Real-World Data Lineage: requirements and experiences

Berber Snoeijer

21 November 2025

Meet the Speaker

Berber Snoeijer

Title: MSc.

Organization: ClinLine

Berber has a degree in Biomedical Sciences and more than 25 years of experience in clinical research. She was the managing director of a data oriented 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 study data process. Drawing on stakeholder input and requirements, she provides input and designs for data structures, solutions, and process optimization.

Agenda

1. Introduction
2. Clinical trials vs Real-World Data studies
3. Transformation nodes
4. Projects and Solutions

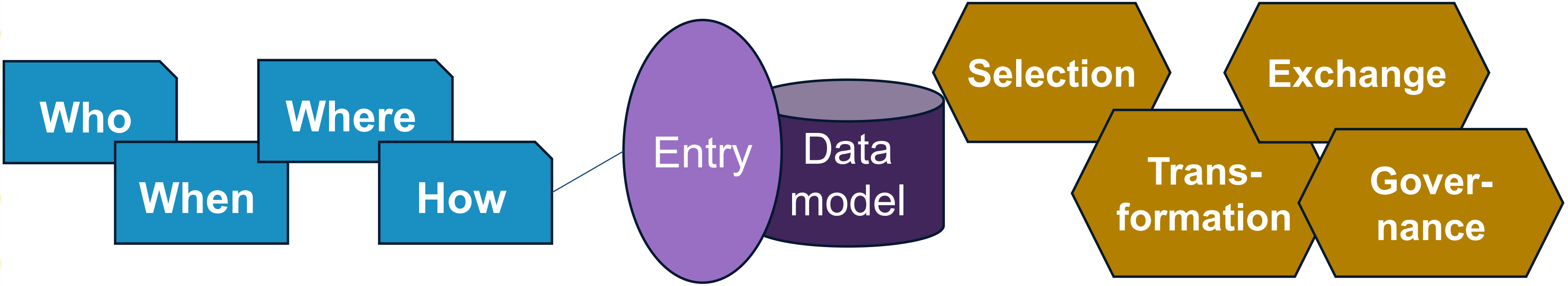
Introduction

What is Lineage?

What are the regulatory requirements?

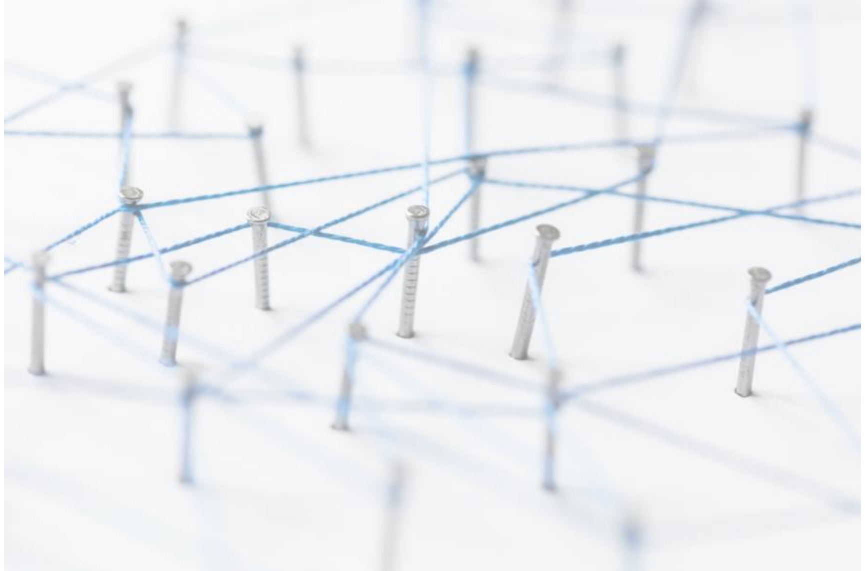
What is Data Lineage?

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. (Wikipedia)



What is data lineage

- Data origin
  - Data collection (Who, When, Where, How)
  - Data model, standards and storage
- Data selection steps AI?
  - Criteria
- Data transformation steps
  - Validation and corrections AI?
  - Standardisation
- Data exchange
- Governance
  - Responsibility, Processes, Procedures, Standards



