Bridging the Standards: Standardized Mapping and Lineage

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

15 May 2025


2025
CDISC + TMF
EUROPE INTERCHANGE
GENEVA
CONFERENCE & EXPO: 14-15 MAY | TRAININGS: 12, 13, 16 MAY

Meet the Speakers

Berber Snoeijer

Title: Manager Innovative Process and Solution Design

Organization: ClinLine

Berber has more than 25 years of experience in clinical research especially clinical data management, data governance, data analysis and reporting. Apart from that, she was the managing director of a CRO for 8 years and R&D manager working with Real-World data for another 8 years. In these roles she designed process-aligned tools and solutions to optimize the efficiency of data flows. In 2018, Berber founded ClinLine, which focuses on optimizing the clinical trial data process. Drawing on stakeholder input and requirements, she provides input and designs for data structures, solutions, and process optimization.

Jules van der Zalm

Title: Manager Operations

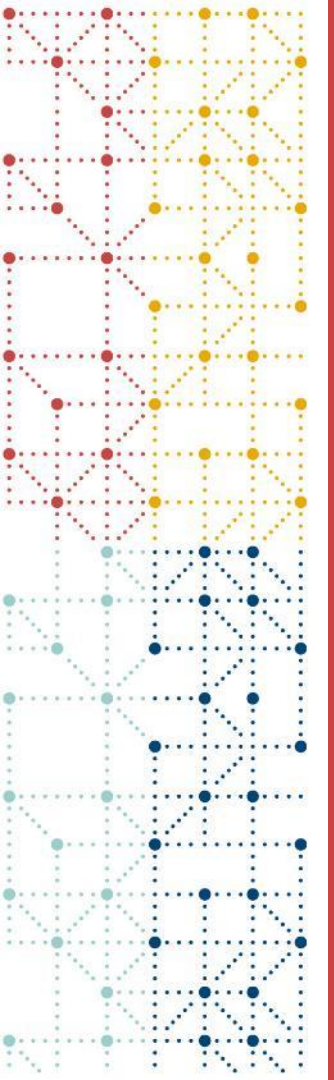
Organization: OCS Life Sciences

Jules van der Zalm is Manager Operations at OCS Consulting in the Netherlands. Combining a strong technical background with a passion for connecting people, Jules is responsible for the successful delivery of all project and training activities that OCS Consulting undertakes for its clients in the life sciences industry. Jules has a strong community presence, with involvement in both global and local organizations, including PHUSE, an independent organization run by a worldwide team of volunteers that aims to bring together industry and regulatory bodies, and PSDM, a Dutch group focused on fostering collaboration among pharmaceutical data and statistics experts.

Disclaimer and Disclosures

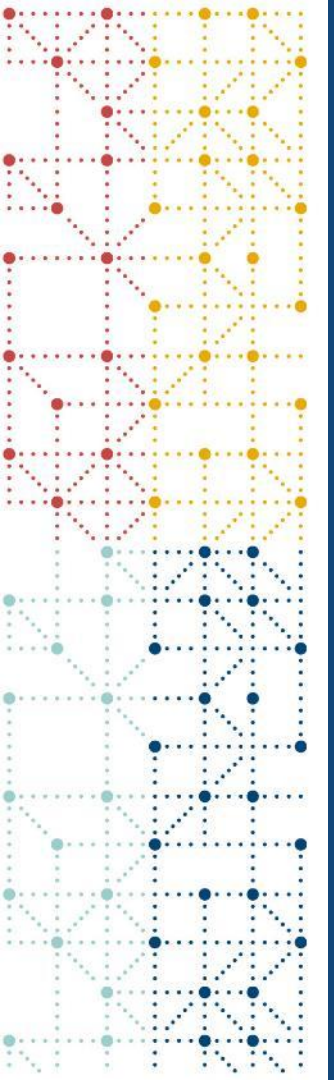
- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.*

- *The author(s) have no real or apparent conflicts of interest to report.*



Agenda

1. Introduction
2. Data Lineage
3. Mapping
4. Automation
5. Conclusion



Introduction

Introduction – Real-World Data

- Data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources. (FDA definition)
- Guidelines:
  - Relevance
  - Reliability
  - Timeliness
  - Coherence
  - Extensiveness

Introduction – Data source types

- Electronic Health Records (EHR) and Claims
  - Data primary collected in the scope of treatment of patients in regular clinical practice
- Registries
  - An organized system that collects clinical and other data in a standardized format for a population defined by a particular disease, condition, or drug exposure.
  - Data can be entered directly into the registry and may also include data from other sources
- Patient Generated Health Data (PGHD)
  - Social Media
  - Commercial Devices

