Paper SI07

Transforming RWD for Regulatory Submissions: How to Use SDTM for RWD

Tasha Nagamine, Droice Labs, New York, USA
Sam Hume, CDISC, Austin, USA

INTRODUCTION

FDA's recent guidance on real world data (RWD) for regulatory decision-making represents a significant milestone towards the use of real world evidence (RWE) in regulatory applications. While the Clinical Data Interchange Stands Consortium (CDISC) Study Data Tabulation Model (SDTM) is the standard representation of randomized controlled trial (RCT) data for regulatory submissions, RWD does not neatly align with SDTM domains (e.g., exposures, adverse events), and thus typically requires complex processing steps to prepare it for research use. As a result, tracking the many layers of human decisions and metadata required to organize RWD into analysis datasets is challenging in practice, resulting in opaque lineage and obfuscating patient and element-level auditability. These nuances of RWD require capturing additional information, metadata, and lineage along the data transformation journey to trust RWE for regulatory submissions. In this paper, we demonstrate ongoing work by CDISC and Droice to effectively use SDTM for RWD.

REAL-WORLD DATA STANDARDS FOR REGULATORY SUBMISSIONS - CHALLENGES

Regulatory agencies have recognized the growing importance of the RWD for generating evidence of drug efficacy and safety to support an expanding set of use cases, such as postmarketing surveillance, label expansion, and new drug approvals. FDA has been active in publishing guidance on how to appropriately use RWD for regulatory decision-making from a scientific standpoint, and also in regards to requirements for submission data standards. FDA guidance titled, "Data Standards for Drug and Biological Product Submissions Containing Real-World Data", has indicated that CDISC SDTM, the current standard for FDA submissions, can be implemented on RWD for regulatory use [1]:

- *Currently, and absent a waiver, sponsors submitting clinical and nonclinical study data (including those derived from RWD sources) in submissions subject to section 745A(a) of the FD&C Act are required to use the formats described in the Study Data Guidance and the supported study data standards listed in the Catalog.*
- *FDA recognizes that a range of approaches may be used to apply currently supported data standards (e.g., Clinical Data Interchange Standards Consortium's (CDISC's) Study Data Tabulation Model (SDTM)) to RWD sources such as EHR or claims data.*
- *With adequate documentation of the conformance methods used and their rationale, study data derived from RWD can be transformed to SDTM datasets and submitted to FDA in an applicable drug submission.*

However, there are significant challenges involved in the transformation of RWD into SDTM because information collected for billing and clinical care does not neatly align with the clinical research variables for which SDTM has been designed such as inclusion/exclusion criteria, exposures, covariates, and outcomes (Figure 1). Therefore, attempts to fit RWD into SDTM result in data and information loss from the source.

