SYNTHEMA + ERN-EuroBloodNet

Joint Training Programme on Synthetic Data Generation in SCD and AML

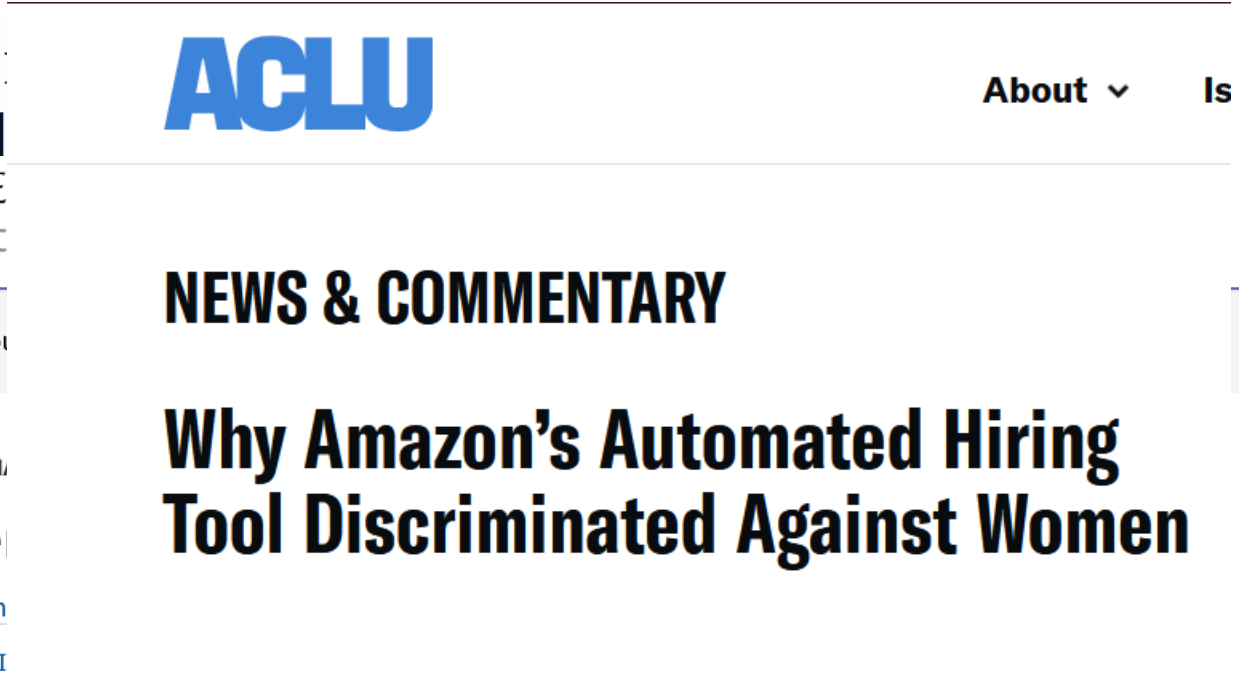
Funded by the European Union

Data Quality, Standardisation, and Interoperability

Sara Reidel- Biostatistician| Data Driven Lead, VHIR

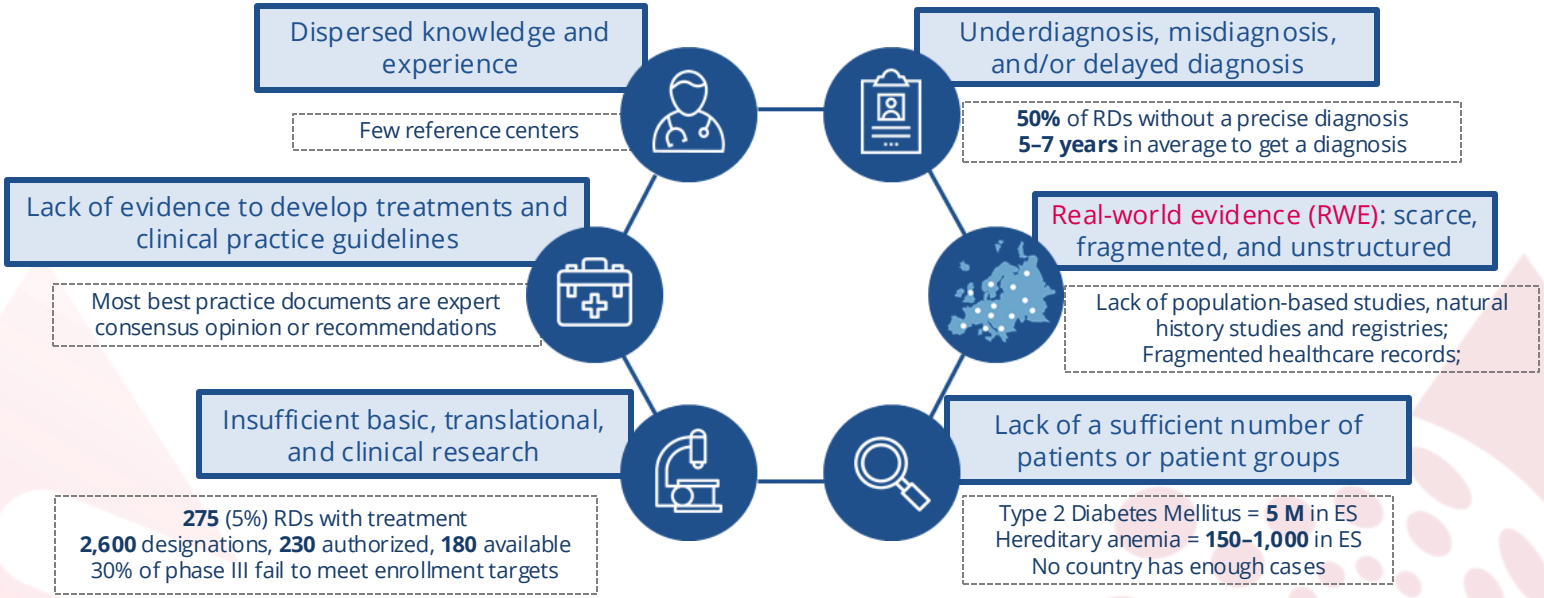
22nd May 2026

Why are we talking about data quality in AI?



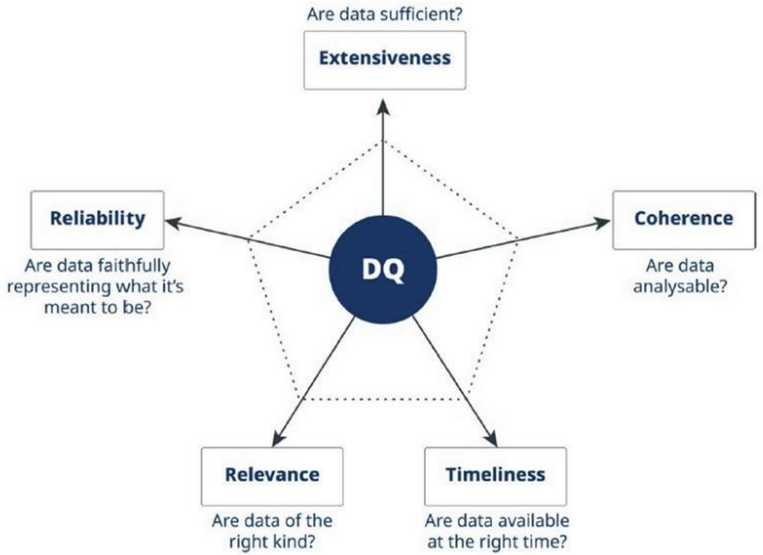
The model learned correctly from the data → the problem was in the data.

Context for RD in EU: +7000 RDs, 30Mi PLWRD in Europe



EMA (European Medicines Agency) Data Quality Framework (DQF) for RWE Generation

It establishes quality dimensions and a maturity model with the aim of guiding the progressive evolution of data quality systems and processes used in regulatory decision-making, defining 4 levels based on functional determinants of data quality.



- **Level 1, Documented:** This is the starting level. Organizations collect essential documentation, such as data dictionaries, operational processes, variable definitions, data capture flows and information transformations, although implementation is not uniform or standardized.
- **Level 2, Formalized / Standardized** This level brings structural coherence to the system and enables reproducibility and comparability across centers or countries.
- **Level 3, Implemented** Once processes are standardized, they can be systematized.
- **Level 4, Automation** A system and infrastructure 'understandable' by machines.

EHDS Article 78: data quality and utility label

1. Datasets made available through health data access bodies may obtain from health data holders a Union label relating to data quality and utility.
2. Datasets containing electronic health data collected and processed with the support of national or Union public funding shall have a data quality and utility label covering the elements referred to in paragraph 3.
3. The data quality and utility label shall cover the following elements, where applicable:
    - for data documentation: metadata, supporting documentation, the data dictionary, the format and standards used, the source of the data and, where applicable, the data model;
    - for technical data quality assessment: completeness, uniqueness, accuracy, validity, timeliness and consistency of the data;
    - for data quality management processes: the level of maturity of data quality management processes, including review and audit processes and bias assessment;
    - for coverage assessment: the time period, population coverage and, where applicable, the representativeness of the population included in the sample, and the average time frame during which a natural person appears in a dataset;
    - for access and provision information: the time elapsed between the collection of electronic health data and their inclusion in the dataset, and the time limit for providing the electronic health data following the issuance of a data permit or the approval of a data access request;
    - for information on data modifications: the combination and integration of data into an existing dataset, including links with other datasets.