Introduction – The Real-World Data Jungle

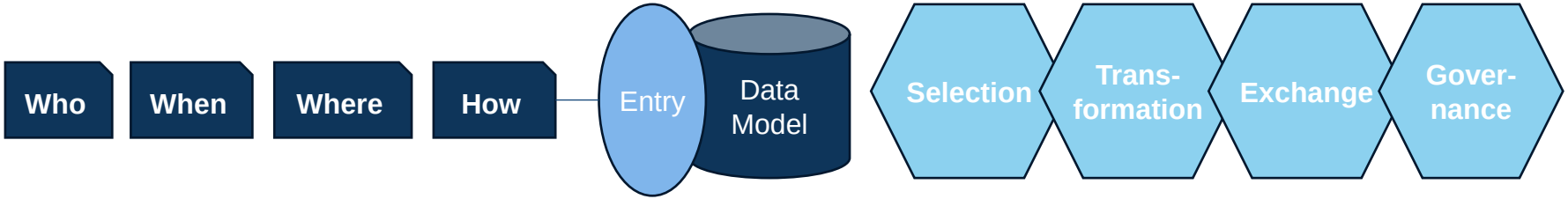
- Variety of sources
  - General Practitioner
  - Hospital Specialists
  - Pharmacy
  - Laboratory
  - Other
- Each source has his/her own scope:
  - Collection practice has effect on **reliability** and analysis
  - What is collected has effect on **relevance** of data
- Different EHR information systems result in variation within and between domains
  - Different data structures
  - Different entry screens

Data Lineage

Data Lineage

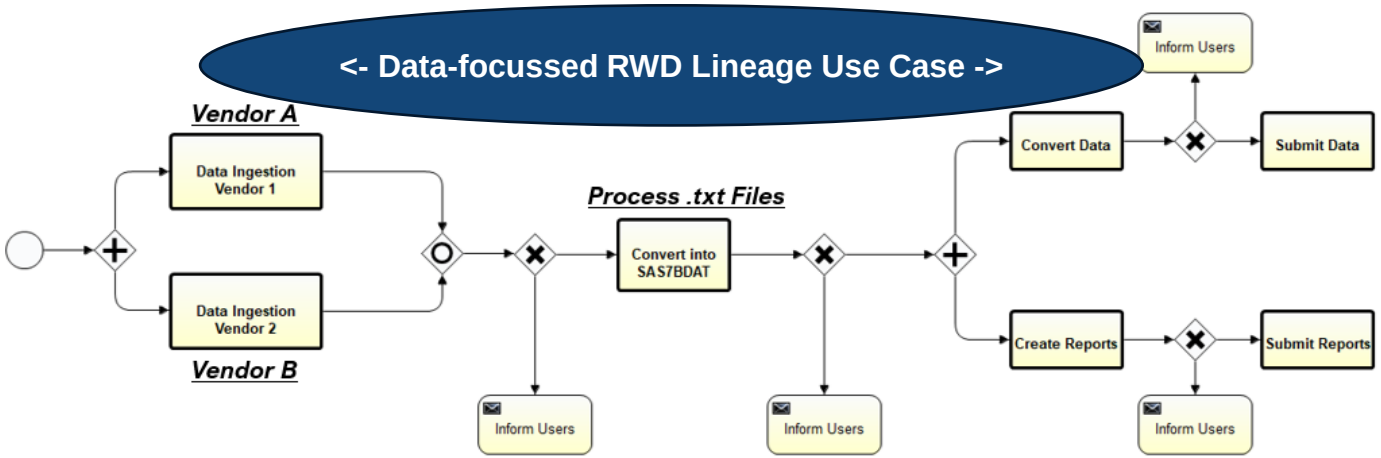
See www.clinline.org/break-learn for the intro to data lineage webinar ->

Wikipedia: **Data lineage** refers to the process of tracking how data is generated, transformed, transmitted and used across a system over time. It documents data's origins, transformations and movements, providing detailed visibility into its life cycle. This process simplifies the identification of errors in data analytics workflows, by enabling users to trace issues back to their root causes.



Process - lineage

Populating a Process Flow Instance - Example 1: Default configuration



Presented by Kai Wanke (OCS LS) at the PHUSE Single Day in Utrecht on 17 April

Use Case (CDISC RWD Lineage)

1. Select a cohort of Parkinson patients from MIMIC deidentified Real-World Data source in FHIR format.
2. Transform FHIR data to a standardized observational data model (OMOP) for selection and imputation purposes.
3. Assess eligibility based on parameterized eligibility criteria.
4. Transform eligible and selected patient data to SDTM submissible datasets
5. Select and Map the data with the OCS LS Mapping Engine
6. Show lineage for every step and transformation using a newly developed Open Source tool.

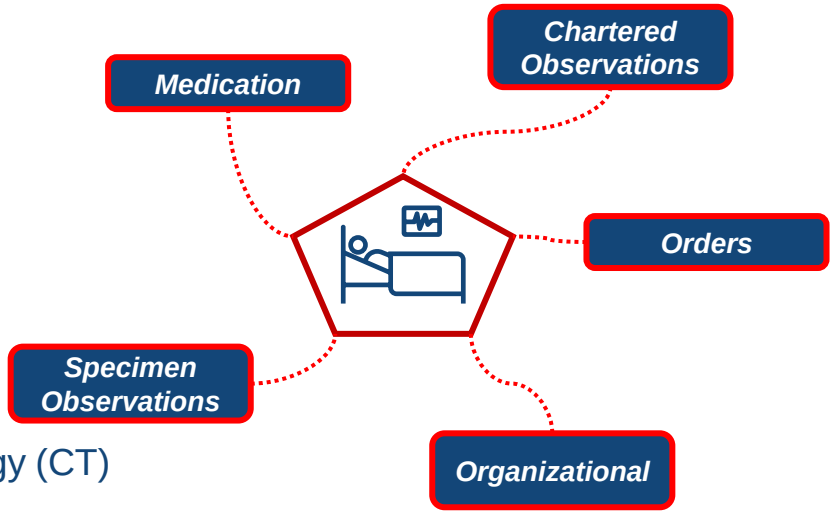
Step 1:
Focus on patient data and corresponding lineage to SDTM DM dataset

