SYNTHEMA + ERN-EuroBloodNet

Joint Training Programme on Synthetic Data Generation in SCD and AML

Funded by the European Union

Synthetic data generation and anonymization methodologies

Sofia Tsekeridou, Netcompany, sofia.tsekeridou@netcompany.com
Imanol Isasa Reinoso, Vicomtech, iisasa@vicomtech.org

May 22nd, 2026

Anonymization for Clinical Data Sharing
Dataset Discoverability – HealthDCAT-AP

Dr. Sofia Tsekeridou, Netcompany S.A.
E-mail: sofia.tsekeridou@netcompany.com

Clinical Data Sharing for Clinical Research and Innovation: Anonymization is a Must!

Siloed Data Centres

- For **data and AI-driven clinical research advancement** and **innovation acceleration**, need of **vital real-world data sharing** imperative!

Data Sharing for Innovation

- **Clinical contexts, heavily regulated** for personal/ sensitive data protection and privacy guarantees (GDPR, EHDS, data protection regulations, etc.), require **full data anonymization before sharing**

Apart from GDPR, EHDS Regulation for Secondary use of Clinical Data

- **European Health Data Space (EHDS)** aims to improve individuals' access to and control over their electronic health data while enabling **secondary use for research, innovation, and policymaking**.
  - Builds on GDPR, Data Governance Act, NIS2
  - Defines **clear roles and responsibilities for digital health authorities and health data access bodies**

Sources: *European Health Data Space Regulation (EHDS) - Public Health* , *EHDS Regulation in a nutshell*

Apart from GDPR, EHDS Regulation for Secondary use of Clinical Data

Secondary Use of Data

Defines **conditions under which health data can be accessed and processed for purposes beyond direct healthcare**, such as research, policy-making, and innovation.

Requires Member States to designate **Health Data Access Bodies (HDABs)** to **facilitate access to such data** while ensuring **security, privacy, and compliance with data protection laws**.

Defines **minimum categories of health data** that **must be made available for secondary use**, including electronic health records, genetic data, and healthcare-related administrative data.

Strict technical and legal safeguards set to **prevent re-identification** and ensure that **data processing remains in a secure environment**. **Anonymization, pseudonymization**, etc.

Researchers and organizations seeking access to health data **must apply for data permits**, and their requests are assessed based on transparency, necessity, and public interest.

Sources: European Health Data Space Regulation (EHDS) - Public Health , EHDS Regulation in a nutshell

Anonymization: Definitions

- **Direct identifiers**
  - **Directly** reveal the **identity** of a person in the dataset
    - Name, Address

- **Quasi Identifiers**
  - Cannot reveal the identity of a person in the dataset directly, but **can lead to identification if combined** with other information
    - Age, Gender, Occupation, etc.

Anonymization: SOTA Approaches

Technique	Description	Example(s)
Aggregation	Replace data values with a single “group” value, after having the data split into clusters.	Exact addresses are excluded from the data set, but the number of people living in that location is known.
Date shifting	Randomly shift a set of dates but preserve the sequence and duration of a period of time	Original Data: Date of Birth Patient1: 15/05/1990 Patient2: 15/11/1985 Anonymized Data: (+30 days) Patient1: 14/06/1990 Patient2: 15/12/1985

Anonymization: SOTA Approaches

Technique	Description	Example(s)
Encryption	Converts sensitive data to an apparently unreadable format. However, owners of the encryption key can convert back the data to their original form.	Original Data: The credit card number is 1234-5678-9012-3456 Anonymized Data: Convert it into Zm53241531XcFd74
Generalization	Reducing the resolution of the data (e.g., from exact date to year, from exact age to age range). Replace data values with approximations, after constructing a hierarchy.	Original Data: Date of Birth is 01/01/1990. Anonymized Data: Change 01/01/1990 to 1990 or to the range 1990-1995.
Masking	Replacing all or partial values with other characters of string data.	Original Data: patient_id is 12345 Anonymized Data: Change 12345 to 12***
Noise Addition	Adding generated “white noise” into original categorical or numerical data. Utility loss concern.	Original Data: Age is 25 Anonymized Data: Age is 27 (with added noise)

Anonymization: SOTA Approaches

Technique	Description	Example(s)
Pseudonymization	Replace directly identifying data (i.e. name, email etc.) with artificial identifiers (pseudonyms, essentially 1-1 mappings of values). Thus, it maintains a link to the original data that (under circumstances) allows for re-identification.	Original Data: Name is John Anonymized Data: Change John to Patient1
Suppression	Removing the attribute from the dataset where it's not necessary for analysis.	Original Data: Attributes include Country of Birth, Current City. Anonymized Data: Remove Current City.
Temporal Relativization	Express dates relatively to a reference date (i.e. the birth date). This method maintains temporal relationships while it obscures the actual dates.	Original Data: Date of Birth is 01/01/1990 and Date of Diagnosis is 03/01/1990 Anonymized Data: Change 01/01/1990 to 0 (date zero) and 03/01/1990 to 2 (2 days after date 0)
Tokenization	Replacing the attribute with a non-reversible token that can be used for joining data without revealing the original attribute.	Original Data: Patient ID as 12345. Anonymized Data: Replace 12345 with a token like ABCD-XYZ-1234-5678.

Anonymization: Metrics

Key to the process: **Utility versus Privacy Trade-off**!

Privacy metrics

- **k-anonymity**: Ensures each patient is indistinguishable from at least k–1 others based on quasi-identifiers.
- **l-diversity**: Ensures sensitive values within each group are sufficiently varied.

