Paper SA09

Multi-Standard Submission – Process, Learnings and Recommendations for Best Practice

Mukesh Garg, Sumptuous Data Sciences, New Jersey, USA
Sanket Kalyankar, Sumptuous Data Sciences, Pune, India

1. ABSTRACT

This paper presents key learnings and best practices from preparing multistandard submissions using both OMOP and CDISC SDTM. A successful submission starts with optimizing upstream activities—rigorous data cleaning, normalization, and ensuring timestamp integrity—to reduce downstream issues. Mapping to the OMOP Common Data Model (CDM) using standard vocabularies (e.g., SNOMED CT, RxNorm) provides a strong foundation.
Implementing a quality dashboard early in the ETL process helps monitor data completeness, consistency, and vocabulary coverage, enabling proactive issue resolution. This enhances the robustness and traceability of both the ETL and the resulting OMOP dataset.
Subsequent transformation to CDISC SDTM requires careful domain alignment (e.g., AE, EX, DM) and terminology compliance. Early validation with Pinnacle 21 improves submission readiness.
Maintaining audit trails, version control, and cross-functional collaboration ensures semantic consistency. These insights highlight practical strategies for managing multistandard data pipelines and optimizing real-world data for FDA and EMA regulatory submissions.

2. INTRODUCTION

Real-world data (RWD) is increasingly central to global drug discovery and regulatory approvals, with agencies like the FDA and EMA leading its integration to generate robust real-world evidence (RWE) for faster product development, stronger oversight, and lifecycle management. This shift drives the need for multistandard submissions (e.g., CDISC SDTM for eCTD Module 5 and OMOP for RWD), as seen in initiatives like DARWIN EU. However, harmonizing heterogeneous data sources poses key challenges in curation, standardization, and analysis. [1], [2], [5], [16]

- Structural differences: Data from multiple sources/formats, complicating curation and standardization.
- Quality issues: Missing values and inconsistencies create noise, hindering bias assessment.
- Volume overload: Massive datasets overwhelm traditional tools, demanding high-end processing.
- Terminology gaps: Inconsistent dictionaries and missing equivalent terms across sources impede standardization.

This paper articulates our experiences, strategic approaches, and lessons learned in the preparation of multi-standard submission. The following sections will provide details like practical examples, challenges encountered, and solutions applied during the data transformation cycle encompassing RAW, OMOP, and CDISC standards. Furthermore, the paper will offer critical insights into submission readiness, with particular attention to the requirements and expectations of regulatory authorities such as the FDA, EMA, and other global agencies.

3. Understanding Multistandard Data Pipelines

Multistandard Data Pipeline, in this case, consists of OMOP and CDISC standards:
OMOP (Observational Medical Outcomes Partnership) is a common data model developed and maintained by OHDSI community. It is an open community data standard for harmonizing observational data for

evidence generation. Transformation of raw data into OMOP involves ETL (Extract – Transform – Load) activities along with vocabularies like (SNOMED CT, RxNorm, LOINC etc.). Benefits of OMOP standard include standardized structure with data quality framework and support evidence generation processes like characterization, safety surveillance, drug comparative analysis etc.[8]

CDISC standardizes the structure of clinical trial data for regulatory submission. It organizes data into consistent domains (e.g., DM, AE, LB) using controlled terminology to ensure integrity, traceability, and compliance. This standardization enables efficient regulatory review and supports interoperability and transparency across clinical research.[17]

This data pipeline transforms raw data into OMOP and CDISC formats, integrating common data models with regulatory requirements to facilitate drug submissions.

4. Preparing Data for Multistandard Submission

4.1 Data Cleaning and Normalization

Efficient analysis and evidence generation is dependent upon data availability as well as quality which is why cleaning data is of utmost importance. Some of the important and most common cleaning parameters are missing values, consistency, duplicate values, accuracy and outliers.[9]

Our sponsor data was cleaned by keeping the end goal of regulatory compliance and submission into consideration. The better the upstream data preparation, the easier it becomes down the stream for regulatory compliance and successful submission.

