

Authors Introduction



Mukesh Garg: Mukesh serves as the Director of Analysis and Operations at Sumptuous Data Sciences, with over 20 years of experience in pharmaceutical data analysis and reporting. He has experience in generating Clinical Study Report (CSR) datasets and reports across a wide range of therapeutic areas, including Oncology, Cardiovascular, Pain, and Central Nervous System (CNS). Mukesh has also worked in regulatory submission processes, having worked with multiple global health authorities. In addition, he has contributed to Health Economics and Outcomes Research (HEOR) through Real World Data (RWD) cohort analyses. Currently, Mukesh is focused on product development initiatives aimed at enhancing the efficiency of clinical and healthcare data analysis.



Sanket Kalyankar: Sanket is currently working as the Director, Technical Operations at Sumptuous Data Sciences. Sanket has 10 plus years of experience in pharma/healthcare primarily focused on application development using R, and Python. His expertise includes electronic submission processes, and he also has significant experience in data standardization, evidence generation, compliance, healthcare data analytics and implementation of OMOP-CDM and other healthcare data standards. His expertise also includes reporting and addressing clinical/real-world data traceability and transparency issues. Sanket is currently co-lead of Phuse Working Group – “Using OMOP and other Real-World data standards for regulatory submissions”. Sanket is based out of Pune, India.



Best Practices for Multistandard Submissions Using OMOP and CDISC SDTM

Mukesh Garg, Director, Analysis and Reporting, Sumptuous Data Sciences
Mukesh.Garg@sumptuous-ds.com

Sanket Kalyankar, Director, Technical Operations, Sumptuous Data Sciences
Sanket.Kalyankar@sumptuous-ds.com



Disclaimer: The views represented in presentation are author's view and do not necessarily reflect the views of their organization.

Background and Key Challenges

Background

As regulatory bodies consider incorporating real-world data (RWD) into drug approvals, validated RWD backed by robust evidence can speed up development while enhancing oversight.

While CDISC is the official regulatory standard, it alone doesn't capture the full RWD picture. Recommended: Multi-standard approach—Common Data Model first, followed by CDISC submission—for better coverage and faster approvals.

Key Challenges

Structural variabilities and inconsistencies

Semi-reliable data quality in most cases - missing data, inconsistent coding etc.

Large volume requiring advanced processing capabilities

Vocabularies and terminology requiring clinical expertise for harmonization



Objective

The objective of this presentation is to articulate 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.

Overview OMOP – CDISC standards

OMOP

Overview: OMOP, developed by OHDSI, standardizes observational data via ETL processes using 130+ vocabularies (e.g., SNOMED CT, RxNorm). Open-source tools support ETL and evidence generation.

Key Benefits: Enables data quality framework; powers analyses like characterization, safety surveillance, and drug comparisons.

Process: Raw data → ETL transformation → OMOP structure for harmonized RWD insights.

CDISC

Overview: CDISC, global standard for structuring and submitting Clinical trials tabulations to regulatory agencies. It defines consistent domains and controlled terminologies, that enables uniformity, traceability and regulatory compliance.

Key Benefits: Consistent and standardized outputs enables efficient regulatory review.

Process: Raw data → SDTM → Apply CT and Metadata → Create submission package → Regulatory submission



Data Preparation

Data Preparation for OMOP & CDISC

Step 1: Detect/resolve quality issues for standards compliance.

Step 2: Fix missing/inconsistent data for reliable RWD submission.

Key Learnings from our project

- CDISC mandates ISO 8601 date formats (e.g., YYYY-MM-DD) for variables like --DTC in SDTM datasets. RWD sources often use inconsistent formats (e.g., MM/DD/YYYY from EHRs), so early standardization during curation prevents validation errors in submissions.
- CDISC datasets like EX or AE require complete start (--STDTC) and end (--ENDTC) dates. RWD frequently has missing ends (e.g., ongoing events), so imputation rules with audit trails (e.g., carrying forward start date or using "ongoing" flags) ensure compliance while maintaining traceability.
- Date Logic: Enforce end date $\geq$ start date via edit checks.
- Early Mapping: Align vocabularies/CT from start to avoid downstream issues.