Utility metrics

- **Information loss**: Measures numerical distortion introduced by anonymization.
- **Statistical similarity** (for numerical variables): Compares means, variances, and correlations between the same attributes in the original and anonymized datasets.
- **Distributional consistency** (for categorical variables): Measures the extent to which category frequencies are preserved after anonymization. It is not available for some anonymization methods (i.e. suppression)

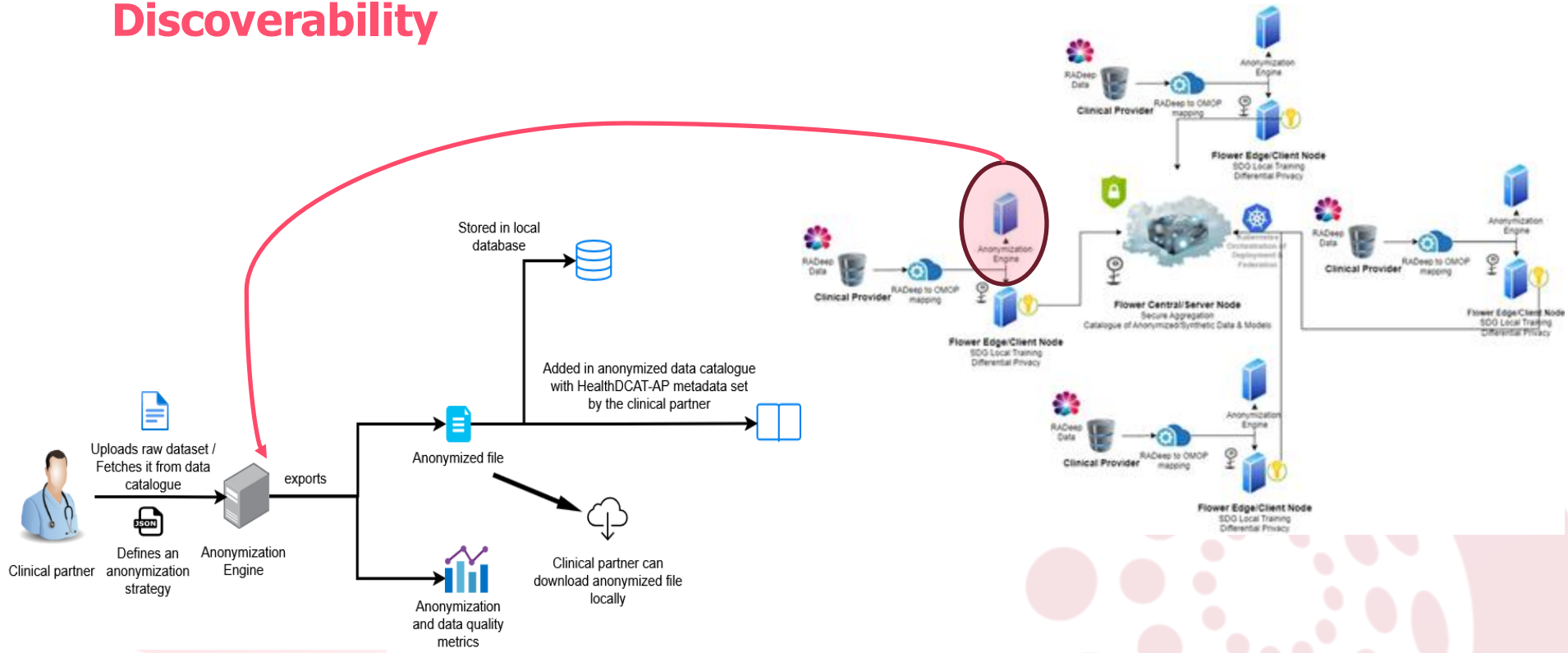
SYNTHEMA: How Anonymization is Addressed

- **Local Anonymization Engine** at each Clinical node with SOTA methods implementation

- **Federated Anonymization** across Clinical nodes network
  - Due to Data Scarcity (limited data at each node) and need for Quality Data of Sufficient Volume and Representability

- **Synthetic Data Generation → Synthetic Data are Anonymized Data as well!**

SYNTHEMA: Data Anonymization and Dataset Cataloguing for Discoverability



SYNTHEMA: Data Anonymization and Dataset Cataloguing for Discoverability

Process

- Define **direct and quasi-identifiers** and the **appropriate anonymization algorithms** per case
- Define **complete Anonymization Strategy for specific dataset** (and its structure/semantics) which leads to **specific utility versus privacy trade-offs** for the released anonymized datasets to ensure their quality for further data processing tasks for e.g. research purposes
- **Perform Anonymization** and **release Privacy and Utility metrics** for the anonymized dataset
- **Publish description of Anonymized dataset** in relevant **Public Catalogue** with access information to dataset

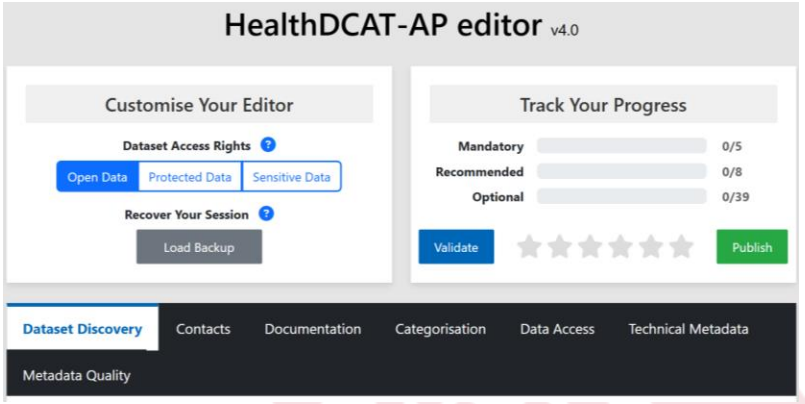
HealthDCAT-AP Compliant Datasets Annotation

For Cataloguing and Discoverability (Secondary Use)

Source: HealthDCAT-AP literacy portal (resulting from EHDS2 pilot and continuing in TEHDAS2)