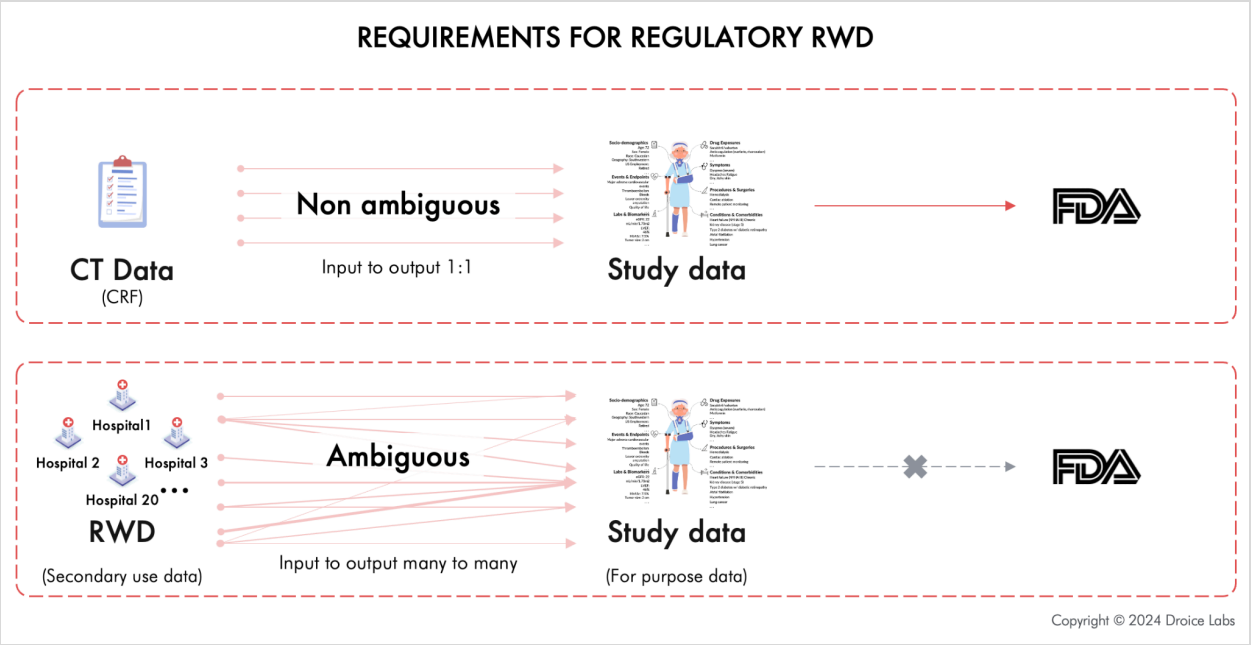

Figure 1: RCT data in case report forms (CRF) maps to study data (e.g. SDTM) by design, whereas it is often ambiguous/unclear how to transform RWD data elements into clinical research data models.

Furthermore, the complexity and diversity of RWD sources, such as EHRs containing tens of thousands of tables in cryptic schemas, significantly complicates the typically manual data transformation process, leading to human errors and omissions of important source data elements and unquantifiable, unaccounted for biases (Figure 2). Therefore, such information loss must be carefully considered when transforming RWD into SDTM for regulatory submissions.

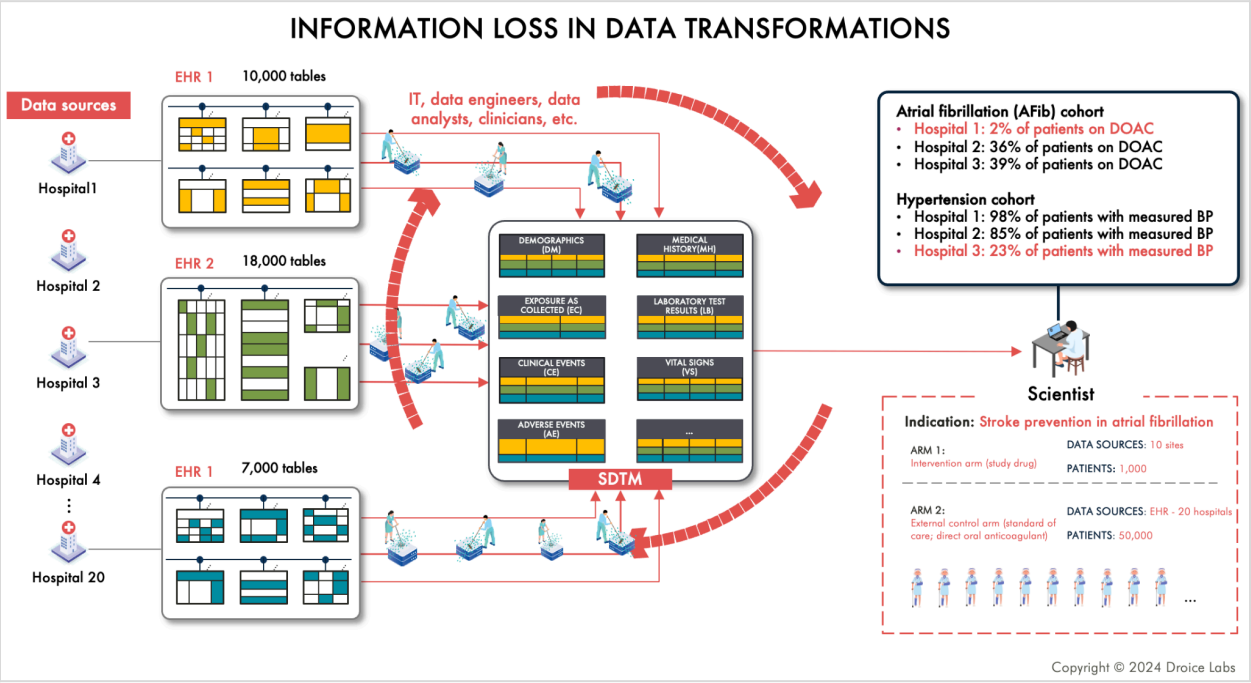

Figure 2: Illustration of the typically manual efforts by teams of data engineers, epidemiologists, and clinicians to transform multiple highly complex RWD sources into SDTM. This process inherently leads to multiple forms of information loss which ultimately create unquantifiable biases in study data that render it unfit for regulatory use.

