Paper DS06

Building High-Quality SDTM Datasets from Real-World Data: Lessons and Solutions

Peishan(Jenny) Han, Bristol Myers Squibb, Lawrence Township, NJ, USA
Lorraine Fang, Bristol Myers Squibb, Lawrence Township, NJ, USA
Lixiao Chen, Bristol Myers Squibb, Lawrence Township, NJ, USA Bristol Myers Squibb,
Po-chun(Kelvin) Lin, Bristol Myers Squibb, Lawrence Township, NJ, USA
Lu Gao, Bristol Myers Squibb, Lawrence Township, NJ, USA

ABSTRACT

Mapping real-world data (RWD) into CDISC SDTM presents significant challenges due to heterogeneous sources such as electronic health records (EHR), claims, and registries, as well as inconsistent or absent coding, nonstandard vocabularies, and variable data quality, etc. This work highlights key barriers and outlines pragmatic strategies to transform RWD into SDTM structures th

We focus on critical considerations, including the application of inclusion/exclusion criteria, construction of study anchors (e.g., RFSTDTC) in the absence of protocol-defined visits, imputation of partial or missing dates, exposure reconstruction, AI-assisted line-of-therapy derivation, and resolution of critical issues identified through P21 validation. We also discuss best practices for building a submission-ready data package and explore how AI may facilitate lab unit harmonization, comorbidity adjudication, and treatment regimen standardization.

Through rigorous curation, terminology alignment, and auditable transformation processes, high-quality SDTM datasets from RWD are achievable. Transparent documentation of assumptions and limitations is essential to ensure reproducibility, regulatory confidence, and the broader utility of RWD in clinical research.

INTRODUCTION

RWD plays an increasingly important role in regulatory submissions and external control arm (ECA) studies. However, unlike clinical trial data—which are collected under controlled, protocol-driven conditions—RWD is inherently heterogeneous. Multiple data sources such as electronic health records (e.g., Flatiron, COTA), registries (e.g., Preamble and Connect-MM), and claims data differ substantially in structure, completeness, coding practices, and timeliness.

In 2023, the FDA issued the industry guidance "*Data Standards for Drug and Biological Product Submissions Containing Real-World Data.*" This guidance emphasized that, starting in late 2025, RWD submitted as study data to NDAs, ANDAs, certain BLAs, and certain INDs must be provided in an electronic format that the Agency can process, review, and archive, unless the submission is exempt or a waiver is granted. While the guidance does not mandate a specific data standard for RWD submissions, mapping RWD to SDTM has been pursued within our organization to support regulatory review and downstream analysis.

This paper presents practical lessons learned from our RWE studies supporting regulatory submissions. These include experiences of harmonizing data from multiple distinct sources, determining line of therapy (LoT), defining inclusion/exclusion (I/E) criteria, standardizing laboratory tests/units and comorbidities, and resolving Pinnacle 21 validation issues. We describe an end-to-end process used in ECA studies and provide recommendations for scaling SDTM development in complex RWD environments.

MAJOR CHALLENGES IN TRANSFORMING RWD TO SDTM

DATA QUALITY ASSESSMENT

RWD often contains a high degree of missingness or lacks key assessments relevant to inclusion/exclusion criteria, baseline characteristics, prognostic factors, and longitudinal outcomes. As a result, substantial effort is required during the early preparatory phase to assess data completeness and fitness for purpose.

LARGE DATABASES AND SUBJECT IDENTIFICATION

ECA studies typically begin with very large RWD databases (e.g., >10,000 subjects) when raw data are not pre-curated for the study. Determining which patients should enter the SDTM pipeline requires early application of high-

level inclusion/exclusion criteria. Applying criteria too late leads to unnecessary processing, while applying them too narrowly risks prematurely excluding eligible patients.

Extensive and Heterogeneous Lab Data

The RWD sources used in our studies often contain hundreds of unique laboratory tests, each with multiple units and naming conventions. Not all tests are relevant to a given indication. Identifying indication-relevant labs and mapping them to standardized LBTEST/LBTESTCD values requires:

- Reviewing free-text test names
- Consolidating variants into harmonized standard tests
- Identifying related tests when required values can be derived by calculation
- Converting units to SDTM-compliant standards

As part of this process, an AI-assisted tool was used to facilitate the review of exported files containing laboratory free-text names and units. However, the AI-generated outputs required iterative refinement to achieve acceptable accuracy. Therefore, manual review is still needed to finalize the mapping and unit conversion.

Free-Text Baseline Assessments

Baseline characteristics such as disease stage, plasmacytomas, CNS involvement, and treatment history are often captured as free text, with inconsistent variable naming across data sources. Harmonization requires predefined consolidation and mapping rules to build a unified, auditable baseline characteristics domain.

Comorbidities Across Multiple Data files

Comorbidities may be recorded across multiple datasets within the same data source, including medical history, adverse events, diagnoses, and hospitalization records. The lack of consistent standard coding complicates direct programmatic derivation.

Inconsistent Missing Date Rules and Abnormalities in Treatment Timing

Different data vendors apply different date imputation rules. RWD also frequently contains implausible dates (e.g., treatment start dates occurring after treatment end dates), requiring manual review or rule-based correction.

Line of Therapy (LoT) Disagreement Across Stakeholders

Clinical teams often disagree with vendor-derived LoT definitions due to:

- Variations in regimen construction
- Inconsistent handling of treatment gaps or overlaps
- Complex post-transplant treatment periods

As a result, additional manual adjudication is required.

