Real World Data, Common Data Models – advantages and challenges

Daniel Rosenberg
Executive Director, Global Epidemiology

November 17th, 2024
ISPOR RWE SUMMIT

Johnson&Johnson
Innovative Medicine

Disclosure

- Daniel Rosenberg is an employee of Actelion Pharmaceuticals Ltd., Johnson and Johnson.

Agenda

- ❑ Background - why do we need CDM for Real World Data, different options
- ❑ Advantages – overcome heterogeneity
- ❑ Challenges in use– provenance, data quality, example
- ❑ Considerations for fit-for-purpose & feasibility data source assessment, example

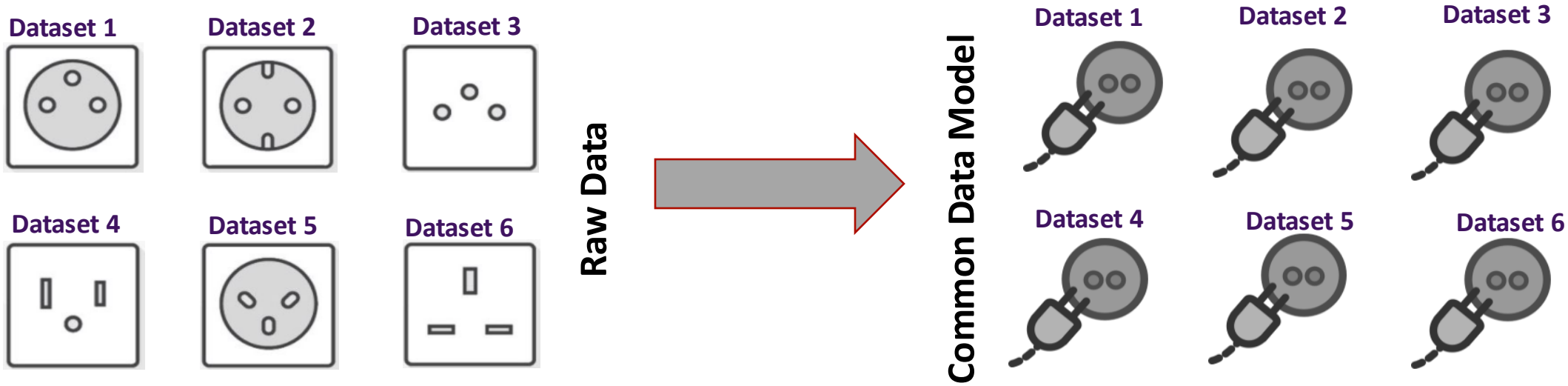
Background

Across patients and healthcare stakeholders, there is an increasing demand for trustful, transparent and reproducible Real World Evidence (RWE)

RWE generation may be restricted due to data scattered across geographic regions, heterogeneous health care settings, limited data accessibility or small patient numbers

Federated analytical approaches as well as pooled analysis will support addressing key research questions

Transforming datasets into a Common Data Model (CDM) will facilitate research under common protocol and standardized analysis plan, facilitate reproducibility and shorten time to results