Example: The Complexity of MIMIC

- Medical Information Mart for Intensive Care (MIMIC-IV)
- A Real World Dataset:
  - 300,000 deidentified critical care patients
  - 430,000 hospital admissions
  - 70,000 ICU stays
- Captures the Full Patient Journey
  - Integrates data from three hospital systems: ED, ICU, and EHR
- Challenges: A Complex & Heterogenous Dataset
  - Data from different sources with multiple terminologies
  - Heterogeneous, unstructured, and inconsistent formats
  - Designed for real-world care, not for regulatory submission

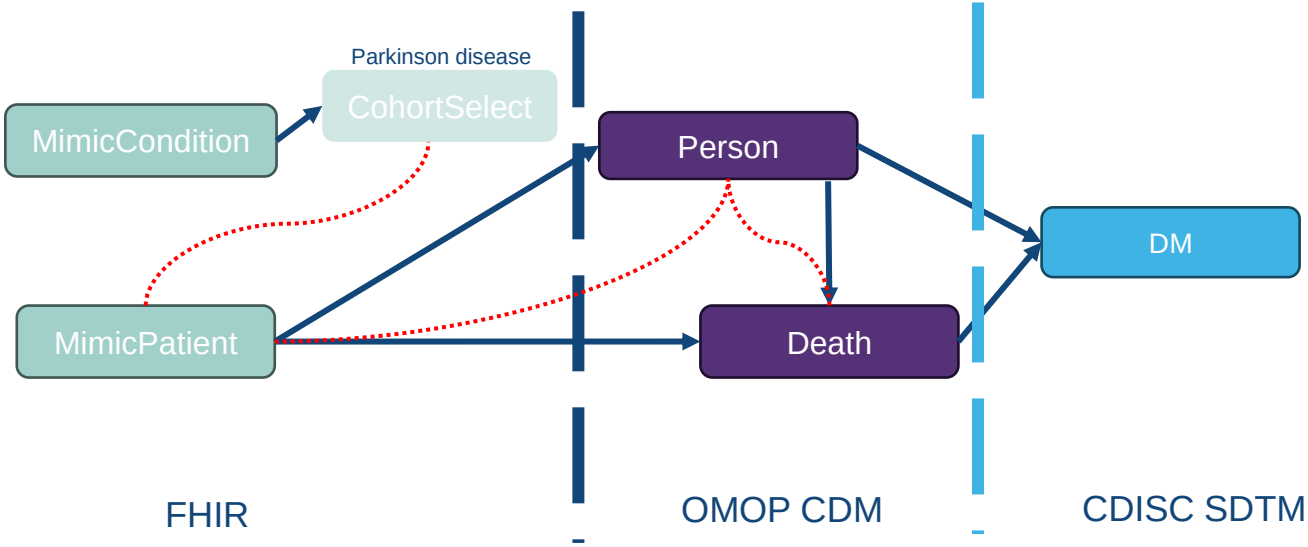
Transforming MIMIC for Submission

- MIMIC-IV-on-FHIR
  - Organizes MIMIC data using FHIR standards
  - Uses 30+ structured dataset (profiles)
  - Retains original terminology
  - No data modifications
- Converting to CDISC
  - Aligns terminology with Controlled Terminology (CT)
  - Maps data to submission-ready formats
  - Ensures CDISC Compliance for regulatory use



Proof of concept: Mapping Patient Data

- FHIR Patient resource to CDISC Standard (SDTM Demographics)



Alignment to CDISC 360i

- A multi-year initiative with the aim to transform the way we develop and use standards within clinical research connected and interoperable information
  - enabling automation,
  - enhancing data integrity,
  - accelerating innovation.
- Use Cases
  - LZZT
  - Breast Cancer
  - Parkinson (aligned to LZZT)