The RADeep registry

- To enable epidemiological and health burden surveillance of RADs in the EU to improve healthcare planning
- To enable translational and clinical research by collecting enough amount of high quality real world data to generate real world evidence for identification of reliable biomarkers for:
  - Disease progression
  - Prognosis
  - Response to treatments

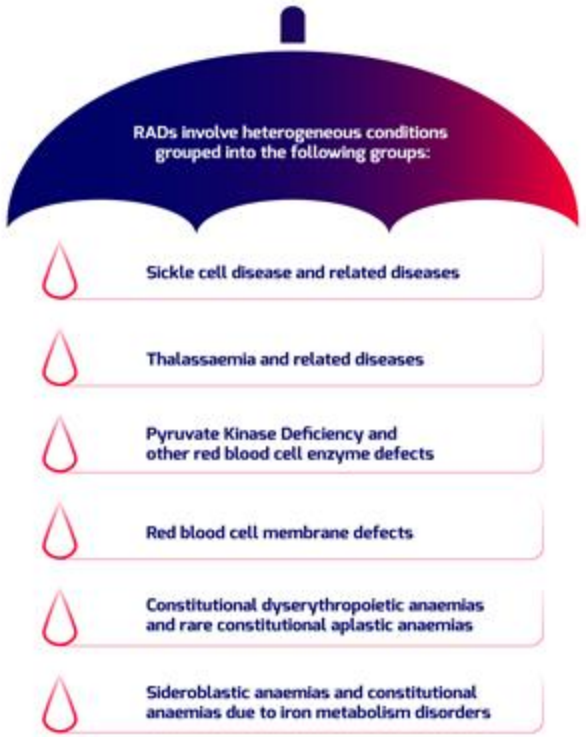
→ possibility to share and pool data

→ reach critical numbers

→ perform clinical trials, research projects

→ knowledge generation (evidence)

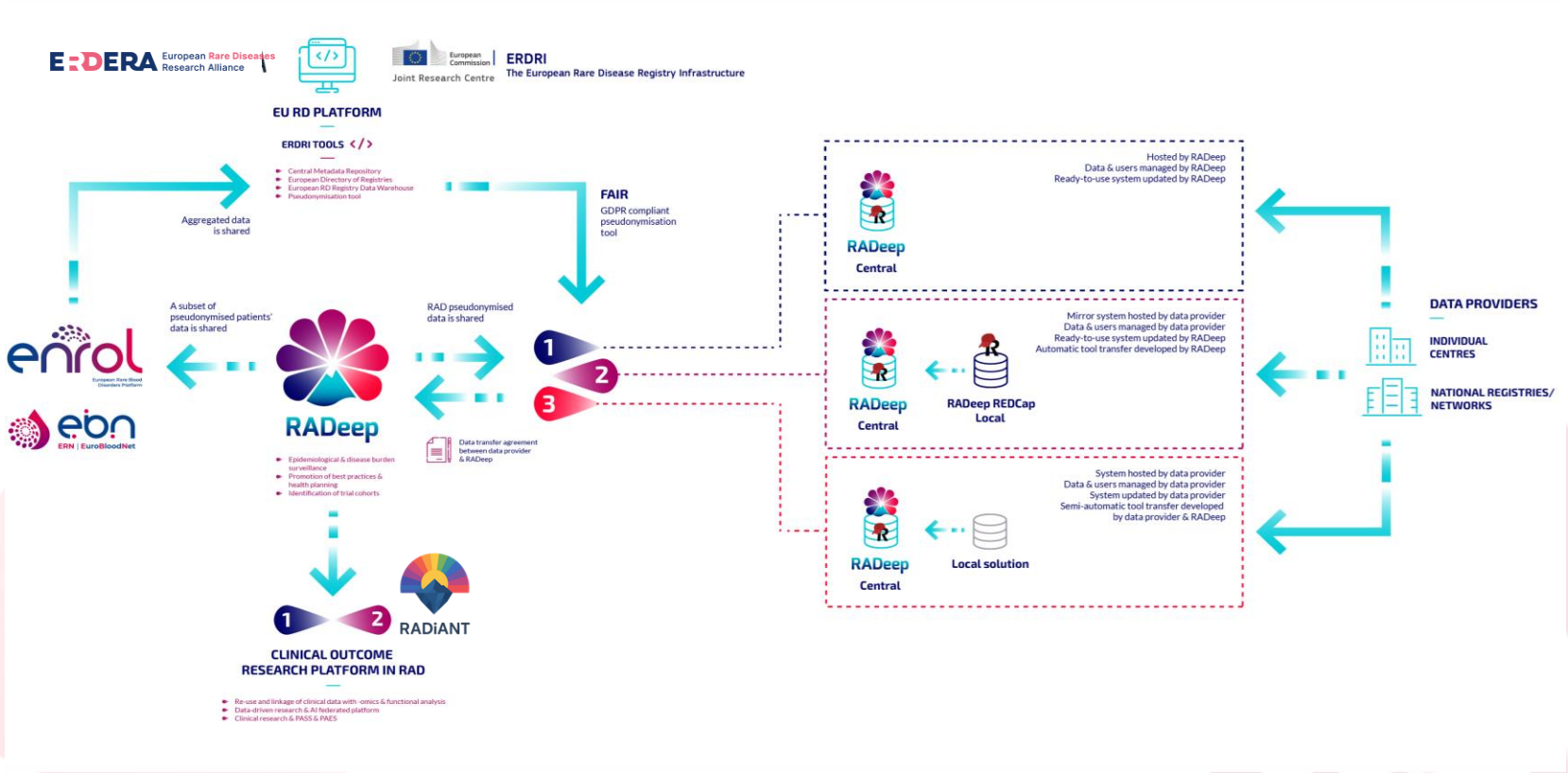
→ **better healthcare for RADs patients**