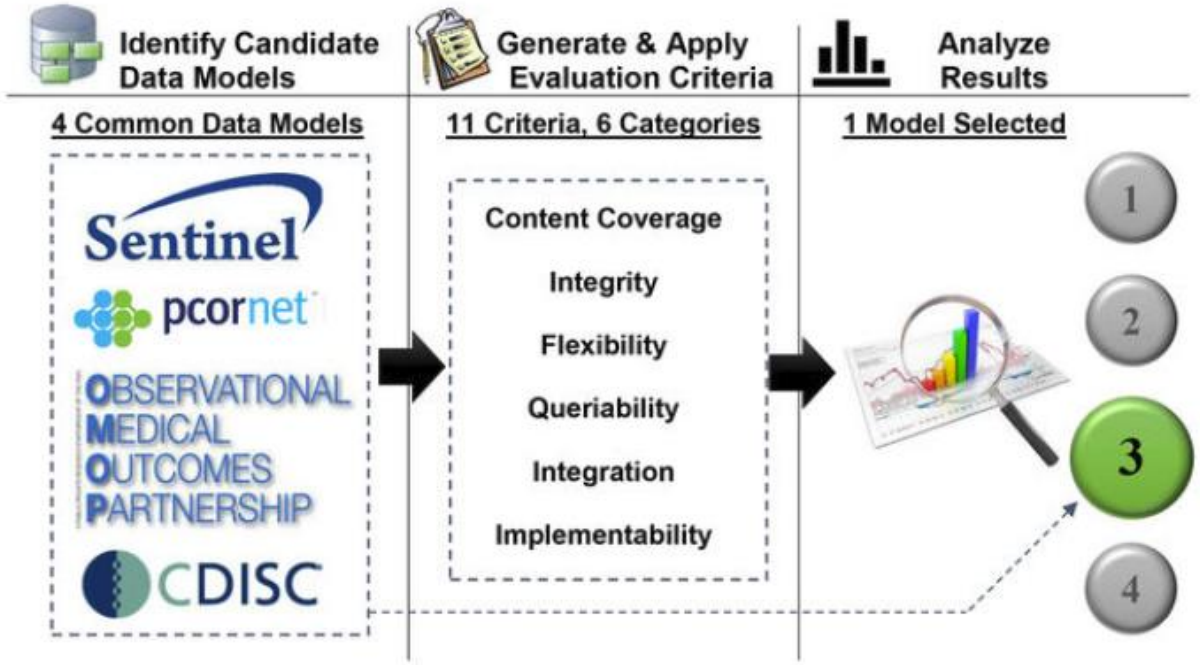
Performance of different CDMs

CDMs are developed for a <u>specific purpose</u>

- data source e.g. data generated for administrative reason vs. study design driven data collection

Fitness for use depends on how closely the CDM matches the planned use

- Specific research question
- Transparency, reproducibility, assess validity
- Completeness, integration
- Ease of implementation and speed of use



Garza M et a.. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform. 2016;64:333–41.

Schneeweiss S et al. Choosing among common data models for real world data analyses fit for making decisions about the effectiveness of medical products. Clin Pharm & Therapeutics 2020; 107(4): 827-833.

Key learnings: Mapping example from STDM to the OMOP CDM (1/2)

To prepare for research questions in an federated data network three registries databases were mapped from STDM to OMOP CDM

OMOP CDM

- Can be used for registry data without loss of essential information
- Open community data standard, with frequently updated vocabularies (OHDSI standardized) and analytic tools

From STDM to OMOP CDM it is essential to define and document imputation rules, in detail

- These imputation rules are critical to know at time of reseach project
- Custom mapping is a key component in registry mapping as there are more clinical data: not all diagnoses, medications, etc., may have a SNOMED code
- Regular mapping refeshing to newer OMOP versions is recommended even for completed registries