- CDISC Digital Data Flow
- RWD Lineage Project
- 360i Project

resources and recordings: https://www.cdisc.org/cdisc-360i

Mapping

Considerations

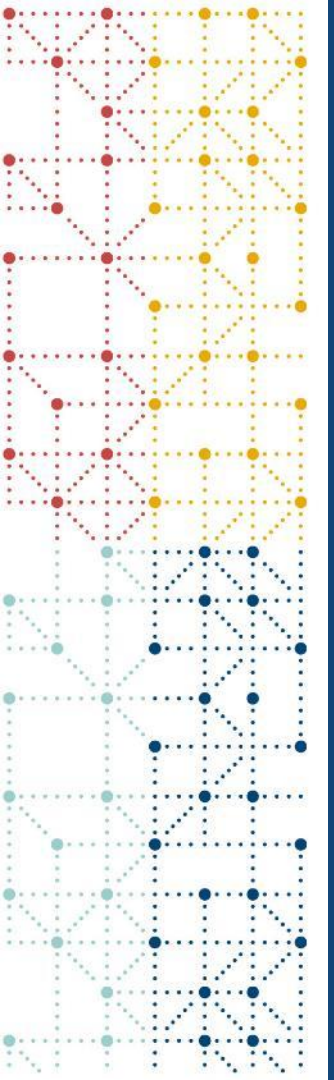
- Data Structure challenges when working with FHIR and SDTM.
  - Different purpose and structure
- Terminology Alignment:
  - FHIR stores terminology in two resources: CodeSystems and ValueSets
  - SDTM uses CDISC-controlled terminology (CT)
- Handling challenges like missing identifiers and standardizing formats
- Data mapping: Determine which information might be relevant for target data
- Ensuring compliance with CDISC standards

Mapping example - Variables

Input Row Id	Input dataset	Input variable Name	Output dataset	Output variable Name	Condition Text	Condition code	Transformation type (select)	Transformation Text	Transformation code	Recode list reference
1	MimicCondition.ndjson	condition.code.coding.code	CohortSelect	ConditionCode	Parkinson conditions (ICD-10)	condition.code.coding.code="G20" and condition.code.coding.system="http://mimic.mit.edu/fhir/mimic/CodeSystem/mimic-diagnosis-icd10"	NONE			
2	MimicCondition.ndjson	condition.code.coding.display	CohortSelect	ConditionDisplay	Parkinson conditions (ICD-10)	condition.code.coding.code="G20" and condition.code.coding.system="http://mimic.mit.edu/fhir/mimic/CodeSystem/mimic-diagnosis-icd10"	NONE			
3	MimicCondition.ndjson	condition.id	CohortSelect	ConditionUID			NONE			
4	MimicCondition.ndjson	condition.subject.reference	CohortSelect	SubjectUID			NONE			
5	MimicCondition.ndjson	condition.encounter.reference	CohortSelect	EncounterUID			NONE			
6	MimicPatient.ndjson	patient.identifier.value	person	person_id			UNIQUE_NUMBER			
7	MimicPatient.ndjson	patient.gender	person	gender_concept_id			RECODE			_CM1,_CM2
8	MimicPatient.ndjson	patient.birthDate	person	year_of_birth			TRANSFORM		YEAR(datetime(birth_datetime))	
9	MimicPatient.ndjson	patient.birthDate	person	month_of_birth			TRANSFORM		MONTH(datetime(birth_datetime))	
10	MimicPatient.ndjson	patient.birthDate	person	day_of_birth			TRANSFORM		DAY(datetime(birth_datetime))	
11	MimicPatient.ndjson	patient.birthDate	person	birth_datetime			TRANSFORM		DATETIME(birthDate)	
12	MimicPatient.ndjson	patient.race.display	person	race_concept_id			RECODE			
13	MimicPatient.ndjson	patient.ethnicity.display	person	ethnicity_concept_id			RECODE			
14	CohortSelect	SubjectUID	person	person_source_value			NONE			
15	MimicPatient.ndjson	patient.gender	person	gender_source_value			NONE			
16	MimicPatient.ndjson	patient.race.display	person	race_source_value			NONE			
17	MimicPatient.ndjson	patient.ethnicity.display	person	ethnicity_source_value			NONE			

