PHUSE EU Connect 2024 DS04

From Source to Submission: **Getting the Best of Multiple Standards**

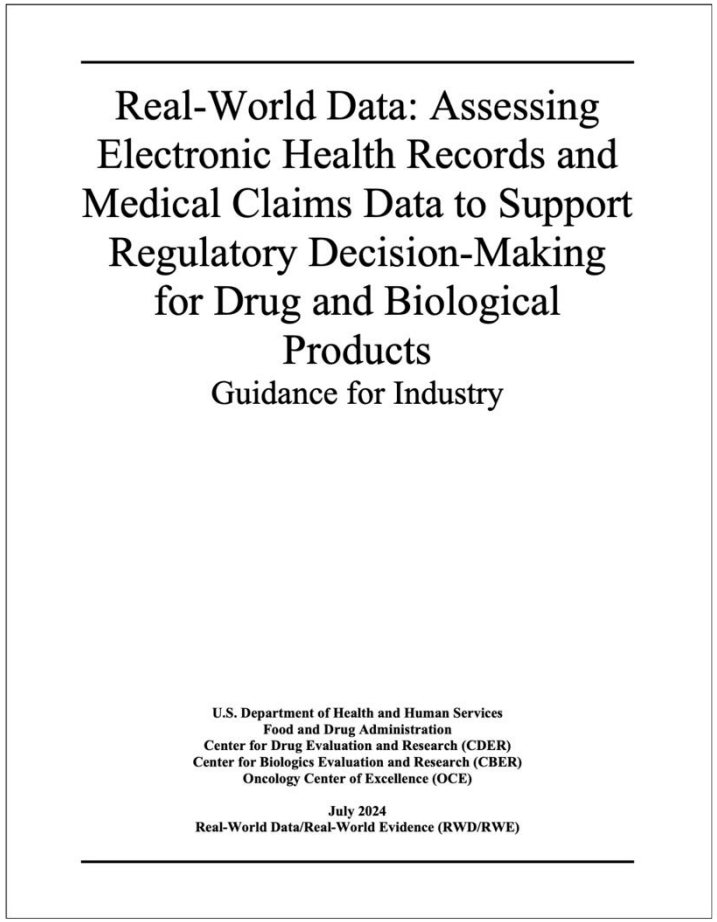
Berber Snoeijer, ClinLine
Jules van der Zalm, OCS Life Sciences

Content

- Introduction
- Data Standards and approach
- Mapping automation
- Defensive programming
- Experiences
- Conclusion

Real-World Data and Standards

- FDA's Definition of Real-World Data:
  - Data that is related to patient health status or the delivery of health care routinely collected from a variety of sources.
- Variety of Electronic Health Record Sources:
  - General Practitioner
  - Hospital / Clinic
  - Pharmacy
  - Laboratory
  - Other



Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products

Guidance for Industry

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)
Center for Biologics Evaluation and Research (CBER)
Oncology Center of Excellence (OCE)

July 2024
Real-World Data/Real-World Evidence (RWD/RWE)

https://www.fda.gov/regulatory-information/search-fda-guidance-documents/real-world-data-assessing-electronic-health-records-and-medical-claims-data-support-regulatory

Main concerns for using Real-World Data

- Can we access the source data?

  => Privacy and Governance

- Do we know how data is collected?

  => Provenance

- Do we have documentation, audit trials and data security in place?

  => Traceability, Consistency, Reliability

- Is our data fit for purpose?