Regulatory Requirements

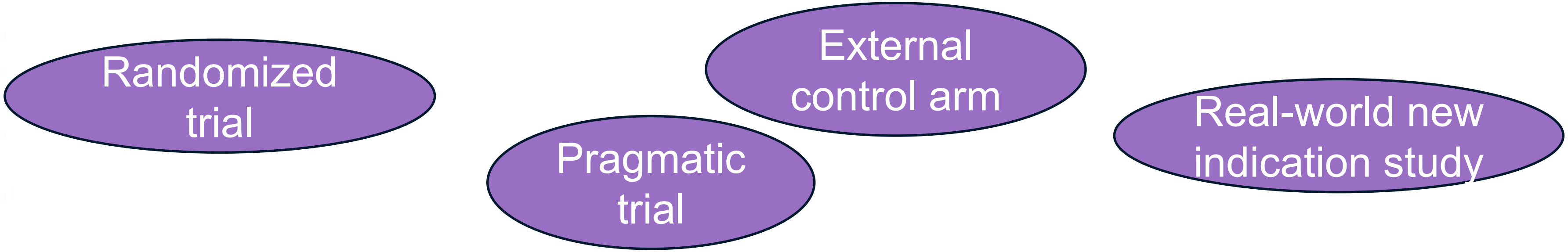
- EMA Data Quality Framework
  - Reliability: fundamentally depends on the systems and process in place for the primary collection of data and its processing.
- FDA guideline: Data Standards for Drug and Biological Product Submissions Containing Real-World Data
  - During data curation and data transformation, adequate processes should be in place to ensure confidence in the resultant data.
    - Documentation of these processes may include but is not limited to electronic documentation (e.g., audit trails, quality control procedures, etc.) of data additions, deletions, or alterations from the source data system to the final study analytic data set(s).
- FDA guideline: Real World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products
  - In general, sponsors should address the procedures used to ensure completeness and accuracy of study data, as well as processes for data accrual, curation, and transformation over the data life cycle.

Clinical Trials vs Real-World Data Studies

Clinical Trials vs Real-World Data Studies

- Clinical trials
  - Usually randomized
  - Data collected for the purpose of the trial .
  - Data collected in EDC
  - Data processes according to decades of standard siloed processes and procedures.

- Real-World Data Studies
  - Non-Randomized
  - Data coming from real-world data sources like EHR and registries.
  - Data collected in source information systems.
  - Silos do not work anymore. More interaction needed.