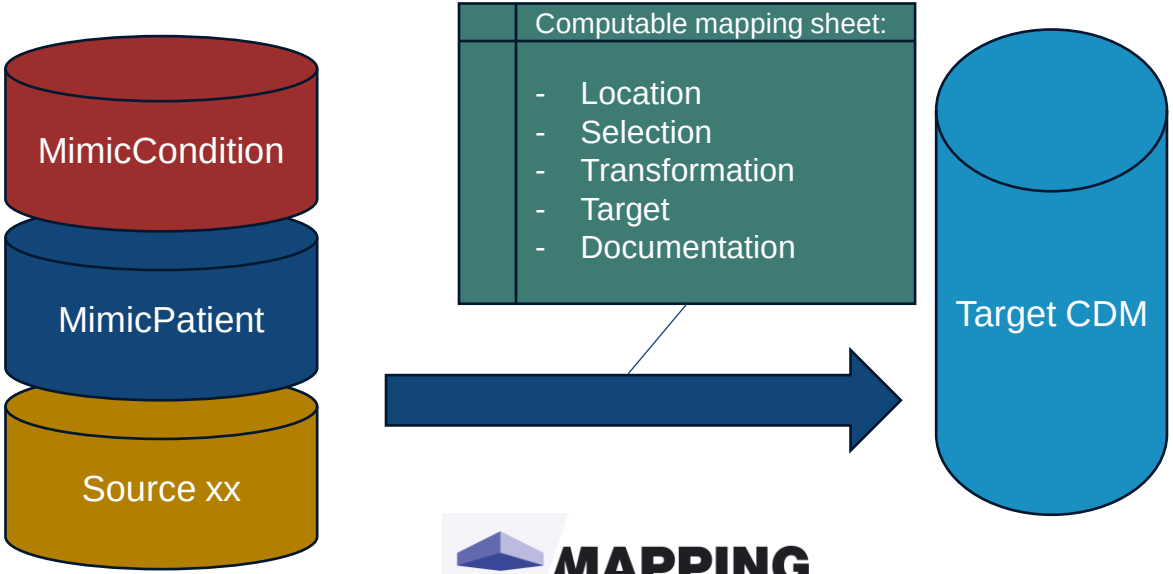
Mapping example - Codes

Recode list reference	From	CodeList	CodeSystem	Decode	To	CodeList	CodeSystem	Decode	Comment
SEX_FHIR_OMOP	F	BIRTH_SEX	US CORE	Female	8532	Gender	OMOP	F	
SEX_FHIR_OMOP	M	BIRTH_SEX	US CORE	Male	8507	Gender	OMOP	M	
SEX_OMOP_SDTM	8532	Gender	OMOP	F	C16576	Sex	CDISC	F	
SEX_OMOP_SDTM	8507	Gender	OMOP	M	C20197	Sex	CDISC	M	
RACE_FHIR_OMOP	2106-3	RACE	US CORE	White	8527	Race	OMOP	White	
RACE_FHIR_OMOP	1002-5	RACE	US CORE	American Indian or Alaska Native	8657	Race	OMOP	American Indian or Alaska Native	
RACE_FHIR_OMOP	2028-9	RACE	US CORE	Asian	8515	Race	OMOP	Asian	
RACE_FHIR_OMOP	2054-5	RACE	US CORE	Black or African American	8516	Race	OMOP	Black or African American	
RACE_FHIR_OMOP	2076-8	RACE	US CORE	Native Hawaiian or Other Pacitic Islander	8557	Race	OMOP	Native Hawaiian or Other Pacific Islander	
RACE_FHIR_OMOP	2131-1	RACE	US CORE	Other Race	9177	Other	OMOP	Other	
RACE_FHIR_OMOP	ASKU	NullFlavor	FHIR	asked but unknown	4129922	unknown	OMOP	Unknown	
RACE_FHIR_OMOP	UNK	NullFlavor	FHIR	unknown	4129922	unknown	OMOP	Unknown	



Automation

Data mapping automation and traceability



MAPPING ENGINE
CDISC DATA CONVERSION DONE FASTER

Automation – Mapping Engine

SDTM mapping

	SOURCE_DS	SOURCE_VAR	TARGET_DS	TARGET_VAR	SPECIFICATION	FUNCTION
1	SOURCE.DEM		DM	DOMAIN	Assign "DM"	**ASSIGN**[DM]
2	SOURCE.DEM		DM	STUDYID	Assign "OCS"	**ASSIGN**[OCS]
3	SOURCE.DEM	INVSITE	DM	SITEID	Copy from source variable	**COPY**
4	SOURCE.DEM	IDSUBNUM	DM	SUBJID	Copy from source variable	**COPY**
5	SOURCE.DEM	IDSUBNUM	DM	USUBJID	Concatenate STUDYID and IDSUBNUM, separated by "-"	**CONCAT**[STUDYID || "-" || PUT(IDSUBNUM, BEST.)]
6	SOURCE.DEM	VISIT			Not mapped	**NOT MAPPED**
7	SOURCE.DEM	IDLIN			Not mapped	**NOT MAPPED**
8	SOURCE.DEM	DT_BTH	DM	BRTHDTC	Copy from source variable, converted to ISO 8601 format	**ISODATE**[DATE=DT_BTH]
9	SOURCE.DEM	AGE	DM	AGE	Copy from source variable	**COPY**
10	SOURCE.DEM	AGEUNIT	DM	AGEU	Copy from source variable if AGE is not missing	**FUNCTION**[IF ^MISSING(AGE) THEN DO; #COPY#; END;]
11	SOURCE.DEM	GENDER	DM	SEX	Recode according to SEX recoding list	**RECODE**[SEX]
12	SOURCE.DEM	RACE	DM	RACE	Recode according to RACE recoding list	**RECODE**[RACE]
13	SOURCE.DEM		DM	COUNTRY	Assign "NLD"	**ASSIGN**[NLD]
14	SOURCE.SIC		DM	STUDYID	Assign "OCS"	**ASSIGN**[OCS]
15	SOURCE.SIC	IDSUBNUM	DM	USUBJID	Concatenate STUDYID and IDSUBNUM, separated by "-"	**CONCAT**[STUDYID || "-" || PUT(IDSUBNUM, BEST.)]
16	SOURCE.SIC	DT_DCM			Not mapped	**NOT MAPPED**
17	SOURCE.SIC	IDLIN			Not mapped	**NOT MAPPED**
18	SOURCE.SIC	DT_CON	DM	RFICDTC	Copy from source variable, converted to ISO 8601 format	**ISODATE**[DATE=DT_CON]
19	SOURCE.DISP		DM	STUDYID	Assign "OCS"	**ASSIGN**[OCS]
20	SOURCE.DISP	IDSUBNUM	DM	USUBJID	Concatenate STUDYID and IDSUBNUM, separated by "-"	**CONCAT**[STUDYID || "-" || PUT(IDSUBNUM, BEST.)]
21	SOURCE.DISP	DT_DCM	DM		Keep to apply poststeps	**KEEP**
22	SOURCE.DISP	TMSDIS	DM		Keep to apply poststeps	**KEEP**
23	SOURCE.END		DM	STUDYID	Assign "OCS"	**ASSIGN**[OCS]
24	SOURCE.END	IDSUBNUM	DM	USUBJID	Concatenate STUDYID and IDSUBNUM, separated by "-"	**CONCAT**[STUDYID || "-" || PUT(IDSUBNUM, BEST.)]
25	SOURCE.END	DT_DCM	DM	RFPENDTC	Copy from source variable, converted to ISO 8601 format	**ISODATE**[DATE=DT_DCM]
26	SOURCE.END	COMPL			Not mapped	**NOT MAPPED**