Problem and necessity

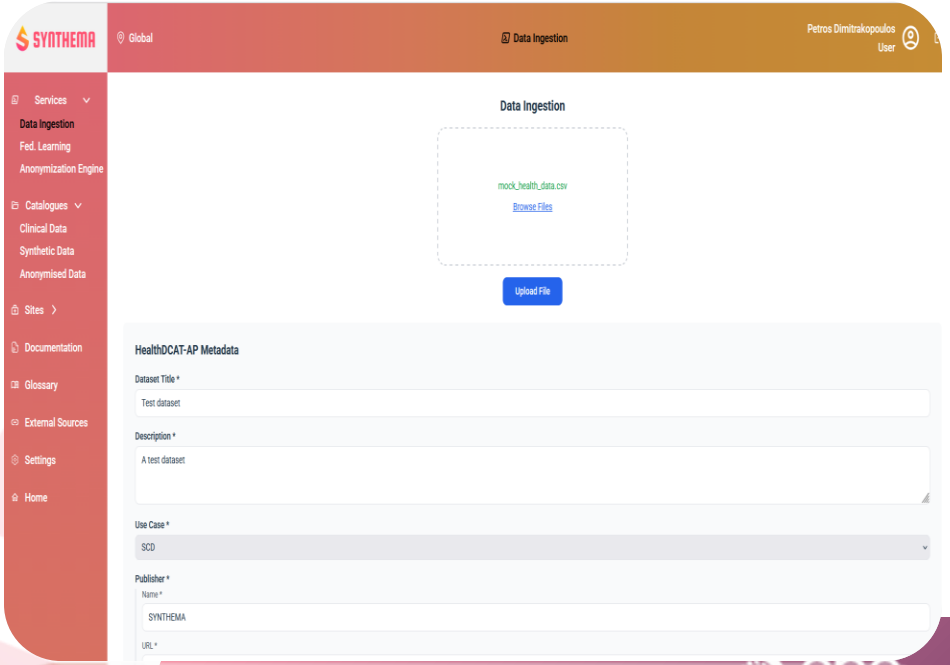
- DCAT-AP standard of choice of European Open Data Portal
- **HealthDCAT-AP standard of choice for EHDS**
  - It extends DCAT-AP to support metadata interoperability within the European Health Data Space (EHDS)

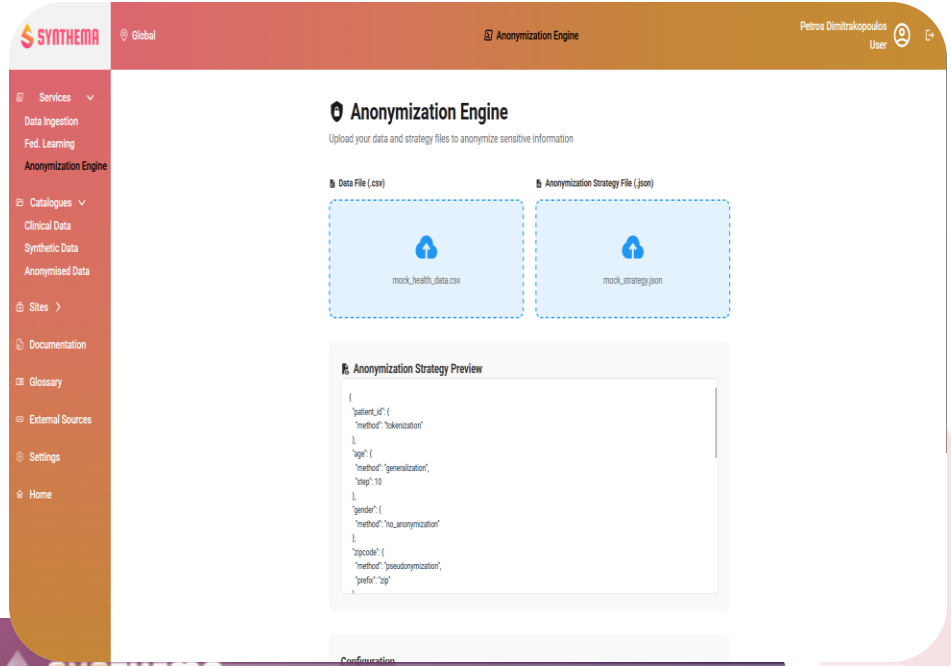


- **HealthDCAT-AP** ensures that **health-related datasets are uniformly described, promoting interoperability and enabling seamless data exchange** among researchers, public health institutions, policymakers, and other stakeholders

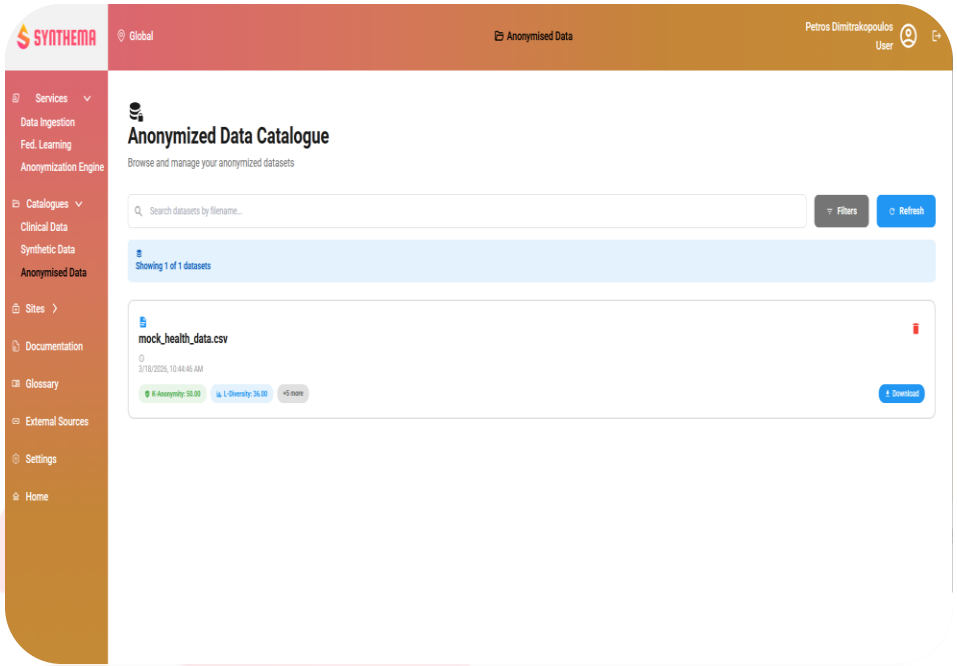
SYNTHEMA: Data Anonymization and Dataset Cataloguing for Discoverability