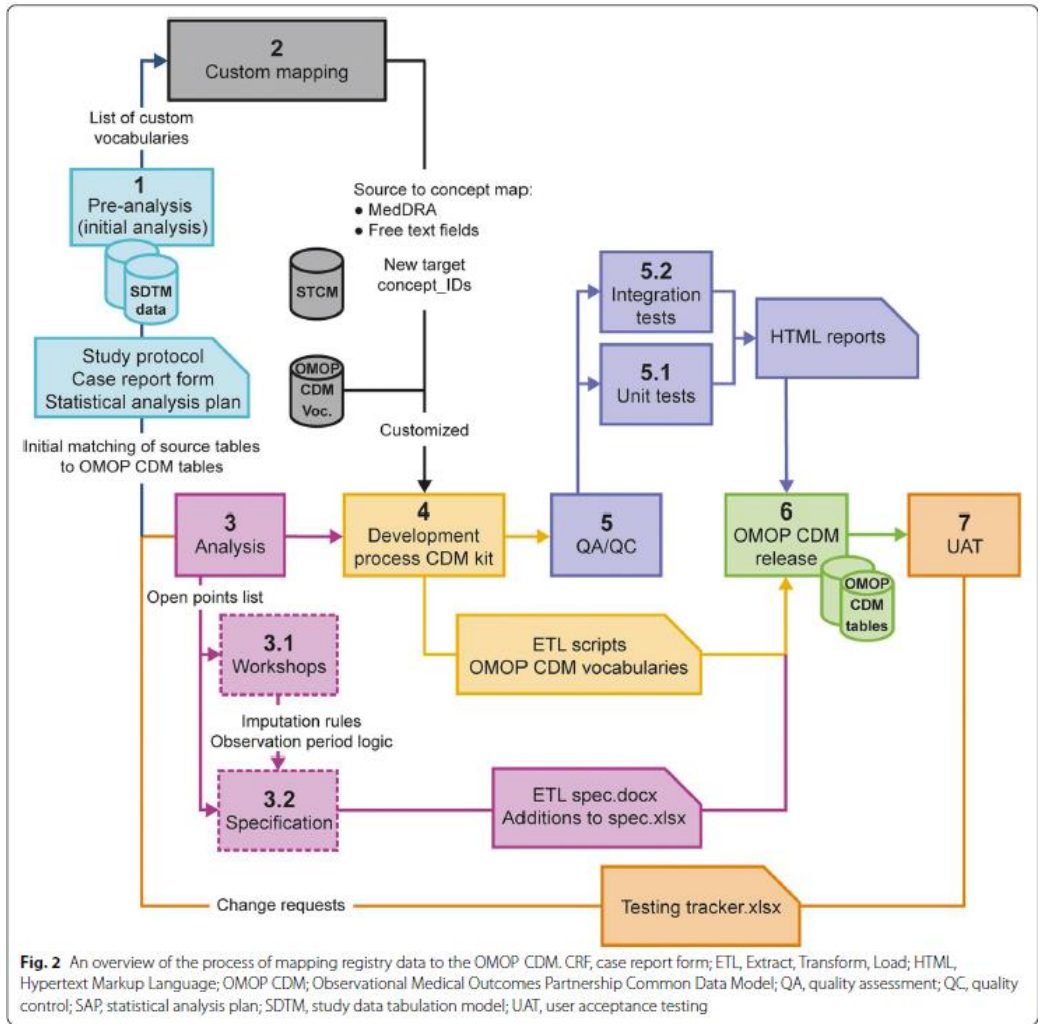


Fig. 2 An overview of the process of mapping registry data to the OMOP CDM. CRF, case report form; ETL, Extract, Transform, Load; HTML, Hypertext Markup Language; OMOP CDM; Observational Medical Outcomes Partnership Common Data Model; QA, quality assessment; QC, quality control; SAP, statistical analysis plan; SDTM, study data tabulation model; UAT, user acceptance testing

Biedermann P, Ong R, Davydov A, Orlova A, Solovyev P, Sun H, Wetherill G, Brand M, Didden EM. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. BMC Med Res Methodol. 2021;21:238.

Homogeneous registry datasets / data after mapping ? (2/2)