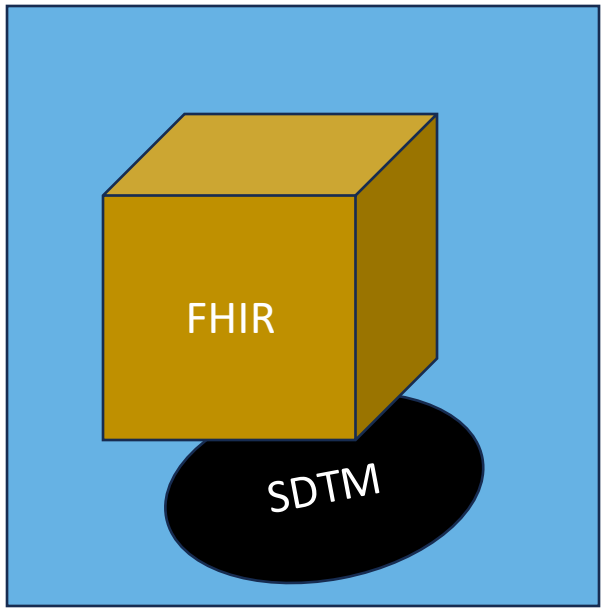
  => Representative, Available, Qualitative

How to assess fit for purpose?

- Know your source!
  - What information is collected and what not?
  - How is information collected?
- Standardize!!
  - Source Documentation
  - Data structure
  - Controlled terminology
- Analyse
  - Representativeness
  - Availability of co-variates, outcomes and safety information
  - Matching

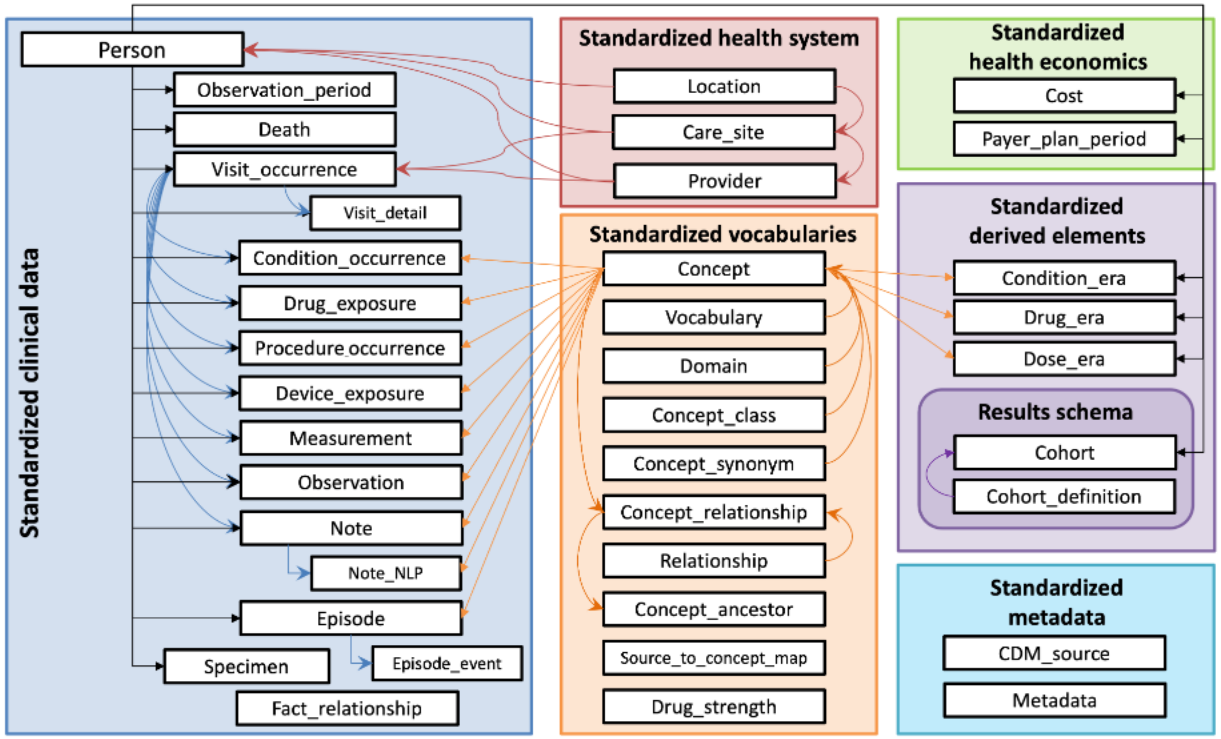
What data standard to use for real-world data?