eha **Officially endorsed by EHA**
07th November 2024

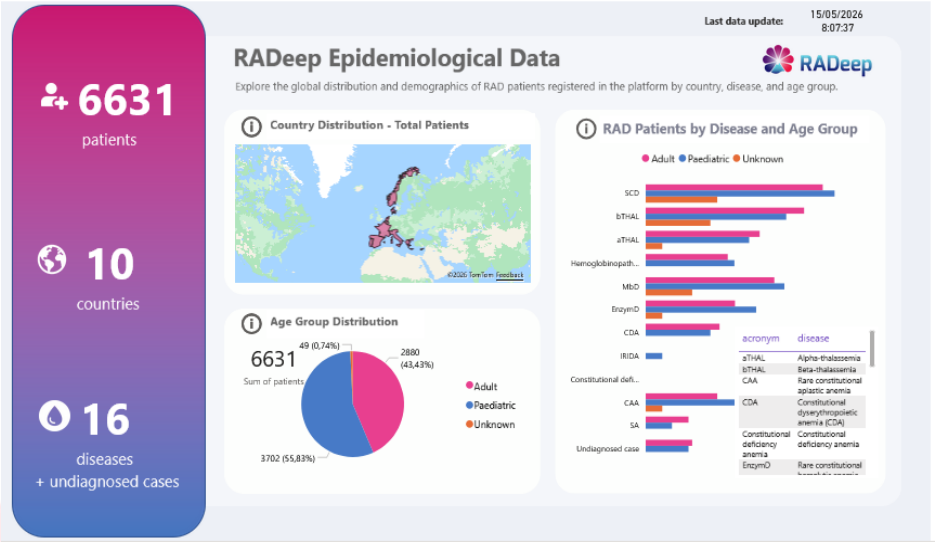
ebn SYNTHEMA

RADeep Registry Ecosystem

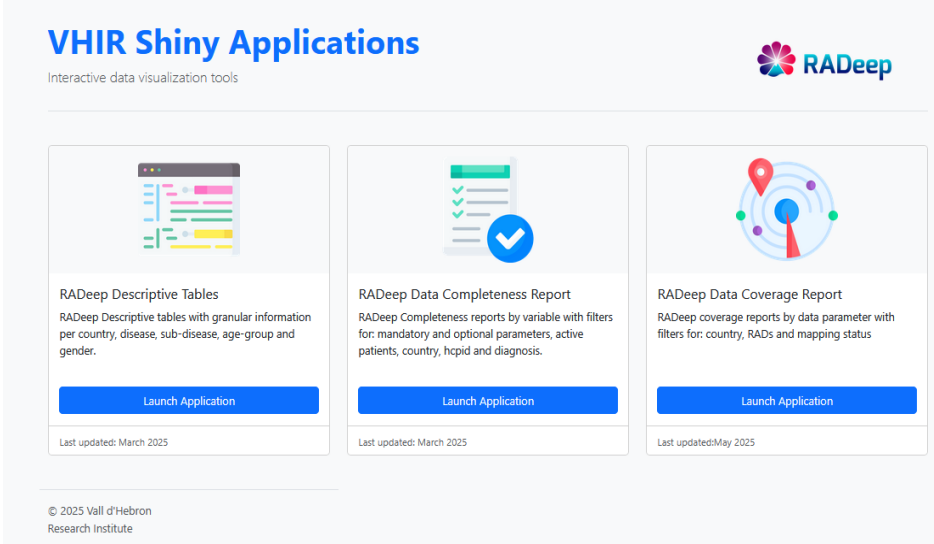


RADeep: EMA Data Quality Framework

Public dashboard



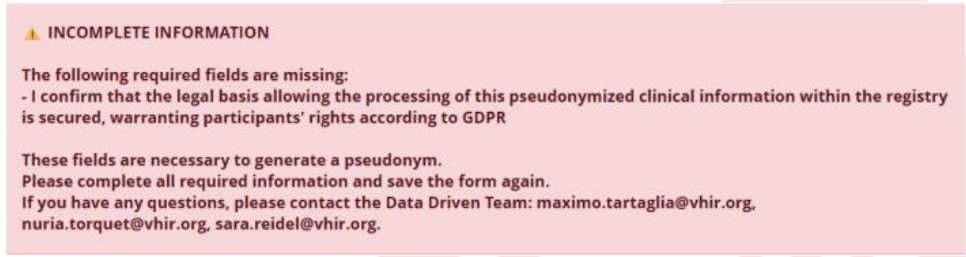
Private dashboards



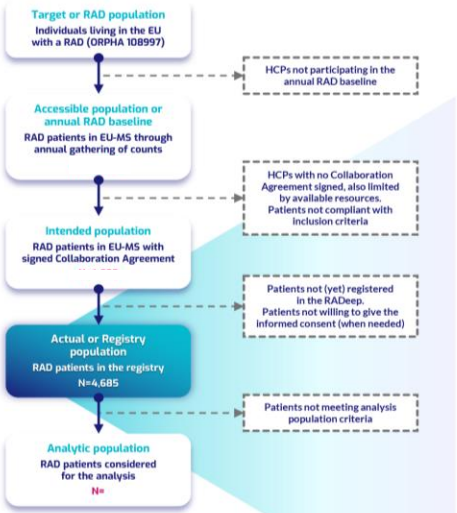
Pseudonym Module

- **Patient Registration** form contains all the information necessary for pseudonym generation and duplicate detection:
  - Inclusion criteria:
    - ❑ Consent registry checked
    - ❑ Patient must be alive at inclusion in the registry
    - ❑ Diagnosis under ORPHA code 108997
  - For duplicate detection:
    - ❑ Diagnosis
    - ❑ Sex at birth
    - ❑ Country of birth
    - ❑ Date of birth
  - For trazability:
    - ❑ Healthcare provider

The module will generate an **alert for review** if any of these required fields are not completed upon saving the form.