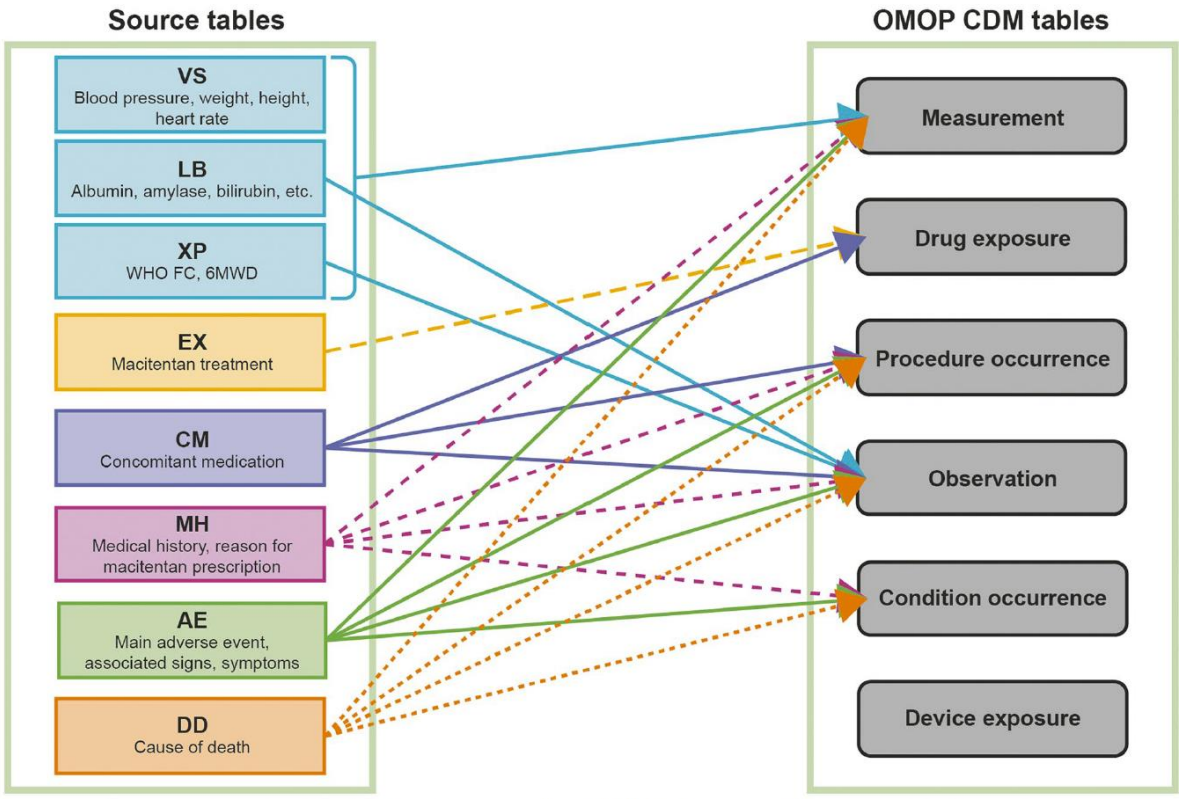


Fig. 3 Example mapping from the OPUS SDTM-format database to the OMOP CDM. For instances where source tables have multiple options for OMOP CDM tables, a decision on which OMOP CDM table to store the record in is decided on a case-by-case basis using OMOP CDM conventions, if available, or customization. AE, adverse event; CM, concomitant medication; DD, death details; EX, exposure to study medication; LB, laboratory data; MH, medical history; OMOP CDM, Observational Medical Outcomes Partnership Common Data Model; VS, vital signs; XP, pulmonary arterial hypertension

- **Expectation**: all three datasets are now homogeneous
  - Data are mapped to the same concepts
  - Data can be found for all datasets in the same OMOP tables

- **Reality**: the datasets are the result of registries which were set up for a purpose
  - Data collection (case record forms) were tailored to the research question of the respective registry
  - Data that were collected systematically in one registry were not in the other registry
  - Source tables may have been slightly different leading to potential risk of mapping to a different OMOP CDM table

The **data format is homogeneous** but the **data itself remain heterogeneous** as the origin provenance of the datasets were PASS (drug or disease registries) in different geographies, centralized vs decentralized health care systems.

Biedermann P, Ong R, Davydov A, Orlova A, Solovyev P, Sun H, Wetherill G, Brand M, Didden EM. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. BMC Med Res Methodol. 2021;21:238.