- FHIR
  - Flexible, flat JSON format
  - 1 file per patient
  - Focus on routine health care information
  - Standard extraction tools are not sufficient
    - Not full set of (coding) information
    - Still a lot of input specifications needed (given the FHIR flexibility)
- SDTM
  - Clinical trial focussed
  - Not all traceability information can be included
  - Not all information needed for fit for purpose will be used for submission
  - Not all information expected for SDTM is available in EHR source
  - Required for submission
  - Clinical study design might not be complete while it is starting point for SDTM creation



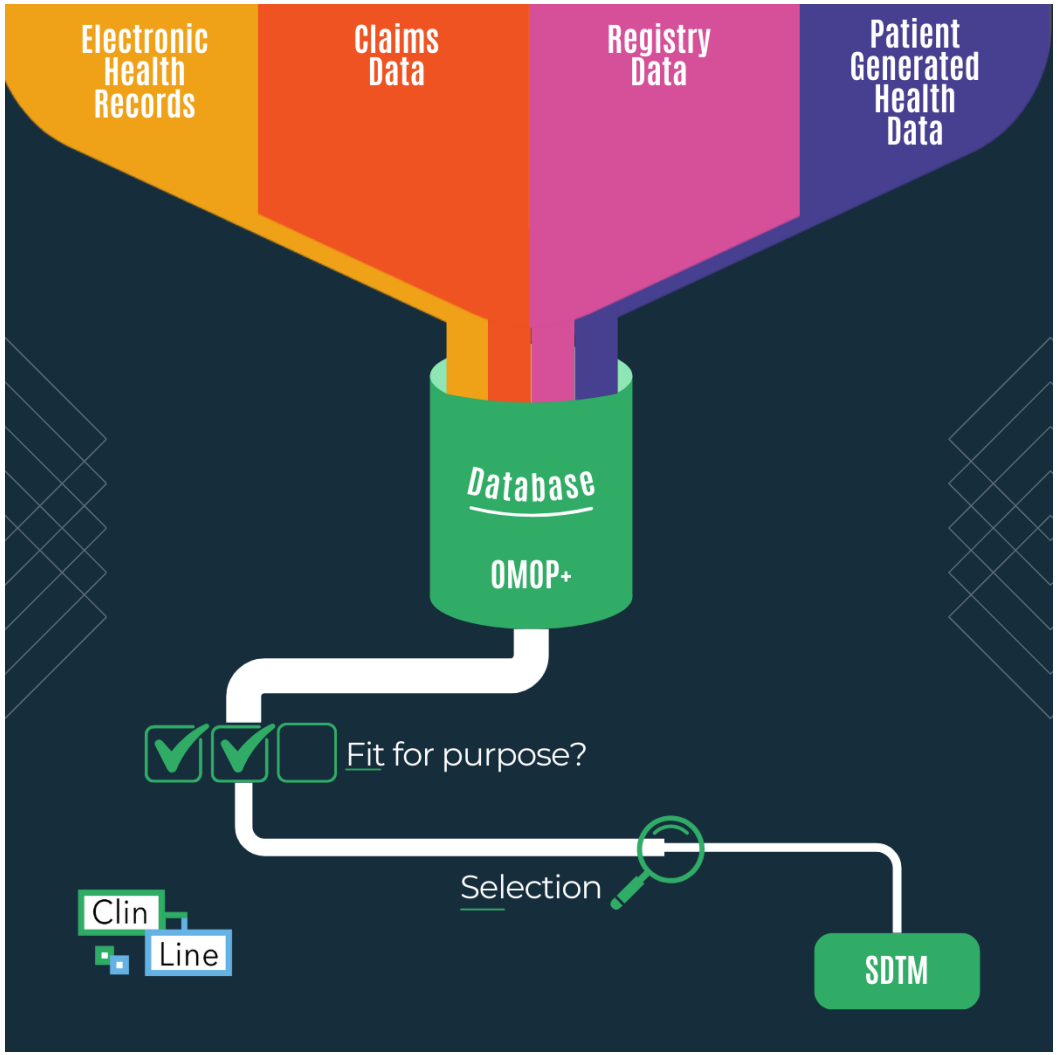
What data standard to use for real-world data?