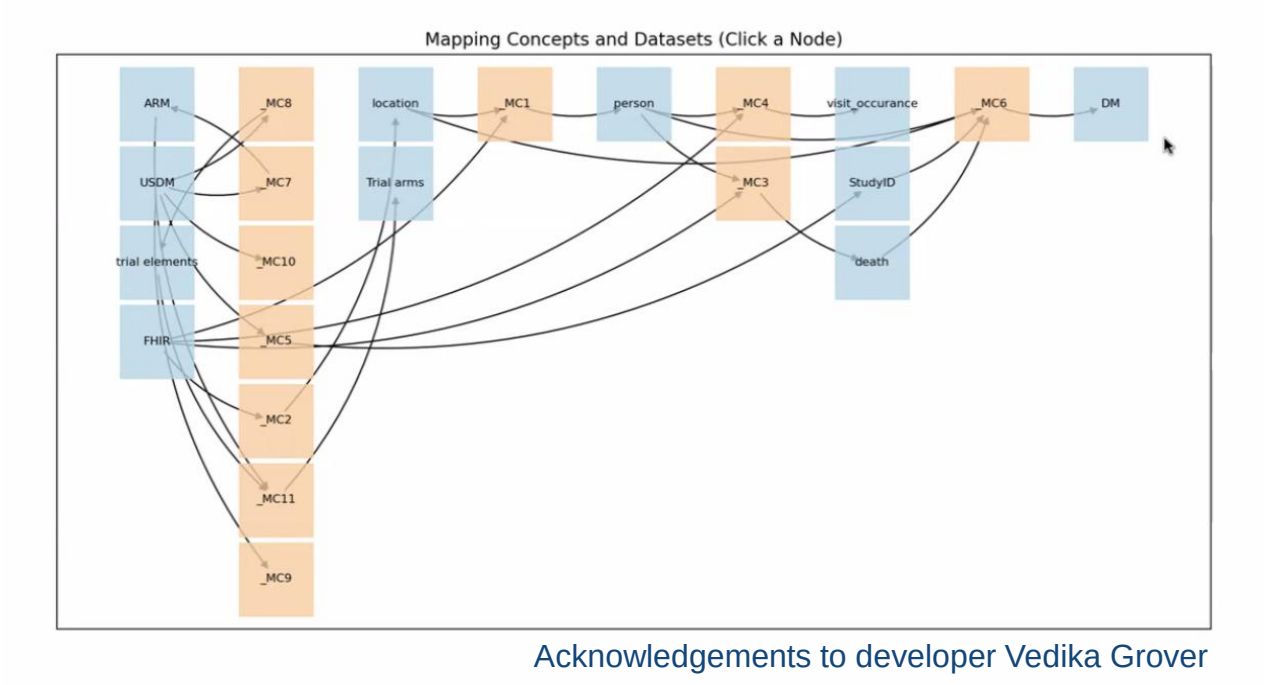
Mapping | Recode | Tables | Columns | Parameters | Libraries

Automation – Mapping Engine

- Can be developed in SAS, R, Python or any other language
- Has the *complete* data mapping in a single artefact
- Uses a library to store repeating mapping rules
- Warns you of missing or unexpected data values