Fit for purpose RWD framework

Along evidence generation cycle

Real-World Study design

1. Formulate research question
2. Design the study

Real-World Data sources fit for purpose assessment

3. Identify the universe of data sources
4. Narrow down to manageable number of data sources
5. Conduct feasibility assessment (qualitative / quantitative) for each dataset
6. Review research question against feasibility assessment and if needed, repeat until a good match found

Implementation

7. Use feasibility results to inform study protocol, analysis plan
8. Document each step in the final report

Feasibility Assessment across datasets – Considerations

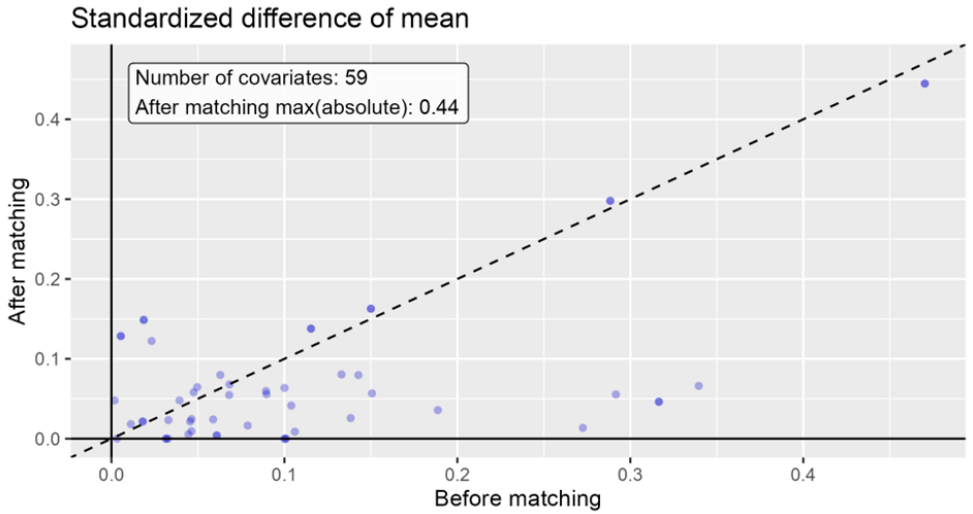
Evidence needed	Decisions supported
Data provenance, data collection for each dataset	Ability to trace back data/audit to source record How representative is the patient population? How generalizable will the results be? How comparable are the databases (meta-analysis?)?
Number of eligible patients	Are there enough patients to move forward?
Key characteristics of eligible patients	Which stratifications would be feasible? How heterogeneous are the patient cohorts within and between the databases? ➢ Are data pooling or meta-analysis feasible
Variable availability, accuracy, completeness, and capture for each dataset	Which variables can be used for the current research question / analysis?
Treatment information (treatment capture, types of treatments, level of detail, start and end dates)	Does the database have sufficient information on treatment? How much can we go into detail with our analysis?
Distribution of year of diagnosis, year of treatment start, and year of enrollment	Do we need to fine-tune our eligibility criteria? How should we define time windows? Which temporal stratifications would make sense?

Use case – Rare Disease – 4 datasets in OMOP-CDM

Table 1: Main characteristics of the data sources

ID	Design	Aim	Patient count	Study period	Region
1.	Prospective observational chohort study	Patients' characteristics, outcomes and safety	2674	April 2014 – June 2020	North America
2.	Retrospective chart review	Patients' characteristics and safety	3031	October 2013 – March 2017	North America
3.	Prospective observational chohort study	Patients' characteristics and outcomes	829	November 2016 - September 2021	North America
4.	Prospective observational chohort study	Patients' characteristics, outcomes and safety	2354	September 2017 - November 2021 (last available data cut)	North America and Europe

Figure 1: Pooled dataset, Standardized mean difference before and after matching



Key Learning: Importance of robust feasibility assessment to identify heterogeneity, in advance of decision to conduct the full study

EM Didden et al. Aggregating and harmonizing registry databases for comparative analyses – lessons learnt. https://www.ohdsi.org/2024showcase-101/

Summary

- The OMOP CDM – one of the best performing CDM for observational health data
  - leads to harmonized data formats
  - allows for standardized programming code and outputs

AND

- **Qualitative** and **Quantitative Feasibility Assessment** specific to the study's reseach question
  - Is important to assess heterogeneity of underlying source data
  - Is critical for reliable, reproducible and tranparent results

Acknowledgements

- Colleagues at Global Epidemiology, Innovative Medicine, Johnson & Johnson
  - Notably Monika Brand and Eva-Maria Didden