Key learnings and suggestions in data cleaning stage:

- CDISC requires date variables in ISO 8601 format, while the RWD has inconsistent formats for date variables, converting all dates into ISO 8601 format at cleaning/curation stage can prevent downstream compliance issues.
- RWD often consists of partial or missing end dates, while CDISC does require start and end dates, defining imputation rules while maintaining traceability can simply transformation as well as achieve downstream compliance.
- Edit checks for maintaining accuracy like end date should not be before start date can help achieve validity.
- Consider vocabularies and controlled terms (CT) mapping from the initial stage for easier transformation with adherence to submission standards.

4.2 Data Mapping

As seen in section 4.1, although data cleaning/curation is an important step to prepare the data for transformation and analysis, a successful submission is not just about quality data but also mapping strategies that include audit trails, detailed documentation to maintain timestamps, integrity, traceability etc. Following are some of the recommendations on submission strategies from our experience:

- Timestamps of data extraction should always be monitored and logged, and it is recommended that checks are in place to determine the latest or the pre-decided snapshot is being used for curation, transformation and subsequently submission.
- Mapping specs with necessary imputations carried out are documented across multiple transformations (raw to OMOP) or (OMOP to CDISC) to maintain traceability.
- Documents like issue trackers and resolutions are of utmost importance and need to maintained for audit trails and integrity purposes.
- Quality checks applied, comments/reasons for data issues still pending post ETL should be documented for better transparency across multiple data cycles.

5. ETL and Quality Management for OMOP

5.1 ETL Design and Best Practices

We built our ETL using a modular approach, as recommended by the OHDSI community, by separating data extraction, vocabulary mapping, and transformation into distinct, well-defined stages. This structure enabled independent design, implementation, and validation of each component, while supporting collaboration among domain experts, medical coders, and ETL developers. The modular architecture improved transparency, maintainability, and reproducibility of the ETL process, consistent with OHDSI best practice guidance. [10] [16]

Considering data extraction as an ongoing process, with multiple data snapshots being used while the study progresses, the ETL should be reusable or reproduceable to minimize efforts of the ETL team and maintain consistency across all the transformations. [11]

The OHDSI recommends validation to be done in a modular approach, similar to building an ETL, to make sure checks are done and findings are resolved at each stage for achieving [12]:

- Consistent and accurate vocabulary mappings. [12]
- Data completeness – majority data is transformed and very little is left out of the ETL. [12]
- Conformance with the OMOP CDM standard. [12]

The scope of our project while transforming raw data into OMOP involved vocabularies like SNOMED, RxNorm etc. as per our sponsors' requirement. Although these are standard and comprehensive dictionaries used across the globe, we still came across many challenges, a couple of them with our proposed solutions are documented below:

- Challenge 1: Lack of 1:1 correspondence for disease concepts
  Solution: Clinically equivalent synonym concepts were identified and mapped to appropriate standard OMOP concepts following review from our experts.
- Challenge 2: Non-English and free-text present in raw data
  Solution: Non-English terms were translated to English, interpreted with domain expertise, and mapped to corresponding standard vocabulary concepts or clinically appropriate synonyms.

Our data quality monitoring strategy combined the use of Achilles and the OHDSI Data Quality Dashboard (DQD) to assess both patient-level characteristics and standardized data quality dimensions, including completeness, conformance, and plausibility. Achilles was used to contextualize DQD findings through patient-level summaries, while DQD checks were parameterized to reflect study-specific characteristics such as age constraints and source vocabulary limitations. This approach enabled the application of study-specific thresholds, recognizing that generic data quality expectations may not always be appropriate for our sponsor's data. [13] [16]

Example - Source data with the developed ETL revealed that several source values did not have one-to-one correspondence with available SNOMED CT concepts like "part-time wheelchair dependent" or "unable to climb stairs", but such milestones are important and cannot be left out of ETLs. In such cases, the standard concept identifier values were set to 0, while the original source value was carried as-is into source_value fields to prevent information loss. Considering the decision and design, the threshold for "*standardConceptRecordCompleteness"* was adjusted to 80%, while "*sourceValueCompleteness"* was maintained at 100%. This adjustment represents the deviation from standard checks, preserved traceability to the source data, and increased confidence in the output, making it fit for use.