Data Preparation - Continued

- Accurate timestamps and event sequencing – Recommended to use the latest or filtered data snapshot for curation, transformation and subsequent submission
- Strategies for verification and correction – maintaining documents like mapping specs, issue tracker with resolutions are very important for traceability
- Aligning disparate sources to standards – robust, reusable ETLs help in establishing and maintaining provenance even if more than 1 data standard is involved
- Documenting metadata for traceability – maintaining robust documentation like mapping specs with accurate transformation logics throughout the data lifecycle helps in traceability and most possible accuracy



ETL and Quality for OMOP transformation



ETL (Extract-transform-Load) was developed using the following pointers:

- Robust ETL for reproducible results with high accuracy and consistency in vocabulary mappings
- Prevent data loss by optimal use of Observation table in OMOP
- Effective use of quality dashboard (DQD) to track and resolve quality/inconsistency issues

Challenges in building robust ETLs and our approach to resolve these are provided 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.

ETL and Quality for OMOP transformation

Data Quality for OMOP was measured via following points:

- Metrics: completeness, consistency, vocabulary coverage
- Early dashboard implementation and case examples

An Example of handling non-mapped source values are as follows

Preserving Critical

Milestones: Source values like *“part-time wheelchair dependent”* or *“unable to climb stairs”* lacked one-to-one SNOMED CT mapping but were retained in source_value fields to avoid information loss.

Adjusted Completeness

Thresholds: Standard concept completeness set to **80%**, while source value completeness maintained at **100%**, ensuring traceability and confidence in ETL outputs.

Fit-for-Use Outcome:

Deviation from standard checks preserved data integrity, improved transparency, and strengthened reliability of the evidence generated.

Transforming OMOP to CDISC SDTM

Mapping and Domain Alignment for key domains (DM, LB, CM) was done, with the following key points to consider:

- Structuring rules
- OMOP vs CDISC conflict resolution
- Ensuring Semantic consistency

Structuring rules:

- One record/observation
- Domain specific required and permissible variables
- Map OMOP concepts to CDISC controlled terms
- Maintain links back to OMOP source tables for traceability

Conflict resolution:

- Crosswalk mapping: Create mapping tables between OMOP concepts and CDISC terms
- CT validation
- OMOP units to CDISC std units conversion

Semantic Consistency:

- Use a Controlled Terminology Crosswalk
- Apply Consistent Clinical Interpretation Rules
- Version Control of Logic and Terminologies

Practical Examples

Data Loss Prevention

Free-text fields (notes/summaries) in raw data hold key details but risk loss during OMOP/CDISC transformation due to volume and limited visibility.

Use NLS techniques with a specialized python-based text extractor to find relevant free-unstructured text from data and include in the ETL to prevent data loss