- OMOP
  - Observational research standard
  - Basic structure - comparable to structures used at data vendor sites
  - Contains source information variables
  - Univocal: No repeated instances of same information
  - Includes mapping capabilities
    - Like SNOMED to MedDRA
  - Includes child/ancestor relationships

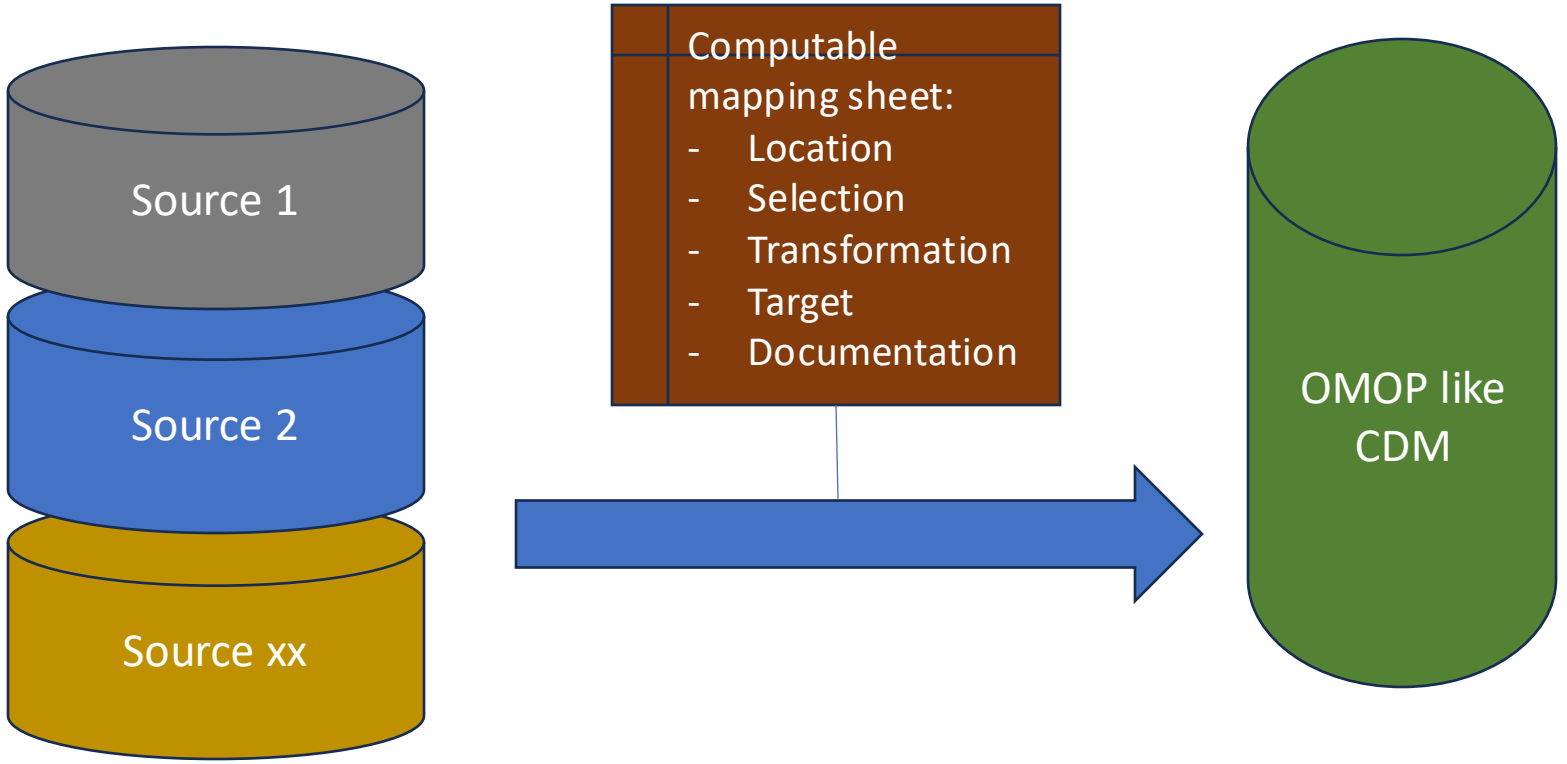


Our approach

1. Standardize from source to OMOP-like datasets
   - Omitted datasets if not needed
   - Added variables if beneficial for traceability and later processes
2. Select cohort
3. Assess fit for purpose
4. Match and impute
5. Transform to SDTM
6. Study specific analysis and reporting



Data mapping automation and traceability



Data transformations

- Mapping of dictionary codes
  - Open OMOP relationship table -> Traceable
  - Automated recognition of unmapped codes
  - Text matching / recognition -> suggest
  - Review unmapped / suggested items
  - Building custom – manual mapped and verified dictionary

- Calculations
  - Needed if level of detail varies between sources
    - Example: full score available in 1 source while only sub scores available in another source
  - Needed if expected (frequently used) endpoint is not at expected detail level
  - Specified in mapping sheet

Data transformations

- Custom codelist mapping
  - Automated transformation
  - Traceable
  - Address misalignment between sources
  - Include full coding trail from source to submission