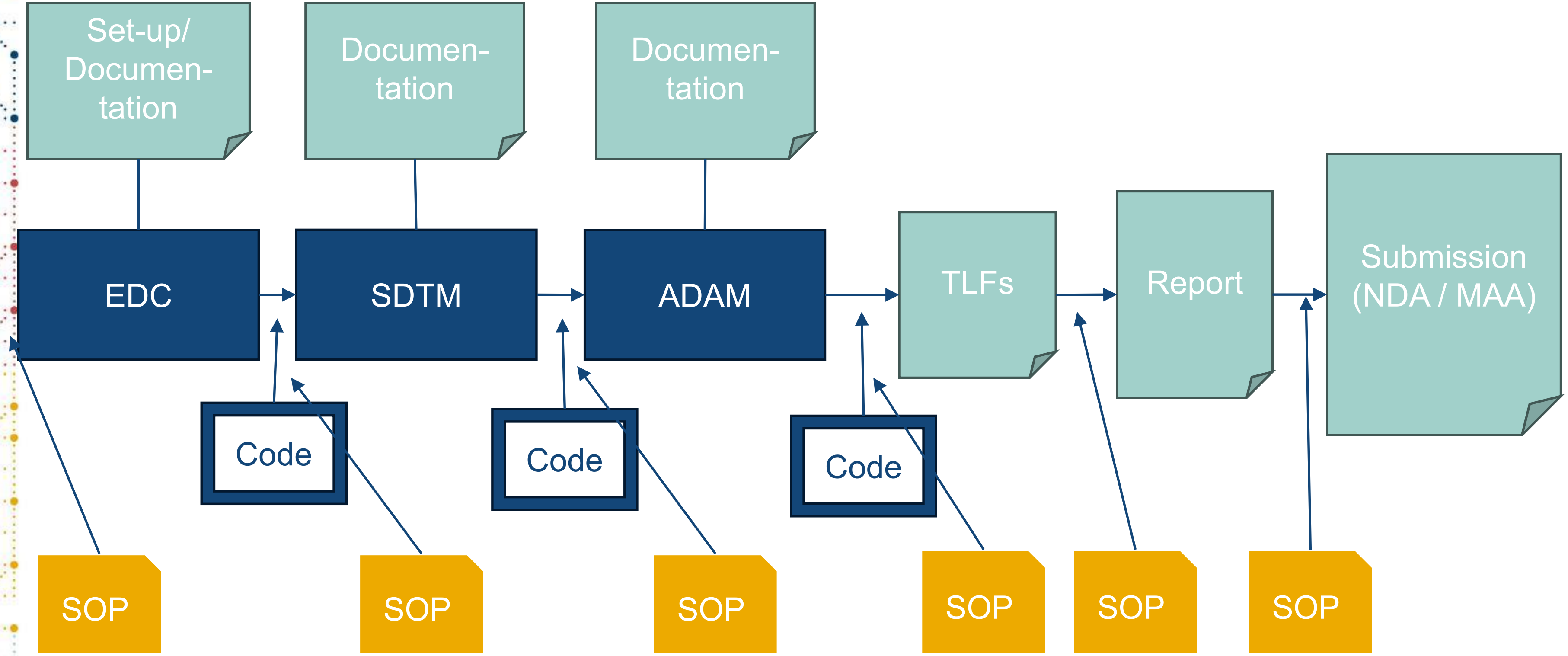
Randomized trial

External control arm

Pragmatic trial

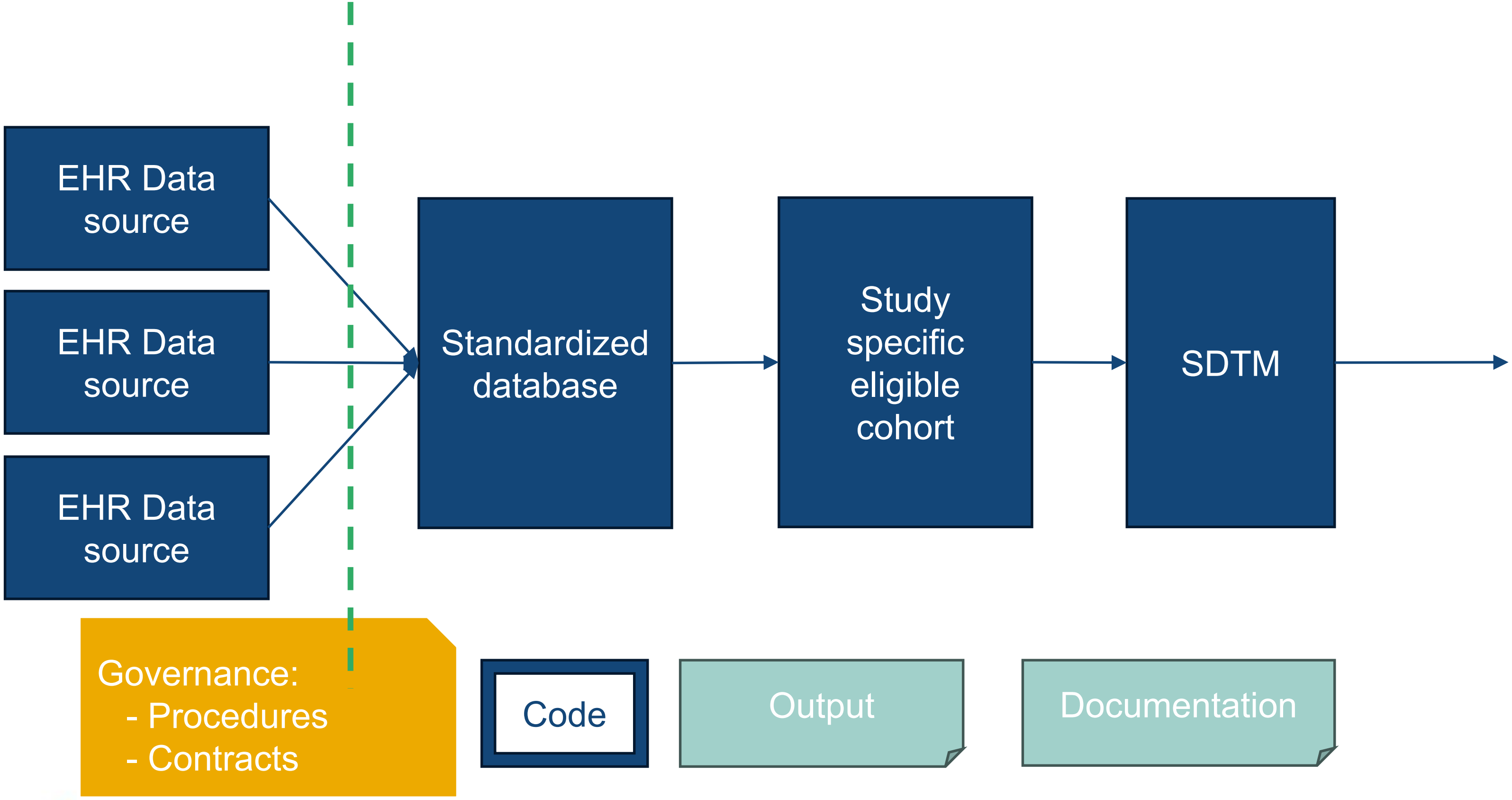
Real-world new indication study

Clinical Trial Lineage



#ClearDataClearImpact

Real-World Data Lineage



