PharmaSUG 2026 - Paper PO-415

Traceability in Real World trials- just an aERD away.

Ashwini Yermal Shanbhogue

ABSTRACT

Traceability of data or provenance is an essential requirement for the regulatory review of clinical trial data. Ensuring the ability to trace data from its source to tabulation datasets to analysis datasets to analysis results is therefore one of the priorities of a sponsor submitting said data. In a traditional, gold standard, randomized controlled trial (RCT), where clinical trial data is collected using Case Report Forms (CRFs), one of the ways to ensure and/or demonstrate this traceability is the annotated CRF or aCRF. In recent years however, the increase in availability of Real World Data (RWD) and the evolution of tools available to analyze it has resulted in the rise of a new kind of clinical trial- one that utilizes Real World Evidence (RWE) or evidence that is generated from the analysis of RWD, to assess the effectiveness and safety of a therapeutic product. RWD, however, is collected in electronic databases rather than in CRFs. Currently, there is no regulatory submission document that demonstrates the traceability of RWD present in electronic databases to analysis results. Therefore, I would like to propose an annotated entity relationship diagram (aERD) as a solution to this problem.

INTRODUCTION

Submission of the provenance of clinical trial data to a regulatory agency is just as vital to its review and the determination of the safety and efficacy of a therapeutic product as the submission of information regarding how the data was analyzed (FDA, 2025a). In a traditional, randomized controlled trial (RCT), paper or electronic Case Report Forms (CRFs) are the sources of data. Annotating these CRFs to show which data collection field maps to which Study Data Tabulation Model (SDTM) variable/s or does not map to any SDTM variable (NOT SUBMITTED) ensures traceability of source data, thus making aCRFs (Figure 1) an integral part of the regulatory data submission package.

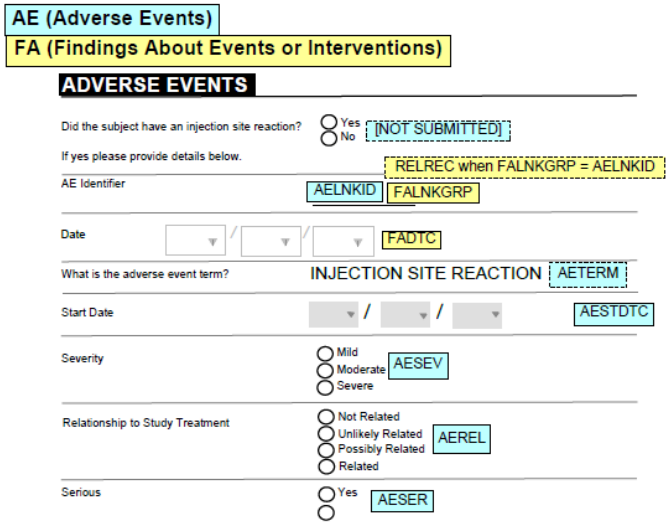

Figure 1 Example aCRF included in the SDTM-MSG_v2.0_Sample_Submission_Package (CDISC 2021)

In recent years however, the increase in availability of Real World Data (RWD) and the evolution of tools available to analyze it has resulted in the rise of a new kind of clinical trial- one that utilizes Real World Evidence (RWE) or evidence that is generated from the analysis of RWD, to assess the effectiveness and safety of a therapeutic product. RWD, however, is collected in electronic databases rather than in CRFs. Currently, there is no regulatory submission document that demonstrates the traceability of RWD present in electronic databases to analysis results. Even if a CRF and aCRF were to be created post-hoc from

RWD, it would not accurately represent the original source of the data and the flow of information from it to tabulation datasets. As a solution to this problem, I would like to propose the creation and submission of an annotated entity relationship diagram (aERD) within the regulatory data submission package for Real World trials.

WHAT IS AN ENTITY RELATIONSHIP DIAGRAM (ERD)?

An ERD is a graphical representation of how components of a database, which are called entities, interact with each other. While there are several ways in which the various pieces that make up the puzzle of an ERD can be defined, for the purposes of this paper, they are defined as follows-

Entity: The tables present within a database. E.g. A 'patients' table within a hospital database that contains demographic information about each patient.

Attributes: Names and data types of the columns that make up a table within a database. These are the properties or characteristics of the table. In the above example, the column name, 'subject_id' and 'INTEGER' data type or the column name, 'gender' and 'VARCHAR(1)' data type are attributes of the 'patient' table.