6. Transforming OMOP to CDISC SDTM

Mapping and Domain Alignment for key domains: DM, LB

Dataset	OMOP Source	Mapping	Alignment
DM	person table.	SEX → gender_concept_id (mapped to CDISC terminology). RACE → race_concept_id. BRTHDTC → birth_datetime.	Validate against CDISC controlled terms for sex and race.
LB	measurement table (for lab results).	LBTESTCD → Lab test concept (mapped to CDISC terminology). LBORRES → value_as_number. LBORRESU → unit_concept_id. LBDTC → measurement_date.	Ensure proper mapping to CDISC LB controlled terms and units.

6.1 Structuring Rules

- One Record per Observation/Event: Each SDTM domain must represent a single observation or event per row (e.g.one lab test per row). [18]
- Domain-Specific Variables: Follow SDTM variable naming conventions (e.g., LBTESTCD) and include required and permissible variables for each domain.
- Controlled Terminology: Map OMOP concepts to CDISC-controlled terms (e.g., MedDRA for AE, CDISC terms for DM, LB, VS).
- Traceability: Maintain links back to OMOP source tables for audit and regulatory compliance.

6.2 Assumptions

- Complete Mapping Exists: Assumes OMOP concepts can be mapped to SDTM domains without loss of critical meaning (e.g., SNOMED → MedDRA).
- Standardized Units: Assumes OMOP measurements have unit concepts that can be converted to CDISC standard units.
- Event Timing: Assumes OMOP start and end dates are accurate and can be aligned with SDTM timing variables (--STDTC, --ENDTC).
- Population Alignment: Assumes OMOP person table demographics align with SDTM DM domain requirements.

6.3 Important notes:

- Use OMOP's standardized vocabularies (SNOMED, RxNorm, LOINC) as a bridge to CDISC controlled terminology.
- Maintain traceability: Document all mappings and transformations for regulatory compliance.
- Validate SDTM datasets using tools like Pinnacle 21 before submission.
- Ensuring that all SDTM variables use the official, standardized terms published by CDISC for regulatory submissions (e.g, AESEV → "MILD", "MODERATE", "SEVERE")
- Mapping from Source Terminology: When transforming OMOP or other source data, local codes (e.g., SNOMED, LOINC) must be mapped to CDISC terms. This often requires:
    - Crosswalk tables or mapping dictionaries
    - Validation against CDISC Controlled Terminology files (updated quarterly)

6.4 Handling conflicts between OMOP and SDTM vocabularies

Conflict occurs because OMOP and CDISC uses Different Coding Systems, Granularity Differences and Unit & Value Representation differences.
Strategies to handle conflicts include Crosswalk Mapping (Create mapping tables between OMOP concepts and CDISC terms), Controlled Terminology Validation, Normalization Rules, Convert OMOP units to CDISC standard units, Collapse hierarchical OMOP concepts into SDTM-approved categories, Document assumptions and Traceability, Maintain detailed mapping documentation and provenance for regulatory audits and Use of Automated Tools (e.g., OHDSI tools, Rabbit-in-a-Hat) to reduce manual errors. [18]

6.5 Validation and Conformance

Use Pinnacle21 to develop and unit test SDTM Output datasets.

During OMOP → SDTM transformation, the following issues commonly appear in Pinnacle 21 reports:

	Common failures	**Mitigation**
Value-Level Metadata & Controlled Value Issues	LBORRESU not matching UCUM units. VSORRESU missing standardized units.	Map OMOP units to UCUM. Use unit conversion rules where needed. Maintain a units harmonization table.
Missing Required Variables	Missing STUDYID, DOMAIN. Missing required date variables such as AESTDTC, LBORRES, EXSTDTC.	Build ETL rules that enforce required fields. Use default placeholders only when permitted by SDTM rules (e.g., --DTC sometimes can be partial). Backfill from OMOP where possible (birth year → DM).
Cross-Domain Inconsistencies	• Dates in AE not matching EX or DM logically (e.g., AE before study start). Mismatched USUBJID across domains.	• Perform subject-level consistency checks. • Apply cross-domain validation scripts pre-P21. Implement a rules engine for visit/epoch consistency.