SOURCES OF INFORMATION LOSS AND BIAS

Transformation of RWD into SDTM is impacted by two major sources of information loss:

1) Data model limitation loss occurs when information in the source data cannot be accurately captured in the target data model/standard [2-6]. Examples of data model limitations:
    a. Clinical study variables (e.g. inclusion/exclusion criteria, exposures, covariates, outcomes) are not directly collected in RWD such as EHR, so computational phenotyping algorithms are used to infer them from source elements such as diagnosis codes, prescriptions, and lab values – such algorithms can be subject to misclassifications and errors
    b. Mismatches and non-matches between RWD terminologies (e.g., ICD-10, SNOMED) and clinical research terminologies used in SDTM (e.g., MedDRA, CDISC terminology) can cause the source information to be altered or lost

2) Data transformation loss occurs due to errors in transformation algorithms [2,8,9] and/or problems in the implementation of transformations on the source data [3,7]. Examples of data transformation loss:
    a. Missing data: locating the required source data elements in complex RWD (such as EHRs) is a human process that can easily miss entire tables or columns containing important information for a study
    b. Transformation errors: bugs and mistakes in human-generated query pipelines, incorrect parsing of target data elements from unstructured and semi-structured fields, and mapping errors can all result in information loss and misclassifications of patient variables

Human decisions on how to handle information loss problems in the transformation of RWD to SDTM can be arbitrary and occur in dataset and study-specific contexts. Metadata around such decisions is often not captured, and different teams may choose to solve the issues in different ways, and thus it can be challenging to reproduce the transformation process across different RWD sources and projects. The ultimate consequence of information loss in transforming RWD into SDTM is that it will not be fit for regulatory decision-making due to unknown levels of unmeasurable errors which can introduce significant and unquantifiable biases (e.g. selection bias, misclassifications of outcomes and exposures) in the analysis dataset.

SUPERLINEAGE™ FOR VALIDATION OF RWD-SDTM TRANSFORMATIONS

To be able to generate reliable RWD in SDTM for regulatory use, additional information is needed to audit source data and quantify the information loss and performance of data transformations in order to calculate error bounds. SuperLineage™ (the superset of Positive and Negative lineage), is a standardized and comprehensive representation of source data containing lineage for each source patient data element that specifies either 1) the location of the element in the output analysis dataset (Positive Lineage), or 2) that the element was not used in the output analysis dataset (Negative Lineage) (Figures 3-4). Furthermore, SuperLineage™ is represented in a standard metadata model that can be submitted along with SDTM in order to be able to use SDTM for RWD. This standardization allows lineage to be combined across disparate data sources and enables unification of downstream tools, workflows, and analytics for validation and audit activities.

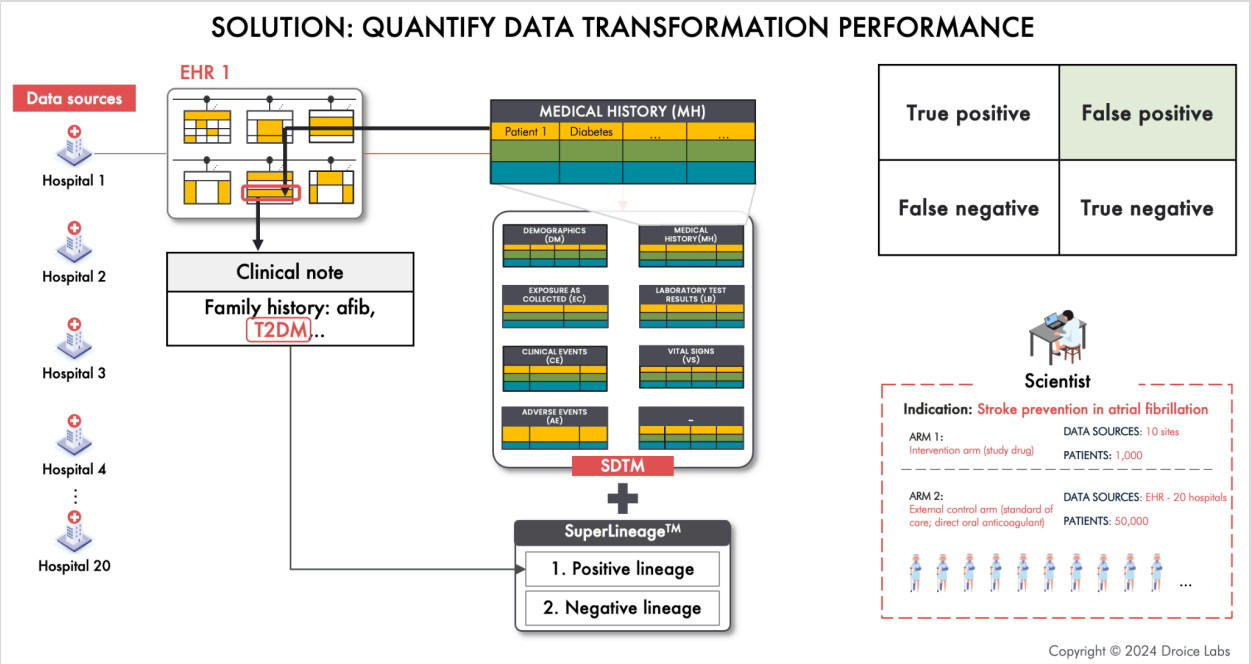

Figure 3: Illustration of using Positive Lineage to validate source data transformed into the output SDTM. In this example, an NLP algorithm did not recognize that diabetes (T2DM) was recorded in the family history and does not positively indicate the patient has diabetes, constituting a false positive in the output SDTM.