Relationship: Lines representing how tables and their attribute/s are related to other tables and their attribute/s. For example, the 'patients' table in the above example may be related/ connected/ joined to another table in the database like, a 'labevents' table, through the 'subject_id' column. This relationship may be depicted with a simple line between the two tables, or with the text, 'subject_id = subject_id' on top of it for further clarity. A crow's foot notation may be used to represent the kind of relationship that exists between the two columns. In the example diagram below, the relationship is a one-to-many relationship with the 'one' side being the 'patients' table and the 'many' side being the 'labevents' table.

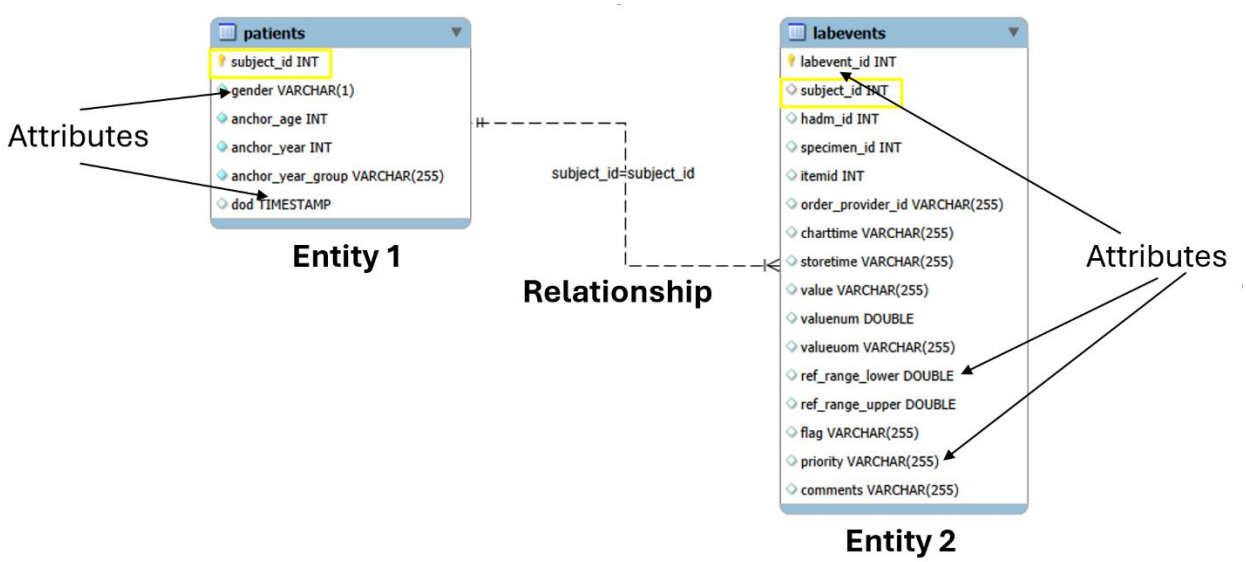

Figure 2 Anatomy of an ERD

Thus, an ERD shows graphically, how all the tables within a database are inter-connected. Annotating an ERD to show which column of an entity/ table maps to which SDTM variable/s or does not map to any variable i.e creating an annotated ERD or aERD, will demonstrate how information flows from these interconnected tables containing RWD to tabulation datasets and onward, all the way to analysis results, ensuring traceability of the Real World Data.

CREATING AN aERD FOR MIMIC IV CLINICAL DATABASE DEMO v2.2

aERDs can be created using various desktop and online tools. I have chosen to create an aERD for MIMIC IV Clinical Database Demo v2.2 using MySQL® Workbench and Adobe Acrobat® Pro.

DATA SOURCE

MIMIC IV is a large, permissibly accessible database containing de- identified electronic health records of patients admitted to the critical care units of Beth Israel Deaconess Medical Center, Boston, MA. MIMIC IV Clinical Database Demo v2.2 (Johnson et al., 2023) is an open access subset of 100 patients from this database, that has been made available on the PhysioNet website (Goldberger et al., 2000) to foster collaboration between researchers and accelerate research progress.

MAPPING TO CDISC SDTM

After downloading the source RWD, I mapped it to CDISC SDTM using Google NotebookLM augmented by manual curation, while following the guidelines provided in FDA (2025a), CDISC (2024a), FDA (2023), CDISC (2019), FDA (2024), FDA (2025b), CDISC (2021), CDISC (2024b), and CDISC (2022).