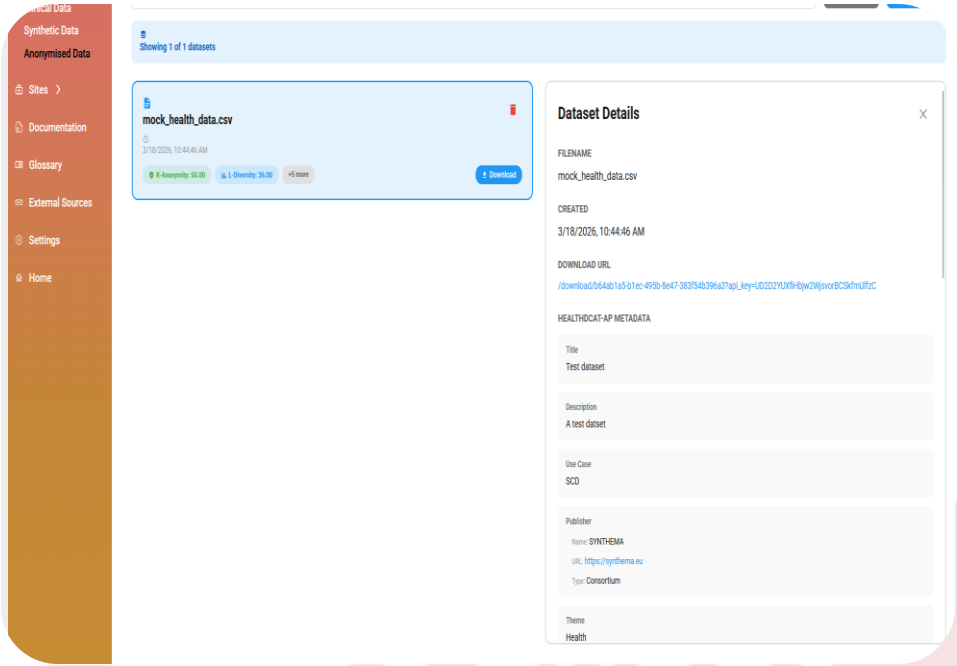
Data ingestion and Anonymization service
Includes HealthDCAT-AP mandatory metadata fields completion form for compliant annotations





SYNTHEMA: Data Anonymization and Dataset Cataloguing for Discoverability





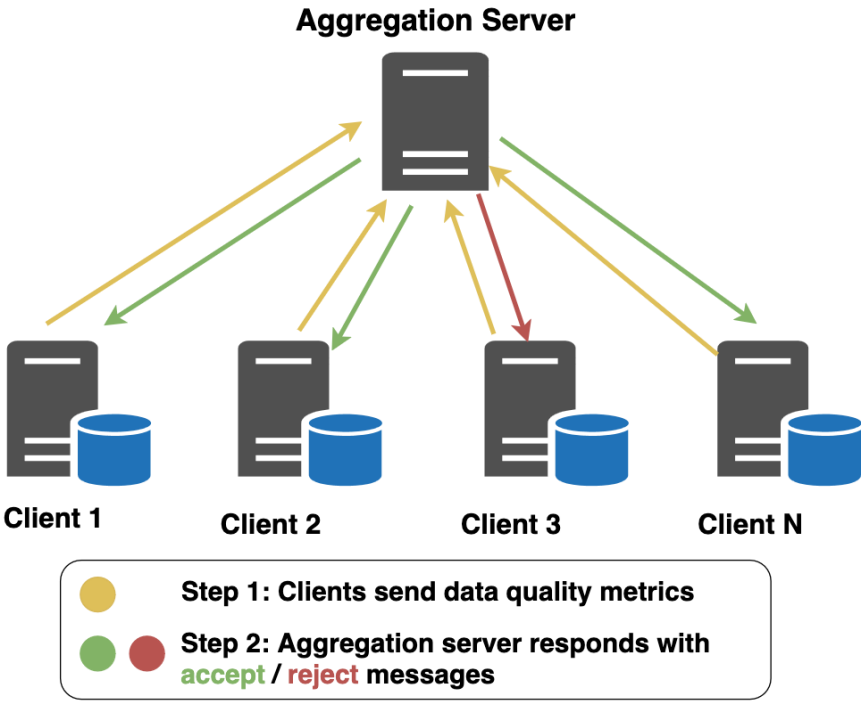
SYNTHEMA: Federated Anonymization

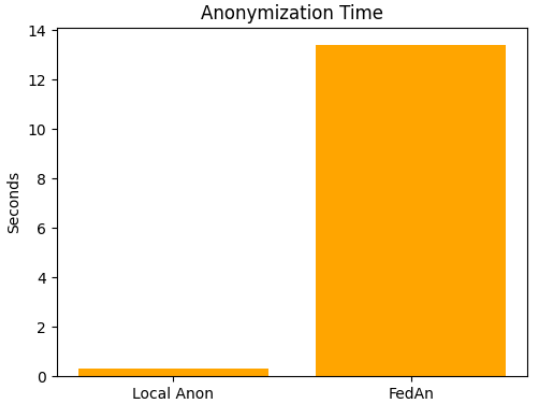
- **Problem**: Effective anonymization cannot always be achieved when data comes from a single data owner
  - Low volume – data scarcity (rare diseases, infrequent situations)
  - Not sufficiently representative
  - Quality issues

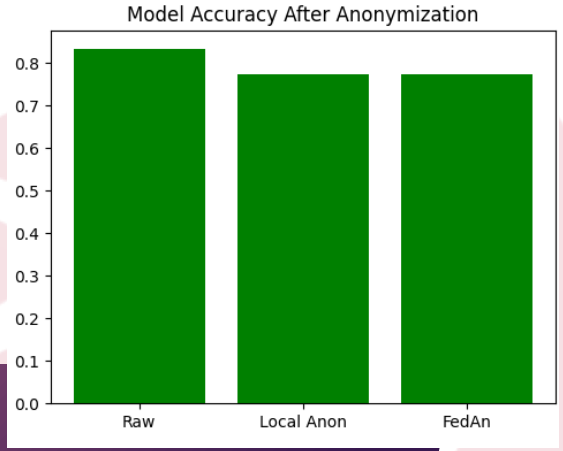
- **Impact**: This leads to sub-optimal anonymization of such data and thus increased re-identification risk, when these are shared