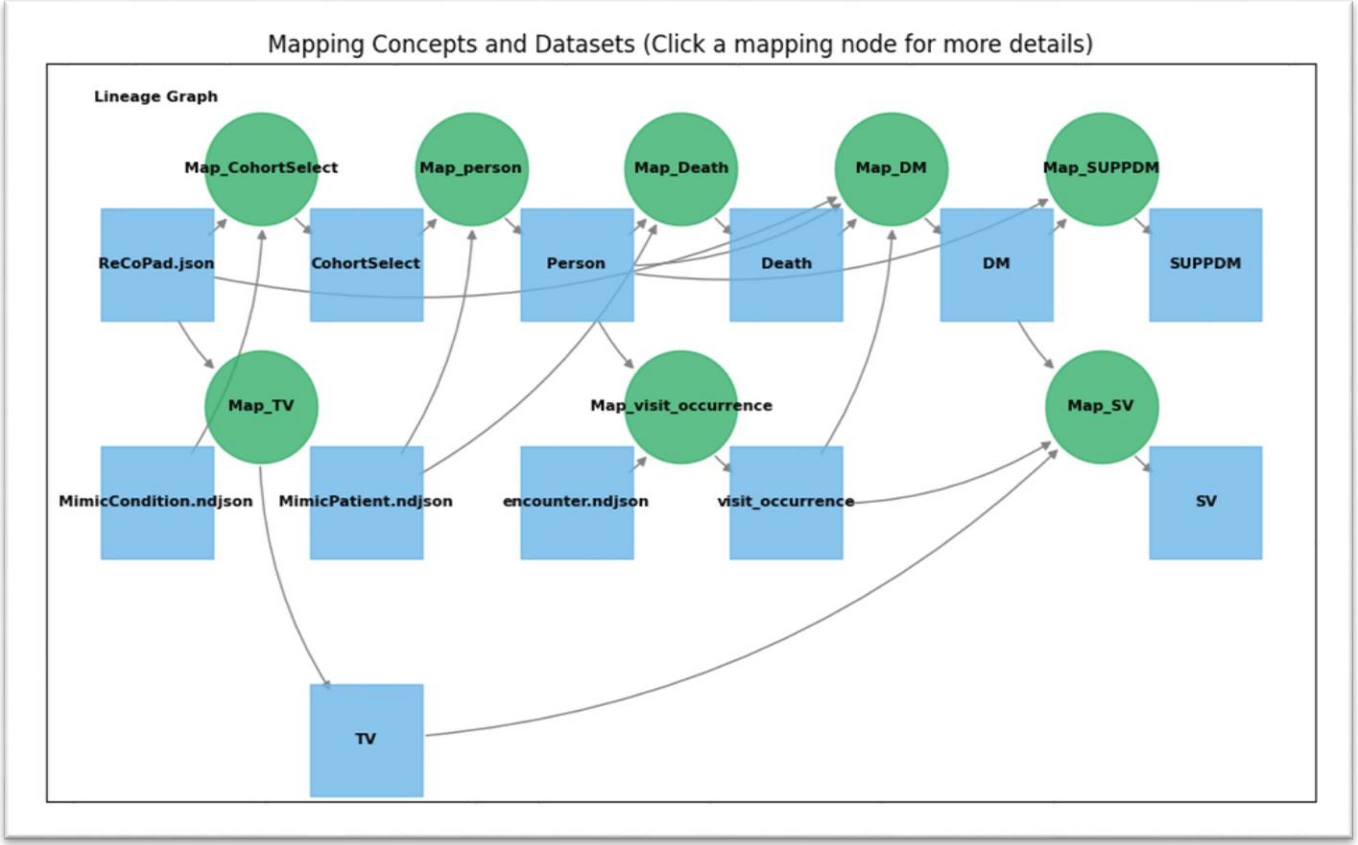
#ClearDataClearImpact

Transformation Nodes

Transformation Nodes

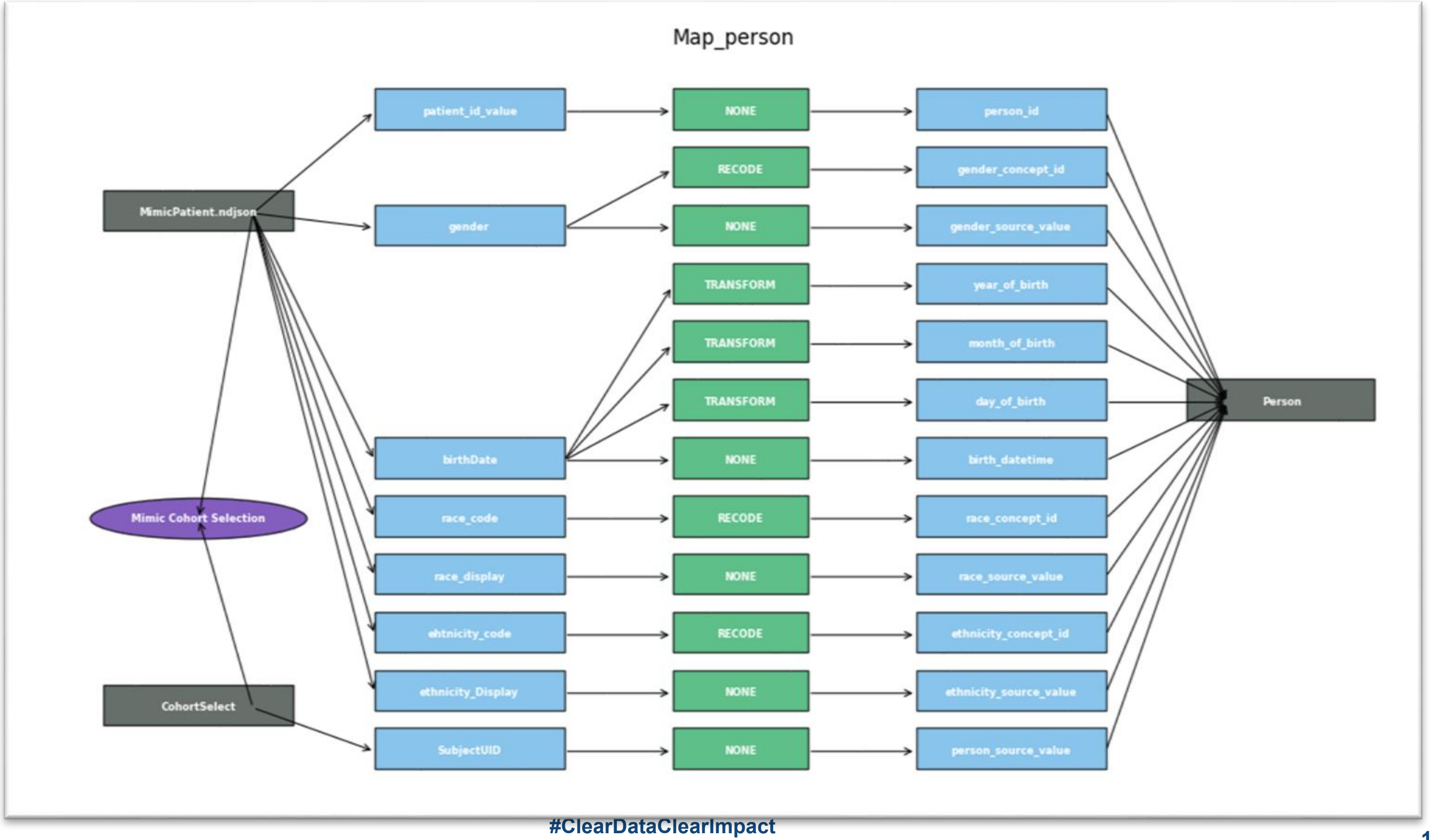
- Transformation between two fixed states in data lineage
  - Data exchange
  - Validation
  - Mapping and data transformations
  - Data Selection (positive / negative)
  - Reporting
- Combines all information defining the transformation
  - Code
  - Input datasets
  - Input specifications
  - Procedures
  - Contracts