MYSQL WORKBENCH

MySQL Workbench is a visual tool developed by Oracle, which can be used for SQL development, database modeling, creation, and management among other uses. I imported the MIMIC IV Clinical Database Demo v2.2 data into the Workbench and organized it into two schemas (which is simply a structured way to organize data) corresponding to the two modules, ‘hosp’ and ‘icu’ delineated within the original data source. ‘hosp’ contains data from the hospital wide electronic health record, while ‘icu’ contains data from the information system used within the ICU. I then color-coded the tables to differentiate between tables belonging to the two schemas. During this process, I made sure that all the attributes of the entities, primary keys and foreign keys were accurate and well defined. Doing so meant that I was able to create my ERD using one click of the ‘Reverse Engineer’ Database option of MySQL Workbench. Below is the ERD I created using MIMIC IV Clinical Database Demo v2.2 data. Blue tables belong to the ‘hosp’ schema and green tables belong to the ‘icu’ schema.

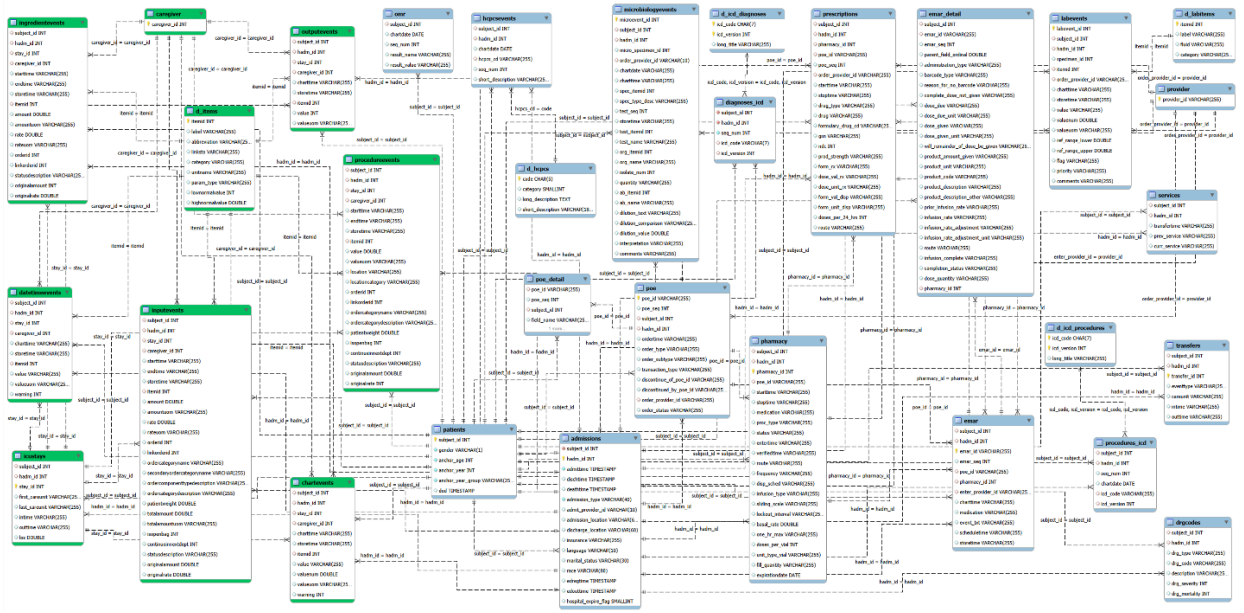

Figure 3 ERD of MIMIC IV Clinical Database Demo v2.2 prior to annotation

MySQL Workbench allows the export of ERDs as PNG, SVG, Single Page PostScript File or as a Single Page PDF. I chose to export mine as a Single Page PDF.

ANNOTATION USING ADOBE ACROBAT PRO

I then annotated the exported PDF using the ‘Add text box’ option within Adobe Acrobat Pro, while following the guidelines from CDISC (2021), a process that is familiar to many within the industry. Figure 4 shows the aERD of MIMIC IV Clinical Database Demo v2.2 data that I have created.