6.6 Ensuring Semantic Consistency

Semantic consistency means that the clinical meaning of the data remains intact even when converted across different data models with different vocabularies, structures, and constraints. It would be very important to:

- Maintaining Meaning Across Transformations.
    - Use a Controlled Terminology Crosswalk
    - Preserve Provenance / Traceability
    - Apply Consistent Clinical Interpretation Rules
    - Semantic Validation
- Version Control of Logic and Terminologies
    - Version Control for Vocabularies
    - Version Control for ETL Logic
    - Governance Process for Changes

7. Practical Examples and Lessons Learned

7.1 Challenges and solutions

Problem 1 - Data Loss prevention:
During the transformation of raw clinical data into OMOP and subsequently into CDISC (SDTM/ADaM), a substantial portion of information was embedded in free-text fields such as prescription notes, discharge summaries and physician comments. Although these notes contained clinically relevant details, their visibility was limited due to the sheer volume of unstructured text. This created a risk of data loss and incomplete representation of patient information if not included in the transformation. [16]

Solution: To mitigate this issue, a Python-based text extraction utility was developed. The tool leverages keyword-driven search to identify and isolate relevant free-text segments from the raw data. Extracted notes are then flagged for mapping into OMOP structures and subsequently transformed into SDTM domains, ensuring that critical information is retained throughout the pipeline. This approach not only prevents data loss but also provides a transparent mechanism for traceability, aligning with regulatory expectations for completeness and accuracy in clinical data submissions.

Problem 2 – Bias assessment and mitigation:
Missing, partial, or overlapping pharmacy and prescription dates limit the ability to accurately define treatment exposure windows, resulting in potential exposure misclassification and bias in downstream safety and efficacy analyses.

Solution: Incomplete or partial dates were carried forward into OMOP as-is, with drugs standardized using the RxNorm vocabulary. OMOP allows source values to be retained without modification, preserving fidelity to the raw data. Although, CDISC SDTM requires visit (full) and exposure information, to meet this requirement, partial dates were imputed to full dates, and in cases where stop dates were missing, the stop date was derived as *start date + days of supply*. This made sure that exposure windows were complete and prevented bias in downstream analysis. All assumptions and derivation rules were documented at each stage and included in the Study Data Reviewer's Guide (SDRG) to maintain clarity and transparency for regulators.

- Effective strategies and solutions
- While developing of a robust, reusable ETL and comprehensive raw-to-OMOP-to-CDISC mapping dictionaries helps prevent rework and inconsistencies. Potential non-1:1 mapping should be identified upfront and documented or discussed with regulators as appropriate.
- Preserving metadata and maintaining end-to-end traceability and data lineage throughout the data lifecycle, supported by accurate and complete documentation, ensures transparency and mitigates downstream compliance risks up to and including regulatory submission.
- Validation should be performed at each stage of the pipeline, using OMOP tools (Data Quality Dashboard and Achilles) and CDISC tools (Pinnacle 21) as early as possible to enable early issue detection and prevent downstream rework prior to submission.

7.2 Key Lessons Learnt

- Raw healthcare data offers significant valuable information to aid or be part of clinical trials at a global scale, OMOP acts as a bridge between raw and CDISC.
- Transparency, traceability, data lineage, and auditability are critical throughout the data lifecycle from raw data to OMOP to CDISC and must be preserved as a top priority as part of regulatory compliance.
- Validation should be planned at each milestone and applied as early as possible in the process, rather than treated as an afterthought.
- Ongoing communication with regulatory agencies helps keep them informed and reduces the risk of confusion in multi-standard data pipelines and regulatory submissions.