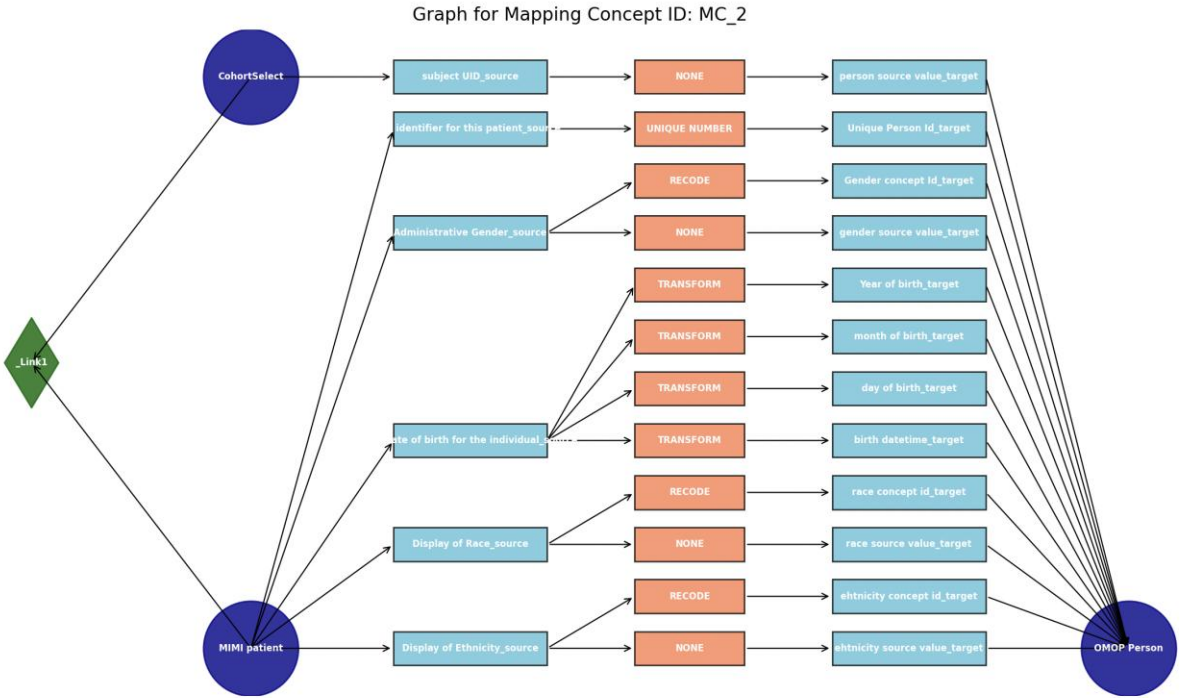
Automation – Lineage Tool

- Open source tool in Python
- Insights on different levels
  - Datasets
  - Transformations
  - Codes