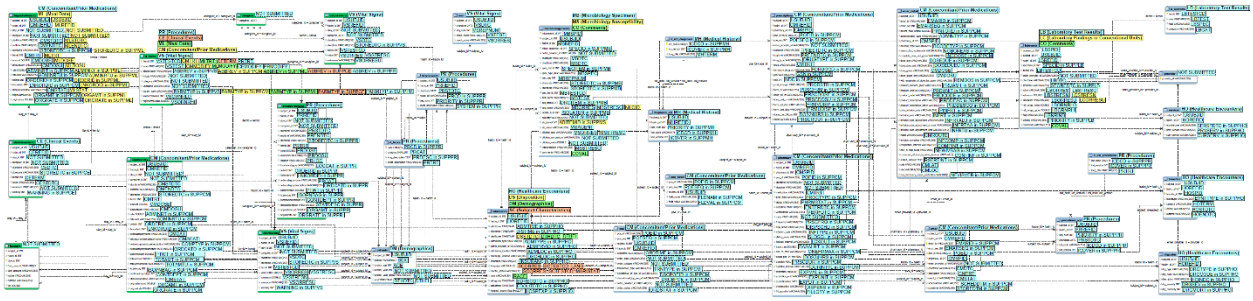

Figure 4 aERD of MIMIC IV Clinical Database Demo v2.2

ADVANTAGES

An aERD can help visualize the flow of data between the inter-connected tables of an electronic database as well as the flow to tabulation datasets, and therefore, to analysis datasets and analysis results. The visualization of the flow of data is particularly impactful in the following cases-

MULTIPLE SCHEMAS

An aERD can help visualize how multiple schemas are inter-connected, how data flows between them and how data flows from multiple schemas to analysis results. The figure below shows how the green tables from 'icu' schema are connected to other green tables within the 'icu' schema and to the blue tables from the 'hosp' schema, and how data is flowing from these inter-connected schemas to tabulation datasets.