- **Solution**: Federated Anonymization
  - enables **collaborative data anonymization** at the **desired quality/utility trade-offs across distributed nodes and data sources**, guaranteeing desired volume and quality of anonymized data,

SYNTHEMA: Federated Anonymization





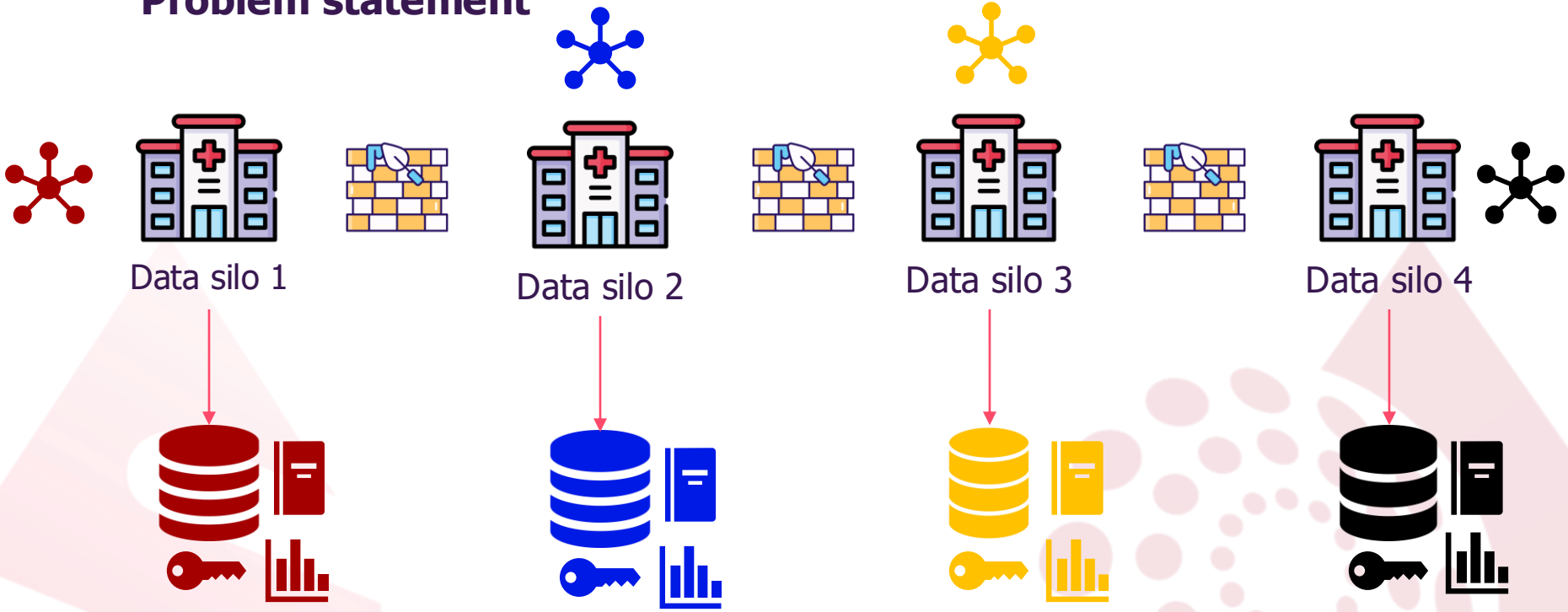


Synthetic data generation and federated learning

Imanol Isasa Reinoso, Vicomtech Foundation, BRTA
E-mail: iisasa@vicomtech.org

Synthetic data generation and anonymization

Problem statement

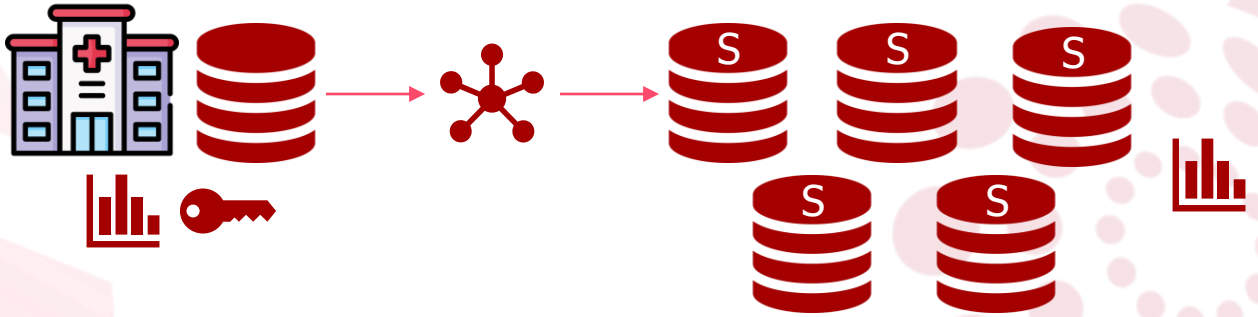


Synthetic data generation and anonymization

What is synthetic data?

"Synthetic data is artificial data that is generated from original data and a model that is trained to reproduce the characteristics and structure of the original data." - EDPS

- SD maintains real distributions and characteristics
- No one-to-one match between synthetic and real samples



Synthetic data generation and anonymization

Anonymization vs. Synthetic Data Generation

Real

ID	Age	BMI	Income
1	24	18.5	1500
2	63	20.0	1200
3	27	22.1	3000

Anonymized

ID	Age	Income
a342gfd25	20-30	1,000-3,000
b453jkd95	60-70	1,000-3,000
g534jda02	20-30	1,000-3,000