8. CONCLUSION

As regulatory agencies explore options beyond CDISC to better accommodate the growing volume of real-world data in a more structured and harmonized manner, multi-standard submissions will continue to play an important role in enabling the use of such data. Although they can be complex, multi-standard data pipelines are strategically important for incorporating broader data sources to support more robust safety and efficacy assessments in clinical trials. By applying best practices such as data loss prevention, bias assessment and mitigation, terminology harmonization, and thorough validation, data quality can be maintained and even improved across multiple transformations, supporting successful regulatory submission.

Building on the above, it is encouraging to see regulatory bodies taking steps to accommodate real-world data through initiatives such as FHIR-to-CDISC mapping and the Darwin EU framework. This suggests that the industry is likely to evolve from relying solely on traditional randomized clinical trials (RCTs) toward more inclusive, real-world data–based clinical trials, enabling global studies with enhanced safety and efficacy assessments. Advances in artificial intelligence further support this evolution, as AI can process large volumes of data and generate actionable insights, potentially expediting drug development and approval. However, a cautious approach with appropriate human oversight remains essential to ensure reliability, maintain regulatory compliance, and safeguard patient safety.

REFERENCES

1. https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence
2. Real-world evidence provided by EMA
3. https://darwin-eu.org/
4. https://www.cdisc.org/standards/real-world-data/fhir-cdisc-joint-mapping-implementation-guide-v1-0
5. https://www.ich.org/page/ctd
6. https://www.ijfmr.com/papers/2023/6/10300.pdf
7. https://pmc.ncbi.nlm.nih.gov/articles/PMC8339486/
8. https://www.ohdsi.org/data-standardization/
9. Guo M, Wang Y, Yang Q, Li R, Zhao Y, Li C, Zhu M, Cui Y, Jiang X, Sheng S, Li Q, Gao R. Normal Workflow and Key Strategies for Data Cleaning Toward Real-World Data: Viewpoint. Interact J Med Res. 2023 Sep 21;12:e44310. doi: 10.2196/44310. PMID: 37733421; PMCID: PMC10557005.
10. https://www.ohdsi.org/web/wiki/doku.php?id=documentation%3Aetl_best_practices
11. Quiroz JC, Chard T, Sa Z, Ritchie A, Jorm L, Gallego B. Extract, transform, load framework for the conversion of health databases to OMOP. PLoS One. 2022 Apr 11;17(4):e0266911. doi: 10.1371/journal.pone.0266911. PMID: 35404974; PMCID: PMC9000122.
12. Session 3_J.Duke_OMOP CDM - the Validation Approach
13. Vaclav Papez, Maxim Moinat, Erica A Voss, Sofia Bazakou, Anne Van Winzum, Alessia Peviani, Stefan Payralbe, Elena Garcia Lara, Michael Kallfelz, Folkert W Asselbergs, Daniel Prieto-Alhambra, Richard J B Dobson, Spiros Denaxas, Transforming and evaluating the UK Biobank to the OMOP Common Data Model for COVID-19 research and beyond, *Journal of the American Medical Informatics Association*, Volume 30, Issue 1, January 2023, Pages 103–111, https://doi.org/10.1093/jamia/ocac203
14. OHDSI Achilles - https://github.com/OHDSI/Achilles
15. OHDSI Data Quality Dashboard (DQD) https://github.com/OHDSI/DataQualityDashboard
16. ChatGPT OpenAI. (2026). ChatGPT (GPT-4o) [Large language model]. https://chat.openai.com used for rephrasing sentences.
17. https://www.cdisc.org/
18. https://www.ohdsi.org/wp-content/uploads/2023/10/3-AndersonBriefreport-Wes-Anderson.pdf

CONTACT INFORMATION

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

Author Name: Mukesh Garg
Company: Sumptuous Data Sciences
Address: New Jersey, USA
Work Phone: +1 317-656-9264
Email: Mukesh.Garg@sumptuous-ds.com
Website: www.sumptuous-ds.com

Author Name: Sanket Kalyankar
Company: Sumptuous Data Sciences
Address: New Jersey, USA
Work Phone: +91 9370366186
Email: Sanket.kalyankar@sumptuous-ds.com
Website: www.sumptuous-ds.com