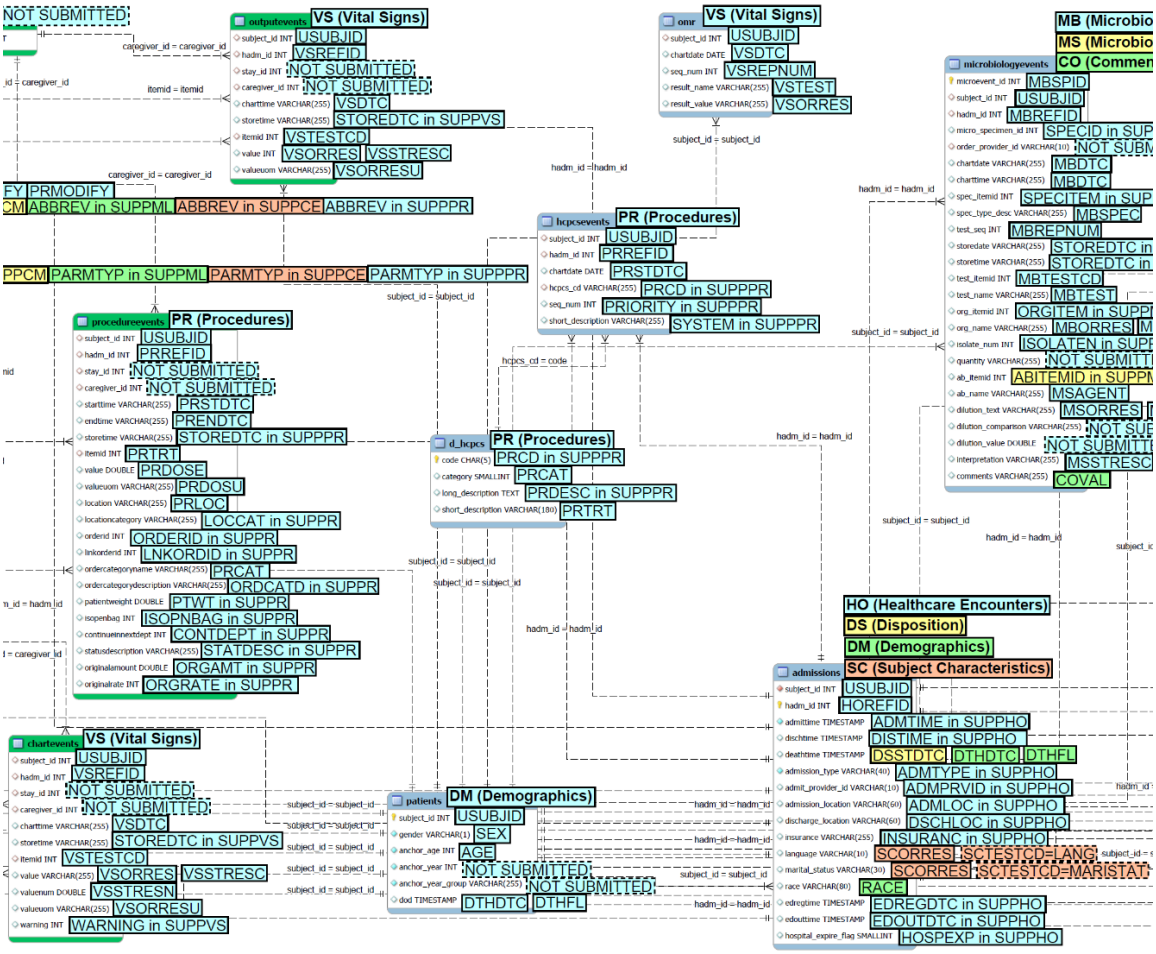

Figure 5 Data flows between multiple schemas and to tabulation datasets

ONE TABLE MAPPED TO MULTIPLE SDTM DOMAINS

An aERD can help visualize how RWD from one table flows to multiple tabulation datasets as shown below.

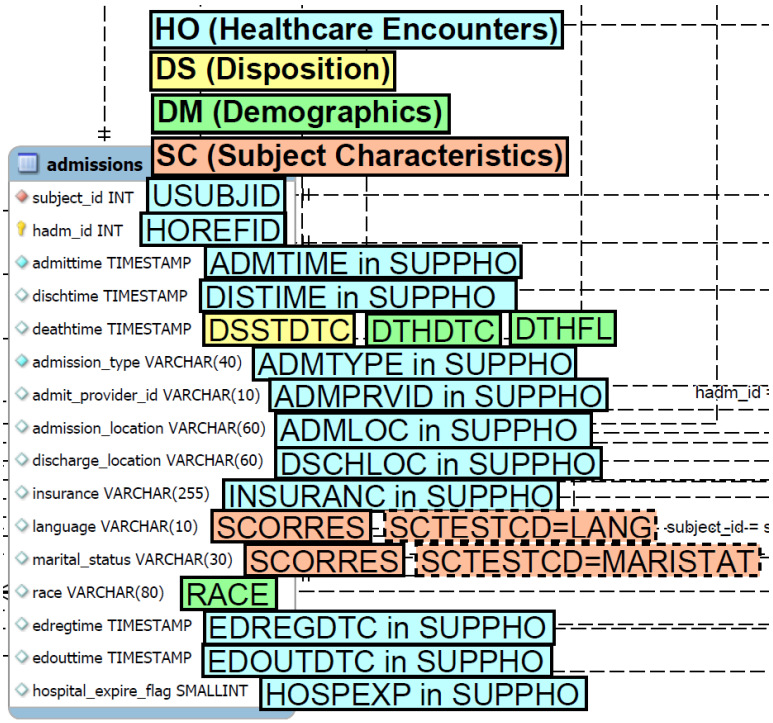

Figure 6 Data flows from one table in the electronic database to multiple tabulation datasets

MULTIPLE TABLES MAPPED TO ONE SDTM DOMAIN

An aERD can help visualize how RWD from multiple tables flows to one tabulation dataset as shown below.