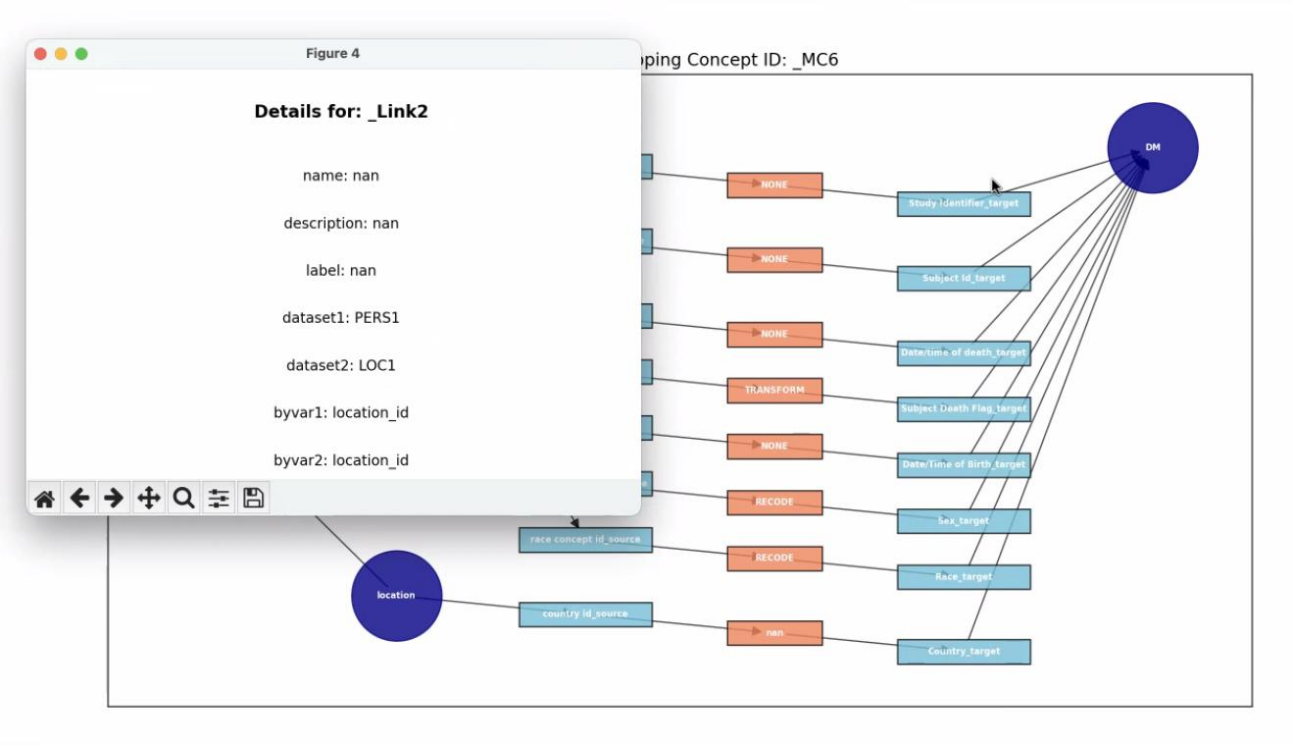


Acknowledgements to developer Vedika Grover

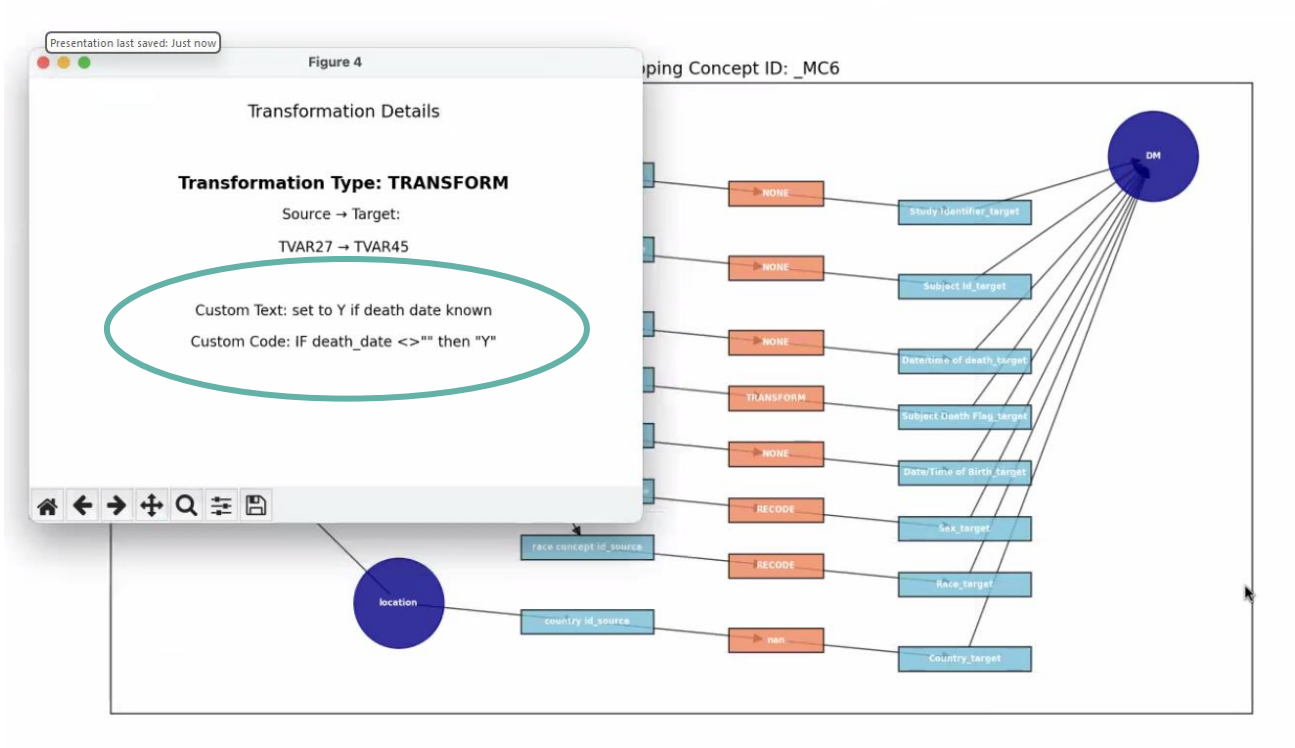
Automation – Lineage tool



Lineage tool output



Lineage tool output



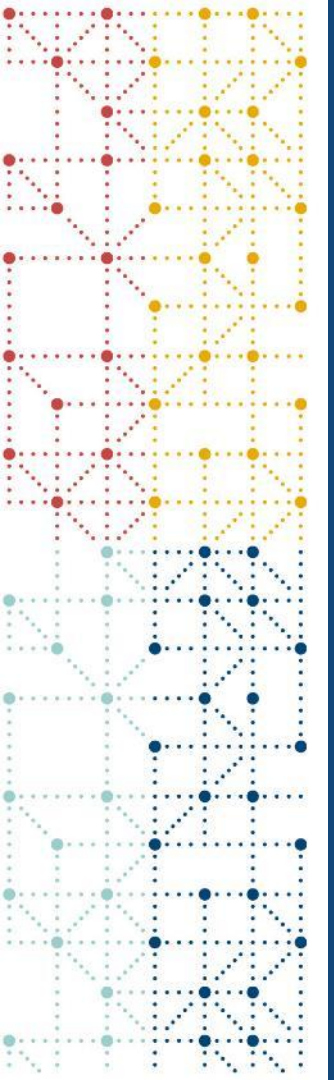
Next steps

- Lineage tool
  - Improve presentation
  - Add more mapping details
  - Make open source and available via Github
- Mapping
  - Add more OMOP domains to drive other SDTM domains
  - Add more SDTM domains
  - Add USDM as input to drive the process
  - Align with CDISC RWD lineage team on mapping logic and output
- Test
  - Test on other indications

Conclusions

Conclusions

- Account for high diversity in source data
- Ensure lineage and information readiness
  - Clear mapping methods and mapping overview
  - Inclusion of traceability information
  - Use Visualizations for process insights
- Automation
  - Standard and validated programs
  - Reduce manual tasks
  - Utilize defensive programming
- Future improvements
  - More Automation & Scalability
  - Open source tool
  - USDM as a driver
  - Available coding and BC lists as a driver



Section Header

Custom Slides Using Infographics

Thank You!

Questions?

- Jules van der Zalm: jules.vanderzalm@ocs-consulting.com
- Berber Snoeijer: b.snoeijer@clinline.eu