Transformation Nodes



#ClearDataClearImpact

Transformation Nodes



Security and verifiability

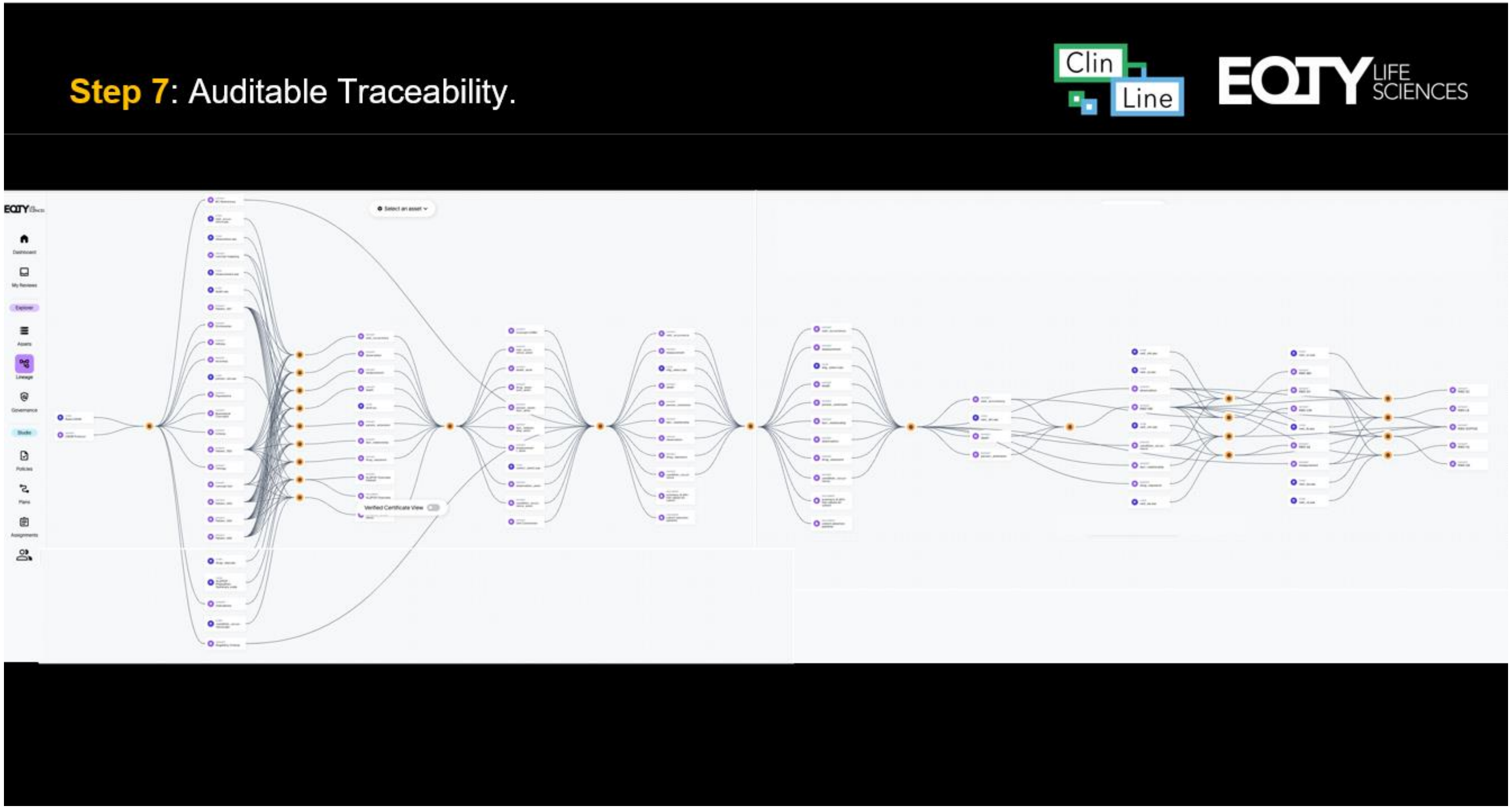
- Transformation nodes and all referred information:
  - Input and output
  - Documentation
  - Code and log / proof of execution
- Immutable
  - Protected environment
  - Proof that it is not changed
    - Date/time file/run / log verification
    - Rerun verification
    - Hashing / cryptography

Challenges

- Source data formats
- Mapping details
- Documentation!!
  - Availability
  - Findability
- Pseudonymization/Anonymization
- AI utilization
- Submission data requirements

