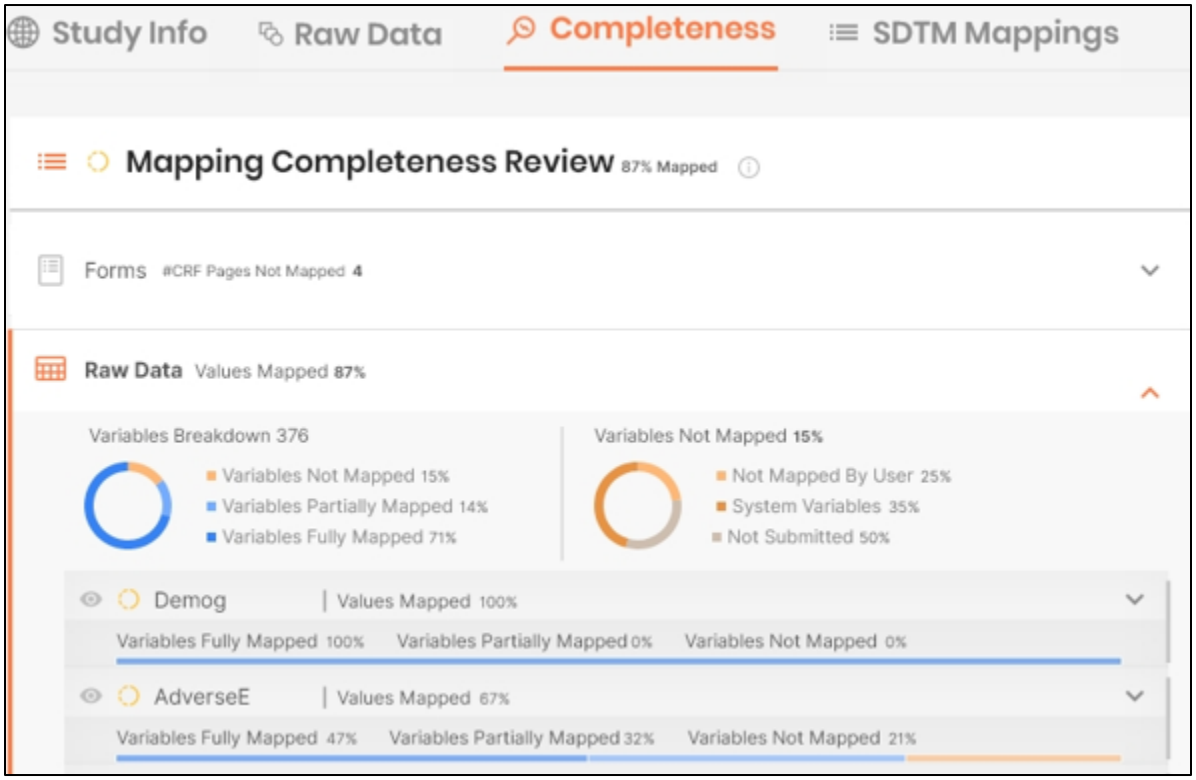
Output 1: Example of SAS macro calls executing the tagging scenarios illustrated above.

The SAS tagging method provides a programmatic solution for completeness validation, but the solution is static and requires manual execution at certain points in the development lifecycle. This prompted the question of whether it was possible to improve upon the current best practice of an inefficient 'waterfall' approach to validation.

Built-in data accountability and data traceability tracking

A tool that provides real-time metrics on data completeness and traceability will enable SDTM developers to confidently account for all source data during the development process without having to execute external scripts or perform manual reviews. This solution monitors the mapping and transformation of data points from ingested source data into the target SDTM structure.

Display 3 illustrates how an SDTM developer can view data completeness metrics for the proportion of source data represented in the SDTM data at the variable and dataset levels in real-time. In addition, a user can mark source records or variables that are not intended for SDTM mapping (e.g., operational or system variables) and such data are accounted for in the metrics. If mapping instructions or source data are modified, the SDTM completeness metric updates to reflect the real-time status.



Display 3: Example of completeness metrics for SDTM mapping.

IMPACT OF REAL-TIME SDTM VALIDATION SOLUTIONS & FUTURE CONSIDERATIONS

An integrated system that incorporates the real-time validation solutions suggested in this paper would:

- Eliminate the need for double programming and diminish reliance on manual review cycles, programmatic checks, and review checklists.
- Reduce the number of programmers required to produce and validate SDTM data; allowing existing programmers to be deployed to other programming activities.
- Provide reliable measures of quality during the SDTM development life cycle.
- Empower SDTM developers to make better decisions about SDTM data mapping and be assured that all source data is represented in the SDTM data.
- Reduce the impact on downstream analysis datasets and output programming activities that are often subjected to iterative maintenance cycles resulting from incomplete or erroneous SDTM data.

Implementing a technology-enabled, zero-programming approach to real-time validation of SDTM mapping may also provide the opportunity to:

- Provide validated real-time data in SDTM format to drive operational oversight and surveillance of clinical trial data. Although it is not the only consideration for implementing this concept, fluid but robust validation would be required in a surveillance system that uses SDTM data.
- Redefine of the traditional SDTM developer role. The industry typically relies on individuals with strong SAS programming abilities to produce SDTM datasets, however, the approach outlined in this paper can enable recruitment from a broader pool of talent not fully reliant on SAS programming skills.

The application of integrated system-based real-time validation can be explored in other avenues, such as:

- The validation of the ADaM datasets is based on pre-defined rules, study design, and analysis requirements.
- The validation of submission materials such as the Define-XML or even specific content of the data reviewer's guides.

CONCLUSION

The importance of a sound validation strategy for transforming SDTM data is clear. The industry has introduced many methods to produce high-quality SDTM data, most relying on a double-programming approach. However, antiquated methods are repetitive and are not responsive to trial and data changes. Can we do it better? Yes, indeed!

An integrated system that incorporates real-time validation solutions by continuously providing feedback on mapping logic accuracy, SDTM compliance and SDTM completeness would be superior to double programming.

REFERENCES

[1] U.S. Food & Drug Administration, Data Standards Catalog, Rockville, 2022. [Online]. Available: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-catalog-v90. [Accessed 22 March 2023].

[2] CDISC, "Guidance for Ongoing Studies Disrupted by COVID-19 Pandemic", 21 April 2020. [Online]. Available: https://www.cdisc.org/standards/therapeutic-areas/covid-19. [Accessed 22 March 2023].

[3] R. Little, et al, The Prevention and Treatment of Missing Data in Clinical Trials, N Engl J Med 2012;367:1355-60.

[4] CDISC, Analysis Data Model Implementation Guide, 12 February 2016. [Online]. Available: https://www.cdisc.org/standards/foundational/adam. [Accessed 22 March 2023].

[5] M. Meyerovich, L. Hazanov, From EDC to SDTM – faster & better!, in PHUSE Connect, Raleigh, 2018

ACKNOWLEDGMENTS

The authors of this paper would like to thank the Bioforum management team for their leadership and guidance. They helped and encouraged the authors to share their thoughts and ideas on this topic. The authors would also like to thank CDISC and the industry for their ongoing efforts to standardize and improve the quality of clinical trial data. Quality data helps accelerate and improve the drug development process and ultimately makes the world a better, and safer, place.

CONTACT INFORMATION

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

Bremer Louw
Bioforum
bremer.louw@bioforumgroup.com
https://bioforumgroup.com/