RADeep: EMA Data Quality Framework

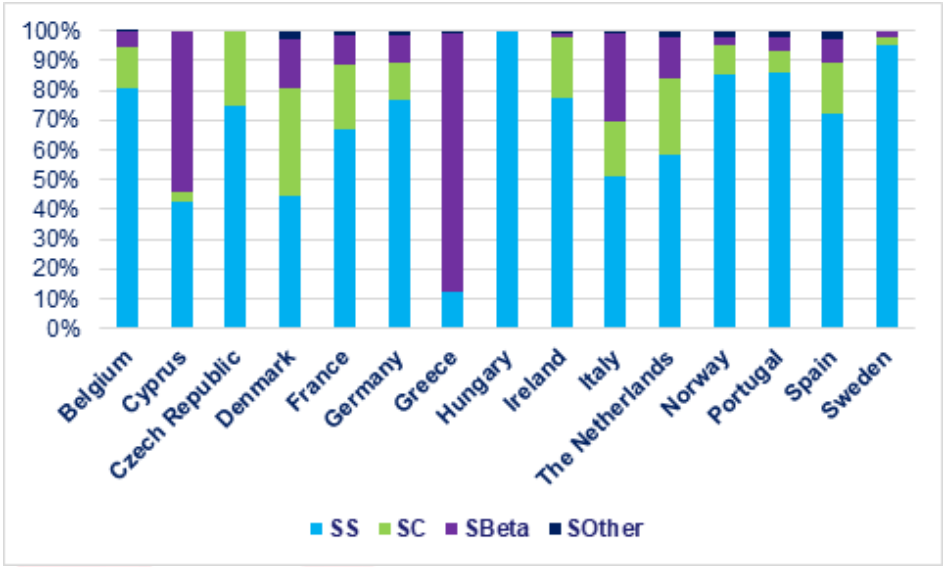


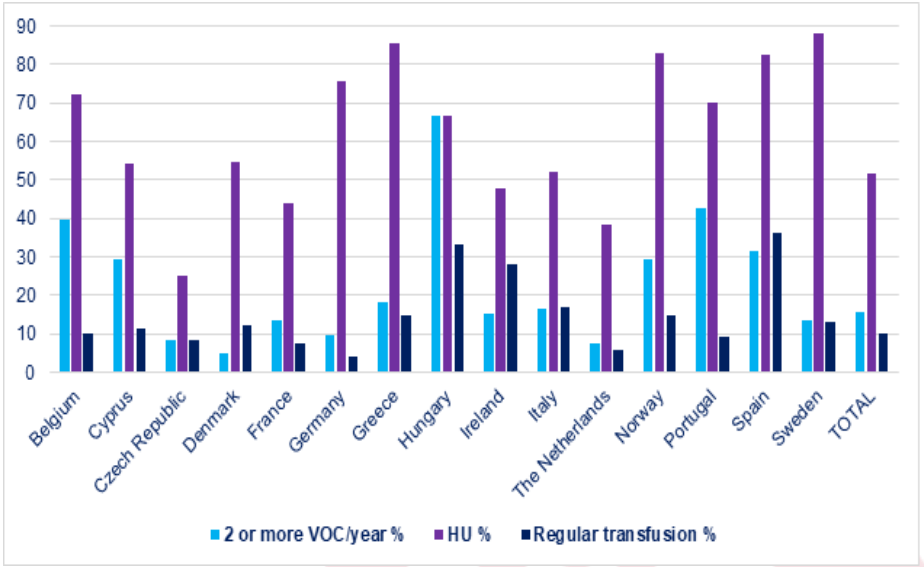
Group		Subgroup		Parameters	Mandatory	Longitudinal
1	Patient pseudonym, permissions and biobanking information	1.1.	Pseudonym	1	1	0
		1.2.	Permissions	5	5	0
		1.3.	Biobanking	3	3	0
2	Demographics	2.1.	Population distribution	12	8	4
		2.2.	Mortality and comorbidities	7	7	6
3	Diagnosis	3.1.	Diagnosis	6	6	0
		3.2.	Genotype	5	5	0
		3.3.	Undiagnosis	2	2	0
		3.4.	Disease Onset	4	2	0
		3.5.	Neonatal Manifestations	5	0	0
4	Physical examination	4.1.	Development	0	0	0
5	Organ Damage	5.1.	Organ damage assessment	14	3	14
		5.2.	Chronic complications of bones and extremities	8	4	8
		5.3.	Chronic cardiac and pulmonary disease	6	5	6
		5.4.	Chronic neurological disease	7	6	7
		5.5.	Chronic endocrinologic disease	8	7	8
		5. 6	Chronic liver and renal disease	8	7	8
		5.7.	Visual and hearing disease	5	4	5
6	Acute Complications	6.1.	Acute complications in RADs (Except SCD) requiring hospitalization or emergency admission for more than 24 hours	4	3	4
		6.2.	Acute complications in SCD requiring hospitalization or emergency admission for more than 24 hours	4	3	4
		6.3.	Intesive Care Unit Admission in the last 12 months	1	1	1
7	Clinical manifestations and surgery	7.1.	Spleen	4	4	4
		7.2.	Gallbladder	3	0	3
8	Treatments	8.1.	Blood transfusion	14	7	13
		8.2.	Chelation	5	5	5
		8.3.	Hydroxyurea	4	1	4
		8.4.	Specific treatment(s)	2	2	2
		8.5.	Haematopoietic stem cell transplantation (HSCT) / gene therapy	5	5	5
		8.6.	Inclusion in clinical trial protocol	2	2	2
9	Fertility	9.1.	Fertility and Offspring	2	2	2
10	Disability	10.1	Disability	3	1	3
11	Laboratory tests	11.1	Complete blood count	17	17	17
		11.2	Biochemical tests	14	14	14
		11.3	Hemoglobin tests	7	4	1
		11.4	Enzyme tests	17	17	0
		11.5	Membrane tests	3	3	0
		11.6	Other Laboratory tests	4	4	0
			Total	221	170	150

- Study populations
- Data dictionary
- Metadata available in catalogues
- OMOP CDM