Projects and solutions

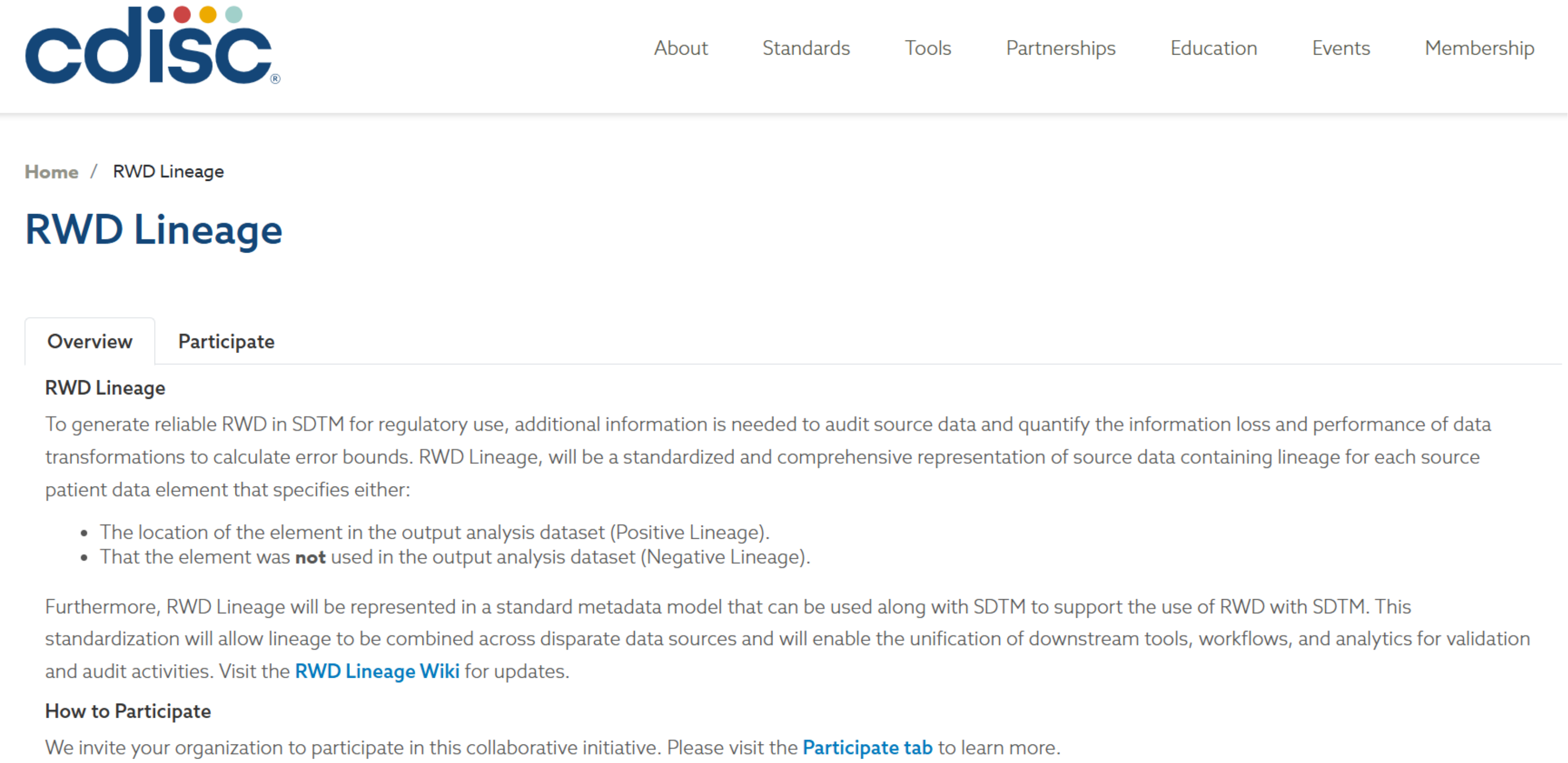
Hashing / dashboard solutions



Step 7: Auditable Traceability.

#ClearDataClearImpact

CDISC RWD lineage



RWD Lineage

RWD Lineage

To generate reliable RWD in SDTM for regulatory use, additional information is needed to audit source data and quantify the information loss and performance of data transformations to calculate error bounds. RWD Lineage, will be a standardized and comprehensive representation of source data containing lineage for each source patient data element that specifies either:

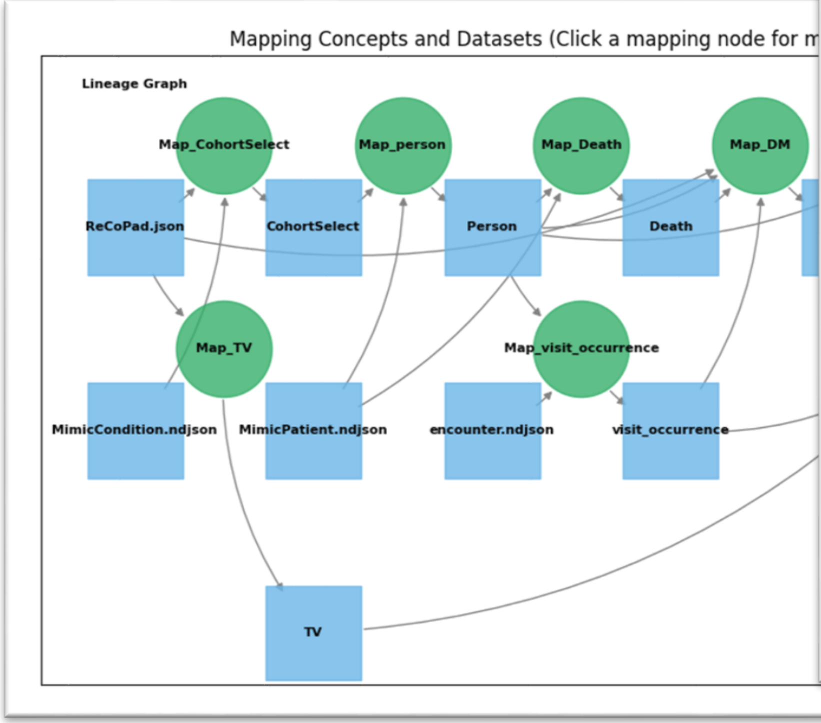
- The location of the element in the output analysis dataset (Positive Lineage).
- That the element was **not** used in the output analysis dataset (Negative Lineage).