Privacy

Utility

Synthetic

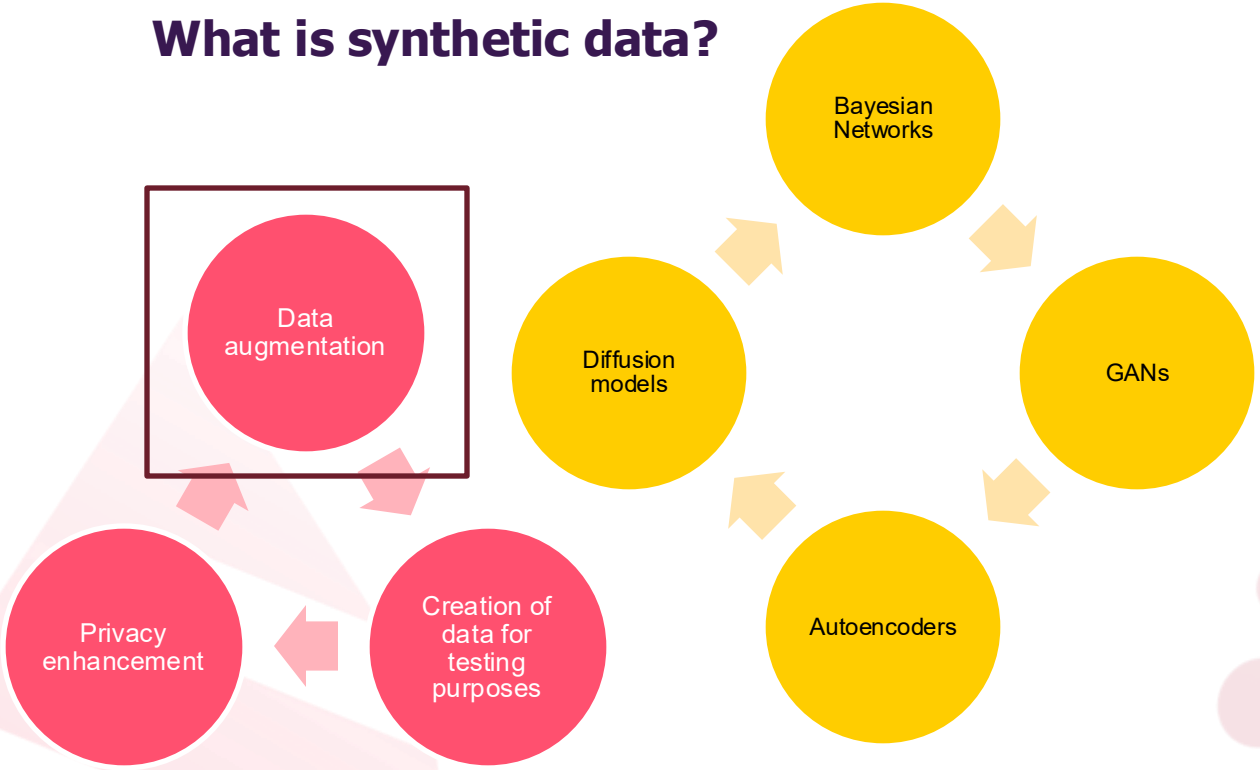
ID	Age	BMI	Income
1	27	18.8	2500
2	60	20.3	3200
3	31	19.2	2000

Privacy

Utility

Synthetic data generation and anonymization

What is synthetic data?

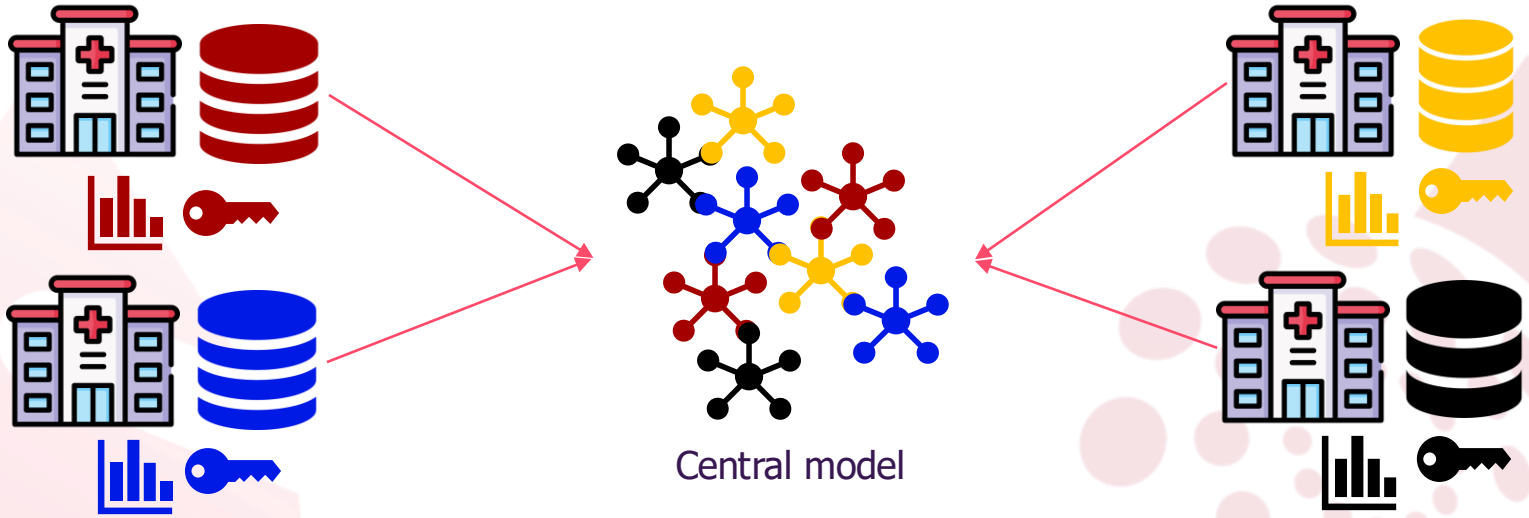


Limited to the representation of the input real dataset

Synthetic data generation and anonymization

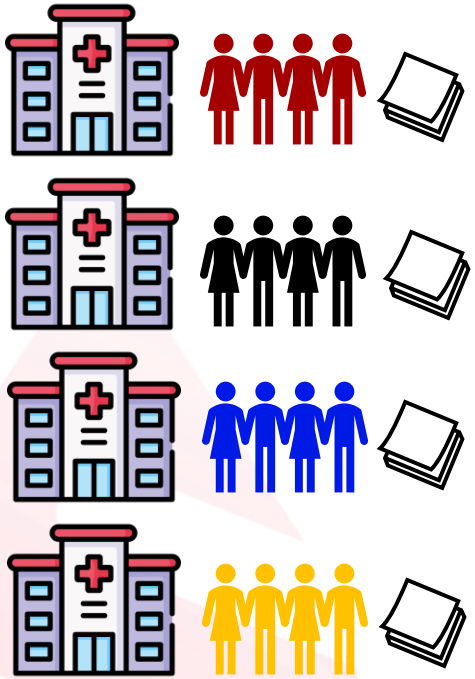
What is federated learning?