Sickle cell disease use case

Clinical outcomes heterogeneity- Accessible population



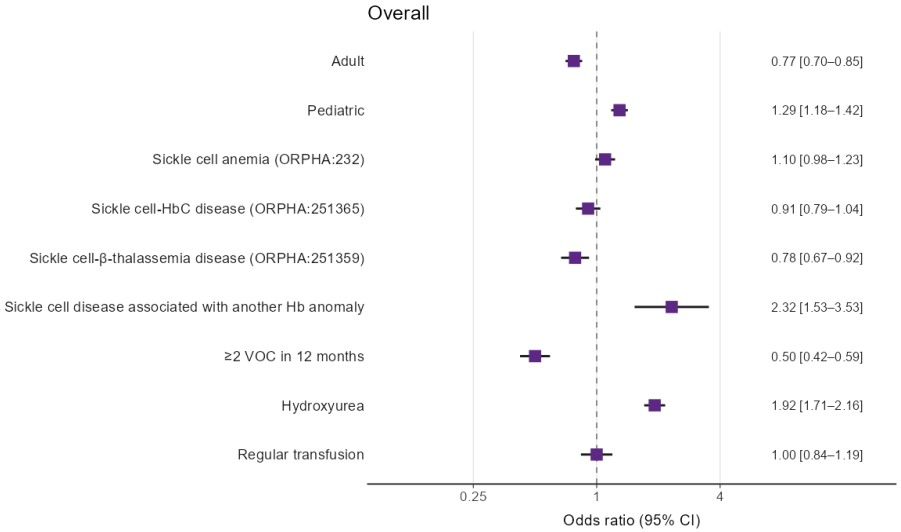


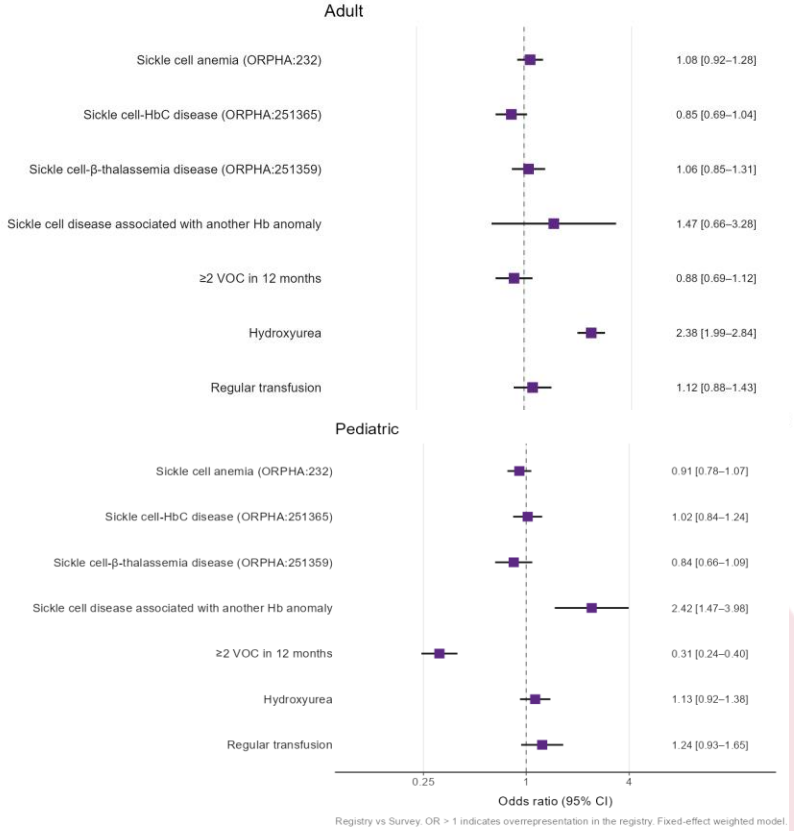
Country distribution of SCD patients by genotype

Country distribution of SCD patients by severity: % of patients with 2 or more VOC per year, on Hydroxyurea (HU) treatment, on regular transfusion treatment

Sickle cell disease use case

Bias assessment- Registry population





Common Data Models

We have the data. A Common Data Model ensures machines can actually read it, turning siloed, inconsistent records into a unified structure that speaks the same language, everywhere.

	OMOP CDM	HL7 FHIR
Primary purpose	Observational research and population-level analysis	Real-time clinical data exchange
Main environment	**Most widely used CDM in hospital and research settings globally**	Clinical systems, EHRs, health apps
Model type	Standardized relational model (SQL tables)	Modular resources in JSON/XML
Vocabularies	Unified (SNOMED, ICD, RxNorm, LOINC… all mapped to a single system)	Flexible; supports multiple terminologies without forced mapping
Interoperability	Across institutions sharing the same model (federated analysis)	Across heterogeneous systems in real time (REST APIs)
Learning curve	High (requires ETL + vocabulary mapping)	Medium-high (requires API implementation)
Typical use cases	Pharmacovigilance, RWE, clinical trials, epidemiology	EHR, telemedicine, wearables
Regulatory alignment	Strongly aligned with EMA and FDA for RWE	Backbone of patient data initiatives (GDPR, 21st Century Cures Act)
Compatible?	**OMOP ↔ FHIR and OMOP ↔ CDISC pipelines exist**	FHIR ↔ OMOP and FHIR ↔ CDISC mappings in development

Introduction to SYNTHEMA & synthetic data in healthcare

Giulio Spinozzi, PhD