  - Address misalignment in the trail from source to submission

Dataset	Variable	Label	Data Type	Value	Data source 1	Data source 2	Remarks
DEMO	RACE	Race	Char	Asian	Y	Y	
DEMO	RACE	Race	Char	African American		Y	
DEMO	RACE	Race	Char	Black or African American	Y		
DEMO	RACE	Race	Char	Other	Y	Y	
DEMO	BIRTH_YY	Birth year	Char			Y	Values are too diverse to present.
DEMO	BIRTH_YY	Birth year	Num		Y		Values are too diverse to present.

	Source 1		Source 2		OMOP				CDISC		
Codelist	Value	Code	Value	code	concept_id	Code	Name	Vocabulary	Codelist	Codelist code	CT
Route	p.o.	0	PO	1	4132161	26643006	Oral route	SNOMED	C66729	C38288	ORAL
Route	p.r.	1	PR	2	4290759	37161004	Rectal route	SNOMED	C66729	C38295	RECTAL
Route	s.c.	2	SC	3	4142048	34206005	Subcutaneous route	SNOMED	C66729	C38299	SUBCUTANEOUS
Route	i.m.	3	IM	4	4302612	78421000	Intramuscular route	SNOMED	C66729	C28161	INTRAMUSCULAR
Route	i.v.	4	IV	5	4171047	47625008	Intravenous route	SNOMED	C66729	C38276	INTRAVENOUS
Route	nasal	5	NASAL	6	4262914	46713006	nasal route	SNOMED	C66729	C38284	NASAL
Route	td	6	TD	7	4262099	45890007	transdermal route	SNOMED	C66729	C38305	TRANSDERMAL
Route			SL	8	4292110	37839007	sublingual route	SNOMED	C66729	C38300	SUBLINGUAL
Route			INH	9	45956874	9011000001100	Inhalation	SNOMED	C66729	C38216	RESPIRATORY (INHALATION)
Route			OTHER	10	9177	74964007	Other	SNOMED			
Route			NOT AVAILABLE	NA	45884396	LA7338-2	Not Available	LOINC			