"Federated learning is a way of developing machine-learning models where each federated device shares its local model parameters instead of sharing the whole dataset used to train it." - EDPS

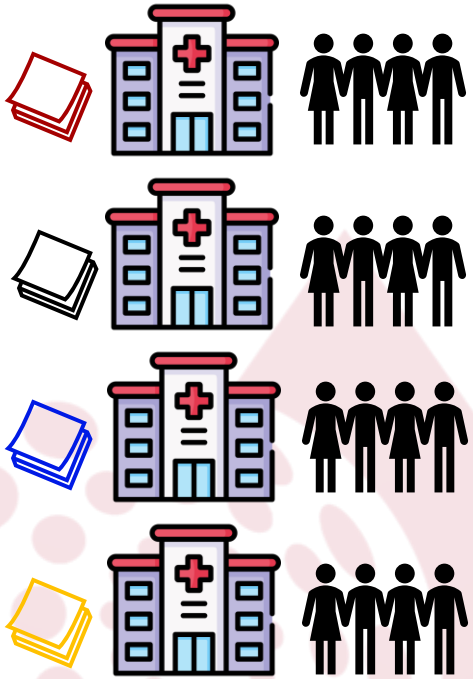


Synthetic data generation and anonymization

What is federated learning?

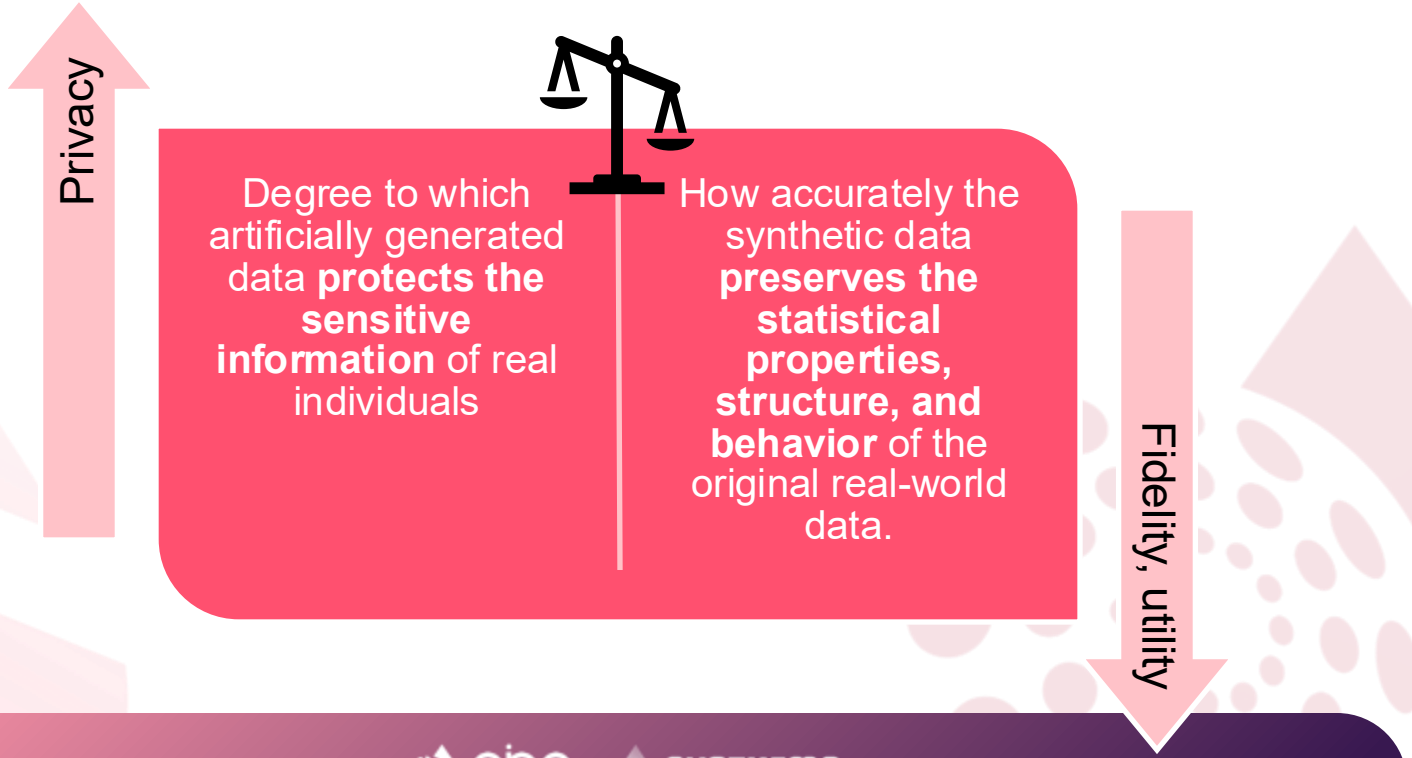


- **HORIZONTAL**: Same features, different origins (individuals).
- **VERTICAL**: Same origins (individuals), different features.