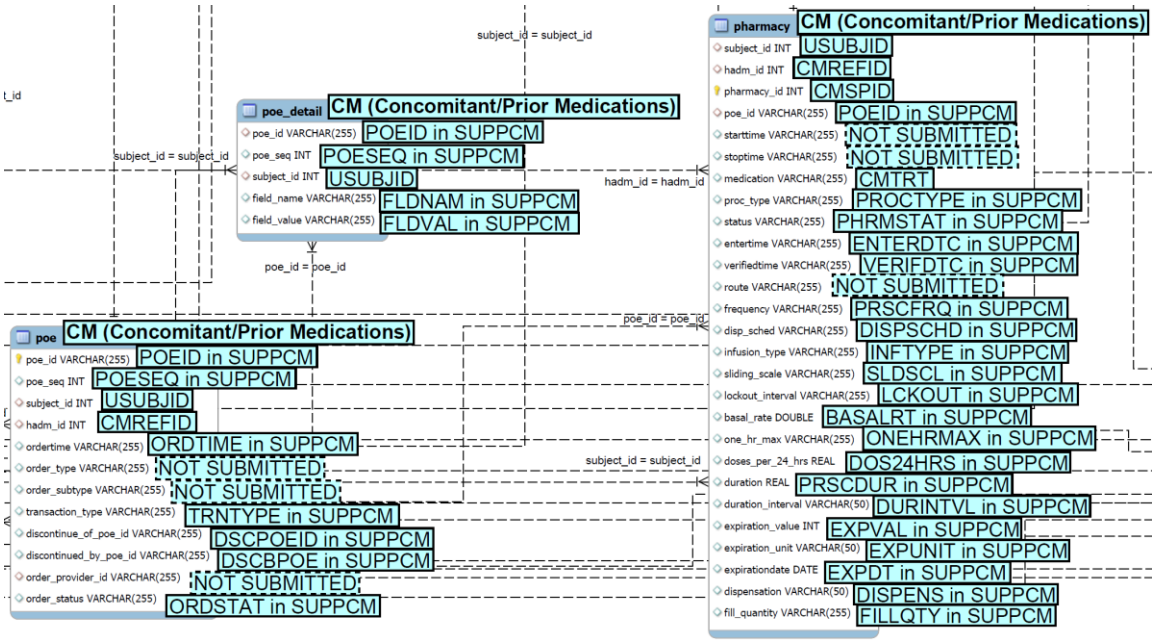

Figure 7 Data flows from multiple tables in the electronic database to one tabulation dataset, CM

CONCLUSION

In this paper, I have demonstrated the creation of an aERD of a database containing RWD and proposed the addition of an aERD to the regulatory data submission package for Real World trials. An aERD will help visualize how data flows between the tables of the source database and from source database to tabulation datasets and onward to analysis results and therefore, ensure reliability and traceability of data in Real World trials.

REFERENCES

1. FDA. (2025a). STUDY DATA TECHNICAL CONFORMANCE GUIDE Technical Specifications Document: Technical Specifications Document. In https://www.fda.gov/. Retrieved March 10, 2026, from https://www.fda.gov/media/153632/download
2. Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). MIMIC-IV Clinical Database Demo (version 2.2). *PhysioNet*. RRID:SCR_007345. https://doi.org/10.13026/dp1f-ex47
3. Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220. RRID:SCR_007345.
4. FDA. (2025b). CBER-CDER Data Standards Program 2024 Annual Assessment. In https://www.fda.gov/. Retrieved March 10, 2026, from https://www.fda.gov/media/187451/download
5. CDISC. (2024a). Considerations for SDTM Implementation in Observational Studies and Real-World Data Version 1.0 (Final). In https://www.cdisc.org/. Retrieved March 10, 2026, from https://www.cdisc.org/sites/default/files/2024-02/Considerations%20for%20SDTM%20Implementation%20in%20Observational%20Studies%20and%20Real-World%20Data%20v1.0.pdf
6. FDA. (2023). Data Standards for Drug and Biological Product Submissions Containing Real-World Data: Guidance for Industry. In fda.gov. Retrieved March 10, 2026, from https://www.fda.gov/media/153341/download
7. CDISC. (2019). CDISC Define-XML Specification Version 2.1 (Final). In https://www.cdisc.org/. Retrieved March 10, 2026, from https://www.cdisc.org/standards/data-exchange/define-xml
8. FDA. (2024). Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products: Guidance for Industry. In https://www.fda.gov/. Retrieved October 24, 2025, from https://www.fda.gov/media/152503/download
9. CDISC. (2021). SDTM and SDTMIG Conformance Rules Version 2.0 (Final). In https://www.cdisc.org/. Retrieved March 10, 2026, from https://www.cdisc.org/standards/foundational/sdtmig/sdtm-and-sdtmig-conformance-rules-v2-0
10. CDISC. (2024b). Study Data Tabulation Model Version 2.1 (Final). In https://www.cdisc.org/. Retrieved March 10, 2026, from https://www.cdisc.org/standards/foundational/sdtm/sdtm-v2-1
11. CDISC. (2022). Study Data Tabulation Model Implementation Guide: Human Clinical Trials Version 3.4 (Final). In https://www.cdisc.org/. Retrieved March 10, 2026, from https://www.cdisc.org/standards/foundational/sdtmig/sdtmig-v3-4
12. CDISC. 2021. “SDTM Metadata Submission Guidelines v2.0.” CDISC. Retrieved January 14, 2026 (https://www.cdisc.org/standards/foundational/sdtm/sdtm-metadata-submission-guidelines-v2-0)

ACKNOWLEDGMENTS

I thank Bhargav Koduru for his support and mentorship.

CONTACT INFORMATION

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

Ashwini Yermal Shanbhogue
ash23shan@yahoo.com
https://www.linkedin.com/in/ashwini-y-shanbhogue/