Outcomes Distributed Across Many Sources

Outcome data may be distributed across multiple sources, including response assessments, reasons for regimen change, treatment discontinuation records, and clinical outcome files. Notation and structure often differ across these datasets.

Domain Derivation Order Is Not the Same as in Clinical Trials

In clinical trials, SDTM programming typically follows a standard sequence in which inclusion/exclusion and demographics are collected on eCRFs at screening, allowing IE and DM domains to be derived early. In contrast, the complex I/E definitions in RWE studies and the need to determine an index date require adjustments to the order of SDTM domain derivation.

SOLUTIONS AND BEST PRACTICES

Early Alignment on High-Level I/E Criteria

Engage statisticians and clinicians early to define broad inclusion/exclusion criteria that can be applied prior to SDTM mapping, based on data availability and study design. This reduces the subject pool to a manageable size while preserving key study requirements. Data quality should then be assessed on this subset to support customized data extraction requests from vendors.

Identification of Key Lab Tests

Among hundreds of lab tests, focus on laboratory tests relevant to the indication, inclusion/exclusion criteria, and prognostic factors. Create a harmonization Excel file documenting:

- Free-text variants.
- Standardized LBTEST/LBTESTCD values.
- Unit conversion rules.

Integration of Baseline Characteristics

Develop a consolidated baseline domain to capture the baseline characteristics, such as diagnosis, ECOG, disease stage, biomarker, etc.. These assessments are typically recorded only once and is represented in different values/formats in RWD. Distributing them across multiple domains adds avoidable complexity, especially at the stage of ADaM derivation.

Standardization of Treatment Regimens

Treatment regimen names may include generic names, brand names, or free text. Develop an Excel file listing unique regimens and review it with clinicians and statisticians to maintain standardized naming conventions throughout the programming process.

Comorbidity Adjudication via Patient-Level File Review

Applying comorbidity criteria using programming logic is not feasible due to the complexity of the data. Therefore, a patient-level file containing ICD-9 and ICD-10 codes, corresponding verbatim terms, and start and end dates will be extracted into an Excel file. Clinicians and statisticians will review this file to determine subject inclusion or exclusion based on each patient's comorbidity profile.

Manual Adjudication of Line of Therapy

An Excel file containing treatment start and end dates, corresponding outcomes, and relevant transplant information will be created by the programming group and provided to clinicians for manual adjudication. Following clinical adjudication, the following steps will be performed:

Post-process line-of-therapy starting and ending dates.

Deduplicate regimens and alphabetize component drugs within the same line of therapy.

An AI-assisted adjudication tool is incorporated into this workflow; however, both the tool and its associated imputation rules remain under active development. As a result, manual clinical adjudication remains essential in this study despite the use of AI-enabled support.

SDTM Domain Creation Order for RWD

Unlike clinical trial workflows, RWE studies in Oncology/Hematology should begin with line-of-therapy derivation where the qualified line or index date will be determined. Once line of therapy is established, other domains will be derived in following order:

- ZT (Line of Therapy)
- IE (Inclusion/Exclusion Criteria Not Met)
- DM (Demographics)
- LB (Laboratory Test Results)
- ZC (Baseline Characteristics)
- MH (Medical History)
- RS (Disease Response)

The line of therapy-derived index date drives the determination of qualified subjects versus non-qualified subjects.

Strong Documentation Practices

Maintain clear documentation of:

- Lab test selection and mapping
- Lab unit conversion table
- Regimen mapping harmonization
- Line-of-therapy adjudication file
- Subject-level comorbidity file

Explicit Missing Data Imputation Rules

Include detailed date imputation rules in the SAP for:

- Partial dates
- Death date imputation
- Regimen start and end date logic
- Prognostic factor assessment windows

Document any rules not covered by the SAP within the SDTM specifications.

SDTM IG Compliance with Necessary Flexibility

Follow the SDTM IG as closely as possible, while introducing pragmatic flexibilities when required. Examples include:

- Mapping disease diagnosis date to SUPPDM
- Mapping ICD-9/10 codes to SUPPMH when no standard MH variable exists
- Storing MHTERM text exceeding 200 characters in SUPPMH
- Consolidating baseline characteristics into a single customized domain (ZC) rather than creating multiple small domains to improve programming efficiency

CONSISTENT SEQ LOGIC ACROSS SOURCES

For BDS domains, apply identical sorting rules across all datasets to ensure consistent SEQ derivation.

EARLY AND ITERATIVE P21 VALIDATION

Run Pinnacle 21 validation early to identify structural issues and critical warnings before ADaM development begins. Collaborate with statisticians and clinicians to resolve findings efficiently.

CONCLUSION

Transforming RWD into SDTM requires careful planning, cross-functional collaboration, and rigorous documentation. Although RWD presents unique challenges—such as heterogeneous coding, missing or inconsistent dates, and complex treatment histories—systematic processes and early clinical alignment can substantially improve data quality. The approaches described here have been applied in ECA studies within our organiztion and provide a scalable framework for future RWD-based regulatory submissions.

REFERENCES

[a]Data Standards for Drug and Biological Product Submissions Containing Real-World Data https://www.fda.gov/media/153341/download

CONTACT INFORMATION

Author Name: Peishan(Jenny) Han
Company: Bristol Myers Squibb
Email: jennypeishanhan@bms.com