Chief Technology Officer

Datawizard Srl

#3

Data model and transformation plan

SYNTHEMA T1.3

Task 1.3 focused on building and testing an ETL process to transform clinical data into the **OMOP Common Data Model (CDM)**.

The aim was to make all data standardized, compatible, and reusable across the consortium.
This standardization allows further activities such as data validation and synthetic data generation.

The ETL pipeline was created and applied to two datasets: **AML** (Acute Myeloid Leukemia – Humanitas) and **SCD** (Sickle Cell Disease - Vall d'Hebron) using a shared architecture and the **OHDSI tools**.

ETL Workflow and Architecture

SYNTHEMA T1.3

The ETL process was designed to be **modular, reproducible, and reusable**, using a Docker based environment with PostgreSQL as the core database.

It integrates both OHDSI tools and custom scripts:

- **WhiteRabbit** → for data profiling and structure analysis
- **Rabbit-in-a-Hat** → for mapping source variables to OMOP concepts
- **Python and SQL scripts** → for data transformation, validation, and export

Recently additional scripts were introduced to:

- perform **data validation and consistency checks**
- generate **flat standardized outputs** for sharing and inspection

The pipeline can be executed on any machine configured with **Docker and the required dependencies**, ensuring reproducibility across environments.

Analyzed Datasets (AML and SCD) - 1

SYNTHEMA T1.3

AML dataset (Humanitas)

A preliminary test dataset was initially used to design and validate the ETL mapping.
This allowed us to verify the structure, logic, and consistency of the transformation process.
Recently, the **real AML dataset** has been received and will be processed in the next phase using the validated mapping and ETL scripts.