Defensive programming

- Check in e.g. OMOP/SDTM creation program for any new source variables or removal of source variables
  - Output warning in log to indicate that there is a change in source variables
  - Make automatic updates to your mapping specification file

Source_dataset	Source_variable	Target_dataset	Target_variable	Specification
SOURCE.DEMO	RACE	DM	RACE	Recode according to codelist RACE
SOURCE.DEMO	BIRTH_YY	DM	BRTHDTC	Generate ISO-8601 date variable from BIRTH_YY, BIRTH_MM and BIRTH_DD
SOURCE.DEMO	BIRTH_MM	DM	BRTHDTC	See derivation at BIRTH_YY
SOURCE.DEMO	BIRTH_DD	DM	BRTHDTC	See derivation at BIRTH_YY
SOURCE.DEMO	RACEOTH	SUPPDM	RACEOTH	Copy from source variable
SOURCE.DEMO	RACEOT			

		only include	additional	Dataset Observation (OMOP variables)				
Source	dataset	WHERE	actions	observation_source_value	observation_type_concep	value_source_value	value_as_num	value_as_text
SOURCE 1	BASELINE		x	"Occupation"	32879	occup33	occup33	*Value(Occupation,* occup33*)*
SOURCE 2	VISITS		x	"Occupation"	32879	fcoccupation	fcoccupation	*Value(Occupation,* fcoccupation*)*
SOURCE 2	VISITS		x	"Occupation"	32879	occupatn	occupatn	*Value(Occupation,* occupatn*)*

Defensive programming

- Check in e.g. OMOP/SDTM creation program for any new source values
  - Use codelists
  - Output warning in log when source value is not found in codelist

Codelist	Type	From	To
RACE	C2C	Asian	ASIAN
RACE	C2C	African American	BLACK OR AFRICAN AMERICAN
RACE	C2C	Black or African American	BLACK OR AFRICAN AMERICAN
RACE	C2C	Other	OTHER
RACE	C2C		

Defensive programming

- Make your program robust
- Expect the worst of your input data (e.g. incomplete data, inconsistent data)
- Build in checks for unexpected data (e.g. data type, out-of-range values, unpredictable values)

Experiences: Standardization to OMOP-like CDM

- Distinction between OMOP measurement or observation unclear
  - Source concept review needed
- Identification of sensitive / non-proportional information
- Some information not accounted for in OMOP standards
  - Additional grouping variables from FHIR source
  - SDTM like visit numbers from registries
  - Status (active/completed)
- Source values already more aligned to SDTM than to OMOP
  - Registries may be partly aligned to SDTM
- Misalignment between sources in information captured and structures
- The need to map information of associated documents
  - Decode values
  - Assessment details

Experiences: Standardization to SDTM

- SV based on index date and windowing – multiple visits within 1 window
- RFSTDTC / RFENDTC definitions needed
- Adverse Events versus Medical History
  - Index date related
- Concomitant Medication versus Exposure
  - Study Objective related
- Source to CDISC dictionaries
  - LOINC is OMOP standard. so LOINC to CDISC transformation can be automated
  - SNOMED to MedDRA goes very well (>90% automated)
  - WHODrug mapping not available in OMOP
  - Unit transformations via standardized UCUM – CDISC unit codelist
  - Character results (NORMAL/ABNORMAL etc) via dictionary mapping sheet

Conclusions

- Account for high diversity in source data
- Ensure traceability and information readiness
  - Clear mapping overview
  - Versioning and Documentation
  - Inclusion of traceability information in resulting OMOP+ datasets
- Utilize defensive programming
- Include SDTM logic if available
  - AE / MH
  - CM / EX
  - SDTM CT

Questions and contact

- Questions?

- Like to learn more?
  - B.snoeijer@clinline.eu
  - Jules.vanderzalm@ocs-consulting.com