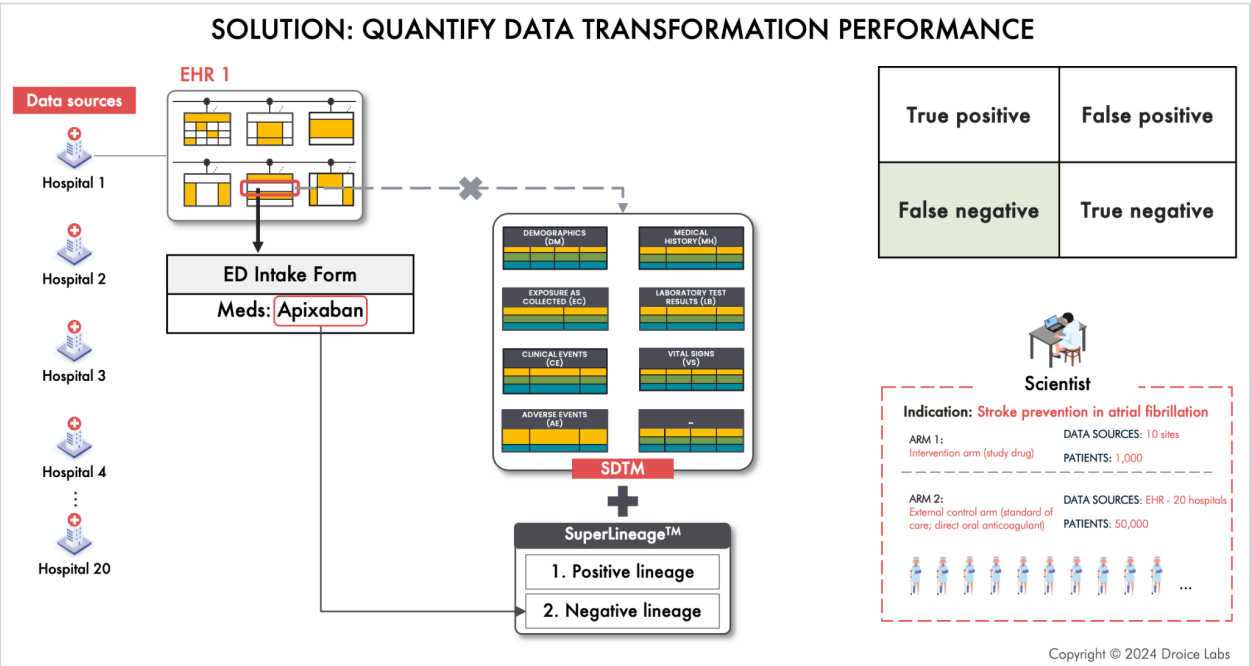

Figure 4: Illustration of using Negative Lineage for source data not transformed into the output. In this example, the ED intake forms were not included as a source for drug exposures, and so the evidence that the patient takes Apixaban on the intake form medication list constitutes a false negative in the output SDTM.

CONCLUSION

While using RWD in regulatory submissions presents challenges for trusting the data stemming from information loss and biases introduced in the transformation process, the solution is to enforce the same type of data quality assessments required in RCTs. RWD in SDTM alone will not have sufficient reliability for use in regulatory submissions. Additional information that comprehensively captures source data for study participants to enable

auditability and quantification of data transformation performance in a standardized way, such as provided by SuperLineage™, should be included along with SDTM datasets to trust the data for high-stakes regulatory decisions.

REFERENCES

1. Data Standards for Drug and Biological Product Submissions Containing Real-World Data
2. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products
3. Choosing Among Common Data Models for Real-World Data Analyses Fit for Making Decisions About the Effectiveness of Medical Products
4. Use of Clinical Data Interchange Standards Consortium (CDISC) Standards for Real-world Data: Expert Perspectives From a Qualitative Delphi Survey
5. An evaluation of the THIN database in the OMOP Common Data Model for active drug safety surveillance
6. Towards quality improvement of vaccine concept mappings in the OMOP vocabulary with a semi-automated method
7. The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform
8. Analyzing the Heterogeneity and Complexity of Electronic Health Record Oriented Phenotyping Algorithms
9. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance

CONTACT INFORMATION

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

Author Name: Tasha Nagamine
Company: Droice, Inc
Address: 31 West 34th Street, Suite 7118, New York City, NY 10001
Work Phone: +1 (929) 224-0049
Email: tasha@droicelabs.com
Website: https://www.droicelabs.com/