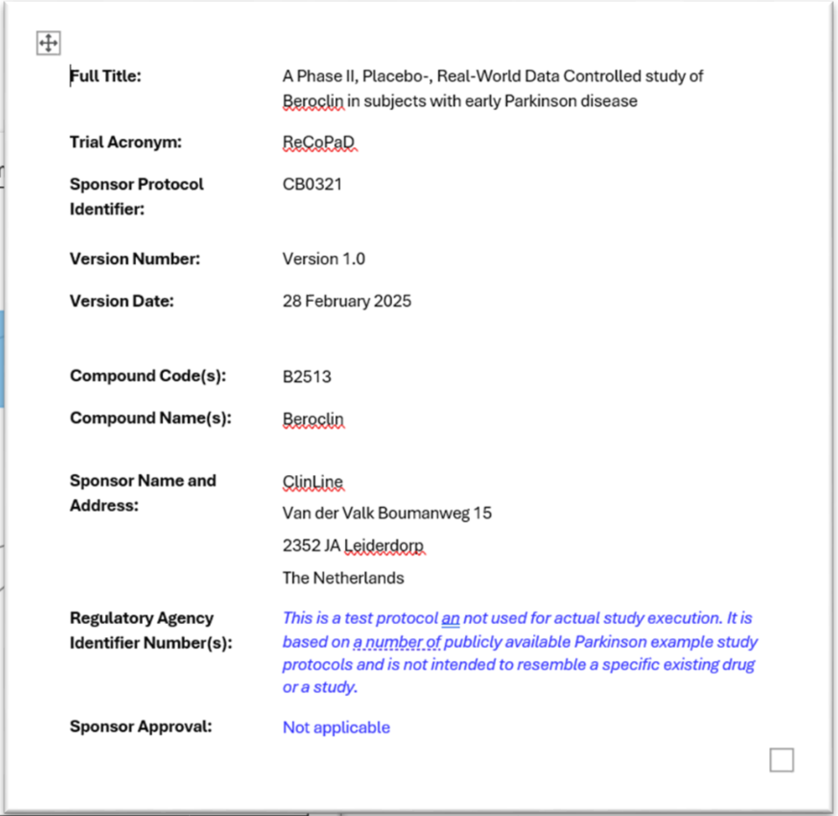
Furthermore, RWD Lineage will be represented in a standard metadata model that can be used along with SDTM to support the use of RWD with SDTM. This standardization will allow lineage to be combined across disparate data sources and will enable the unification of downstream tools, workflows, and analytics for validation and audit activities. Visit the **RWD Lineage Wiki** for updates.

How to Participate

We invite your organization to participate in this collaborative initiative. Please visit the **Participate tab** to learn more.

Parkinson Use Case





Full Title:	A Phase II, Placebo-, Real-World Data Controlled study of Beroclin in subjects with early Parkinson disease
Trial Acronym:	ReCoPaD
Sponsor Protocol Identifier:	CB0321
Version Number:	Version 1.0
Version Date:	28 February 2025
Compound Code(s):	B2513
Compound Name(s):	Beroclin
Sponsor Name and Address:	ClinLine Van der Valk Boumanweg 15 2352 JA Leiderdorp The Netherlands
Regulatory Agency Identifier Number(s):	*This is a test protocol an not used for actual study execution. It is based on a number of publicly available Parkinson example study protocols and is not intended to resemble a specific existing drug or a study.*
Sponsor Approval:	Not applicable

Registry study used for submission

- Rare disease indication
- Registry data used as external control arm
- Lineage
  - Data
  - Documentation
  - Procedures
  - Exchange
- Matching and validity
- Cooperation
- Missingness

Projects and Solutions

- Use cases
- Open-source tool development
- USDM utilization
- Industry initiatives
  - CDISC
  - PHUSE
  - TransCelerate
- Vendor/system requirements and implementation

Thank You!

B.Snoeijer@ClinLine.eu