SCD dataset (Vall d'Hebron)

The SCD dataset, including around 1.000 clinical records, has been fully transformed into the OMOP CDM.
Validation using OMOP data quality tools is currently ongoing.

Analyzed Datasets (AML and SCD) - 1

SYNTHEMA T1.3

Mapping challenges

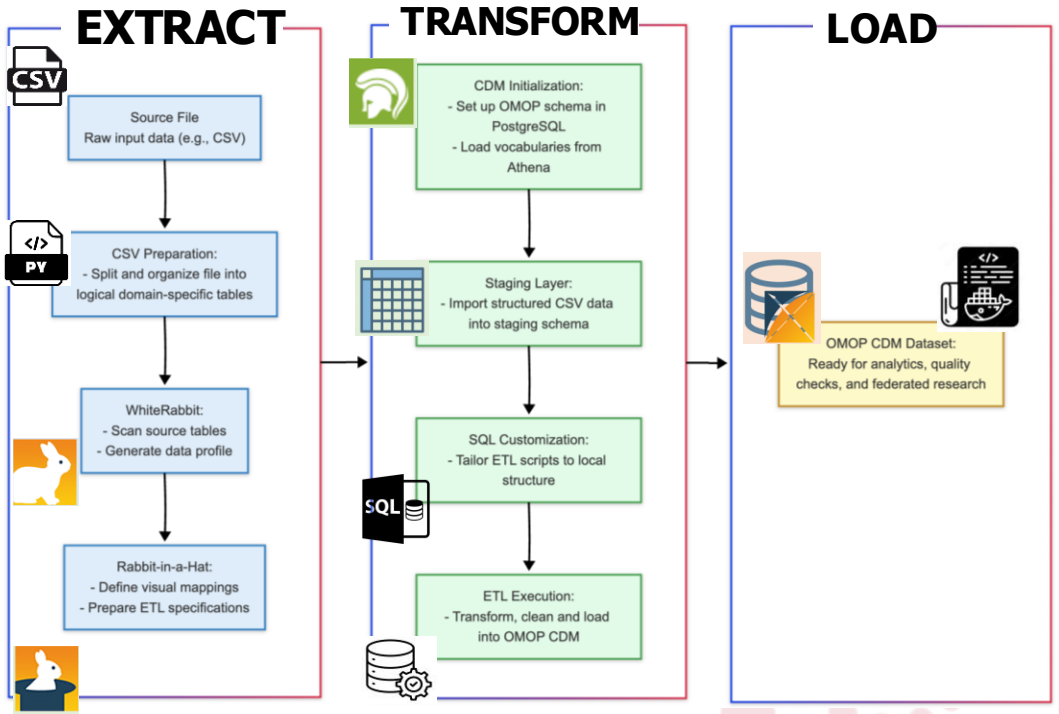
Some genetic mutations in the AML dataset were not available in OMOP vocabularies, so we used broader concepts to keep the information consistent.

In the SCD dataset, some laboratory and clinical parameters were not covered by the standard vocabularies, so we created custom entries in the OMOP vocabulary tables to retain all relevant data and maintain completeness

ETL flow

- **Extract** phase includes CSV preparation, source profiling, and conceptual mapping
- **Transform** phase includes schema setup, staging load, and SQL customization
- **Load** phase concludes the process with data standardized into OMOP CDM

The ETL is containerised and scriptable, ensuring scalability across new datasets or updates.



Validation Process

SYNTHEMA T1.3

After the ETL process, each dataset goes through two validation steps to ensure that the transformed data is correct and reliable.

Clinical Validation

Performed by the clinical partners (Humanitas for AML and Vall d'Hebron for SCD).
This step verifies that all variables are correctly transformed, values are consistent, and the clinical
meaning of the data is preserved.

OMOP Data Quality Validation

Once the clinical review is completed, the OMOP CDM data is validated using **OMOP quality**
tools such as **Achilles**.
These tools automatically check table completeness, data consistency, and vocabulary mapping.

Thanks!

Any questions?

Keep in touch!

eurobloodnet.eu /ERNEuroBloodNet @ERNEuroBloodNet @erneurobloodnet.bsky.social

synthema.eu /synthema @SYNTHEMA_EU @synthemaeu.bsky.social

Funded by
the European Union

Acknowledgements

European Reference Network
for rare or low prevalence complex diseases
Network Hematological Diseases (ERN EuroBloodNet)

Funded by the European Union

This project is supported by the European Reference Network on Rare Haematological Diseases (ERN-EuroBloodNet)-Project ID No 101085717. ERN-EuroBloodNet is partly co-funded by the European Union within the framework of the Fourth EU Health Programme.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

SYNTHEMA

Funded by the European Union

SYNTHEMA is an initiative funded by the European Union's Horizon Europe Research and Innovation programme under grant agreement No. 101095530.