Bias Assessment and Mitigation

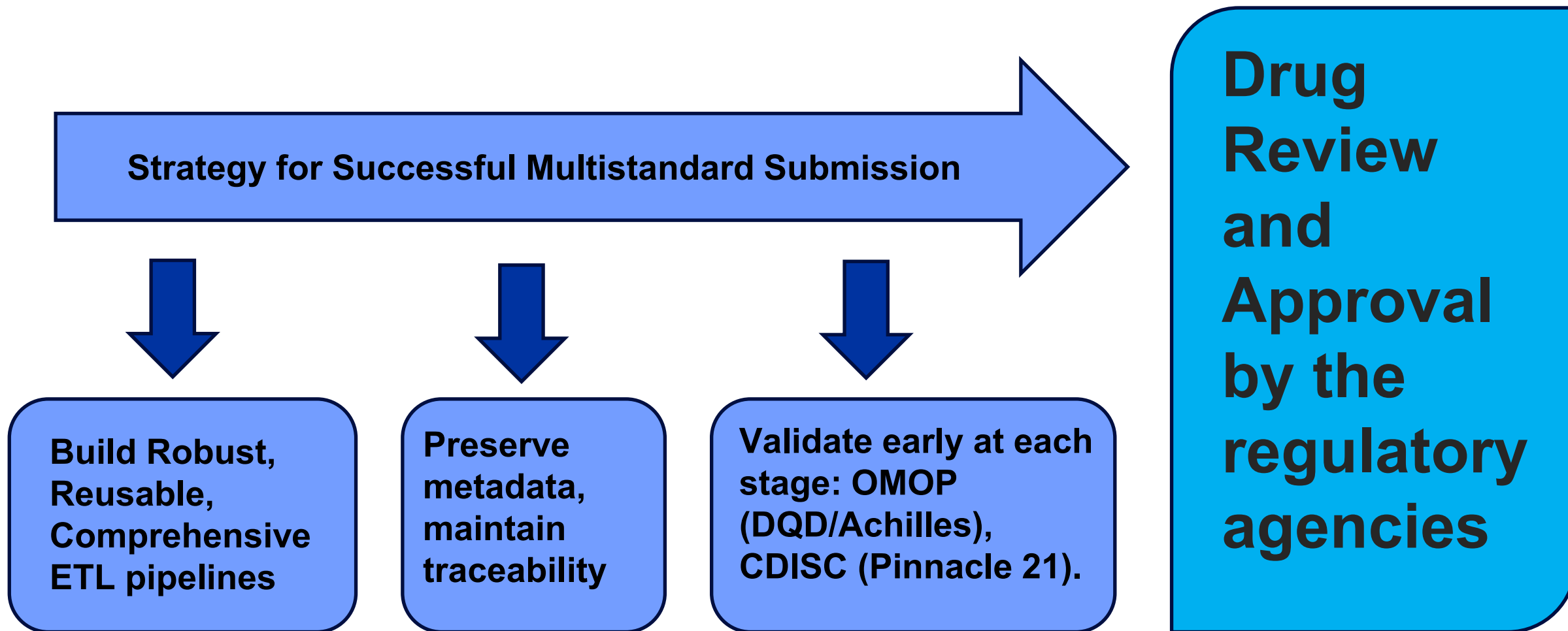
Missing, partial, or overlapping pharmacy/prescription dates limit the ability to define treatment exposure windows, resulting in potential exposure misclassification and bias in downstream safety and efficacy analyses.

- Partial dates carried to OMOP as-is (RxNorm standardized), preserving source fidelity.
- For CDISC SDTM: Impute to full dates; derive missing stop dates as start + days supply for complete exposure windows.
- Document all rules in SDRG for transparency.

Key Learnings and Recommendations

1. Raw healthcare data fuels global clinical trials, with OMOP bridging raw inputs to CDISC standards.
2. Maintain transparency, traceability, lineage, and auditability from raw data → OMOP → CDISC for regulatory compliance.
3. Validation should be planned at each milestone and applied as early as possible in the process, rather than treated as an afterthought.
4. Effective and ongoing communication with regulatory bodies

Effective Strategies and Process Recommendations





Conclusion



- Despite its complexity, real-world data and multi-standard pipelines are being increasingly explored by regulatory authorities. By using best practices, carefully handling known issues, and applying emerging methods to address data loss, bias, terminology harmonization, and transformation validation, data quality can be preserved across standards, enabling successful submissions. Ongoing initiatives such as FHIR-to-CDISC mapping and Darwin EU show that regulators and industry are moving toward a more inclusive, multi-standard submission ecosystem.



THANK YOU

For any questions, please contact:

Mukesh.Garg@sumptuous-ds.com

Sanket.Kalyankar@sumptuous-ds.com



**The Global Healthcare
Data Science Community**

Contact Channels

✉ office@phuse.global
✉ phuse.global

Social Media

✕ [@phusetwitta](https://twitter.com/phusetwitta)
f [/phusebook](https://facebook.com/phusebook)
▶ [/phusetube](https://youtube.com/phusetube)
in [/company/phuse](https://company.phuse.com)