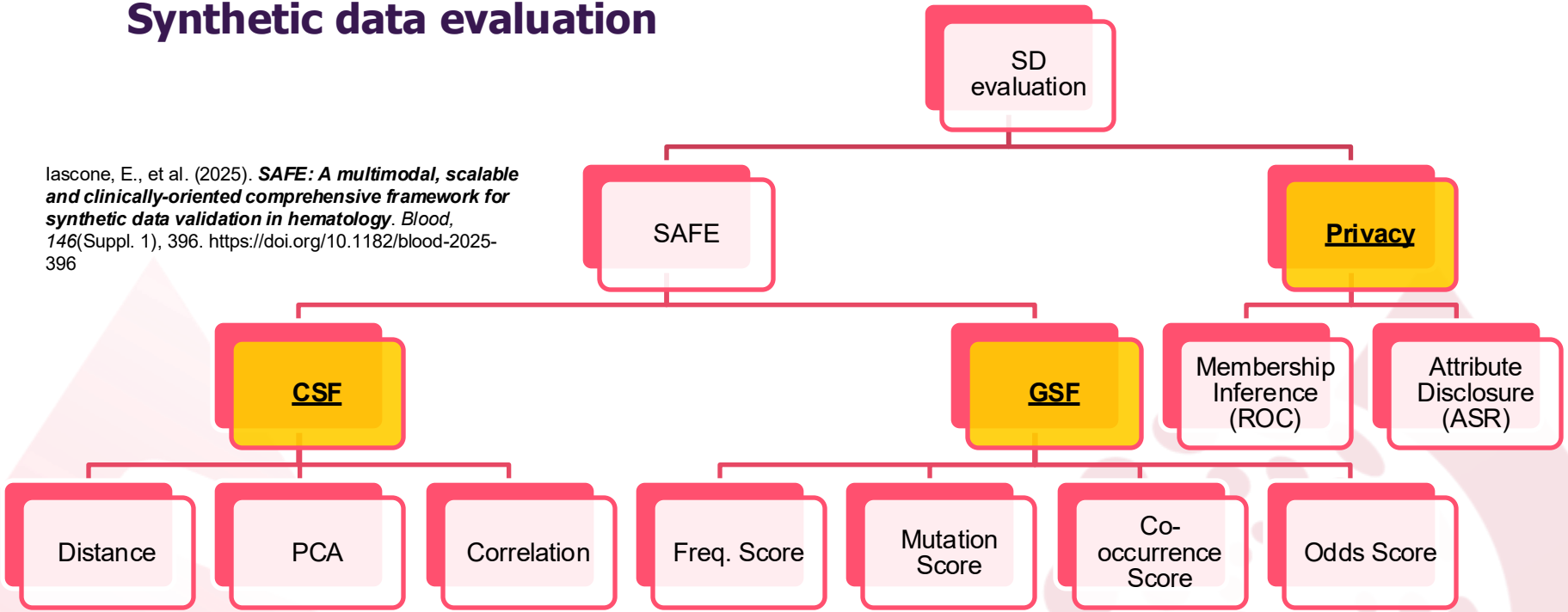
Synthetic data generation and anonymization

Synthetic data evaluation



Synthetic data generation and anonymization

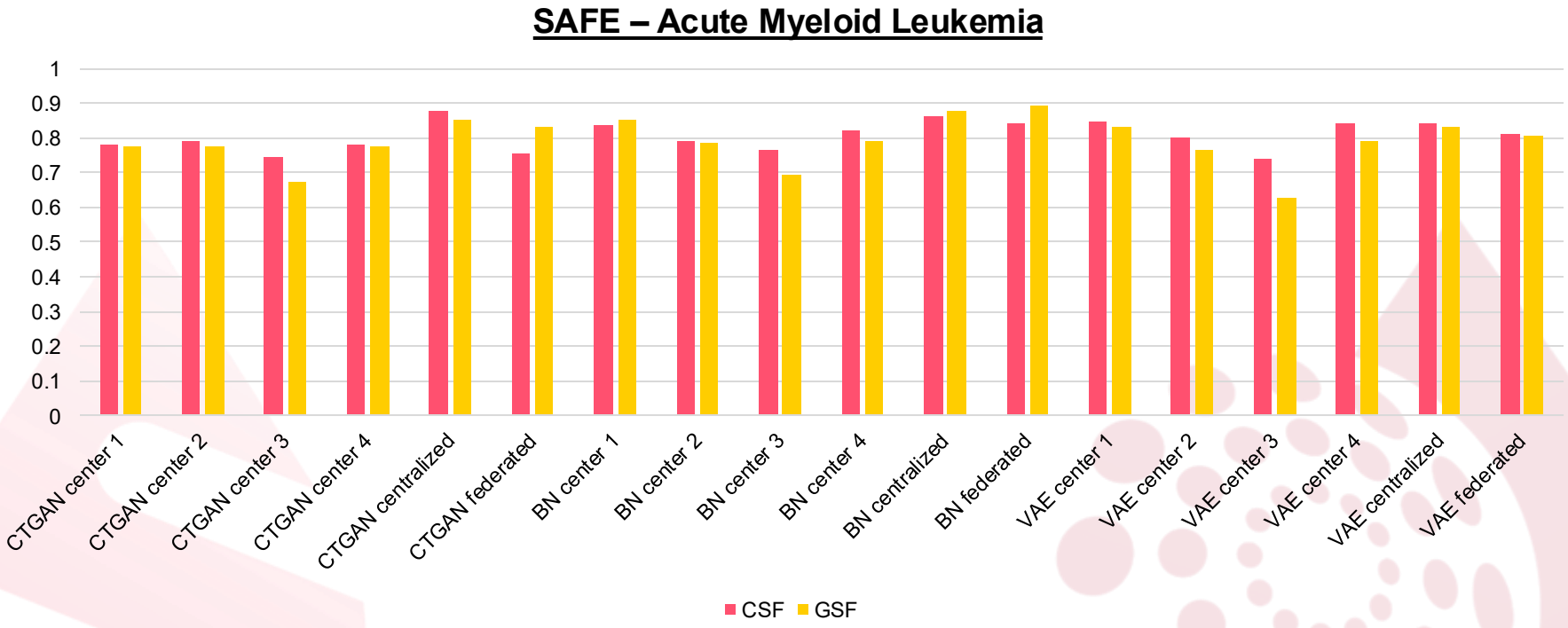
Synthetic data evaluation



Iascone, E., et al. (2025). ***SAFE: A multimodal, scalable and clinically-oriented comprehensive framework for synthetic data validation in hematology***. *Blood, 146*(Suppl. 1), 396. https://doi.org/10.1182/blood-2025-396

Synthetic data generation and anonymization

Real impact: SYNTHEMA as a use case

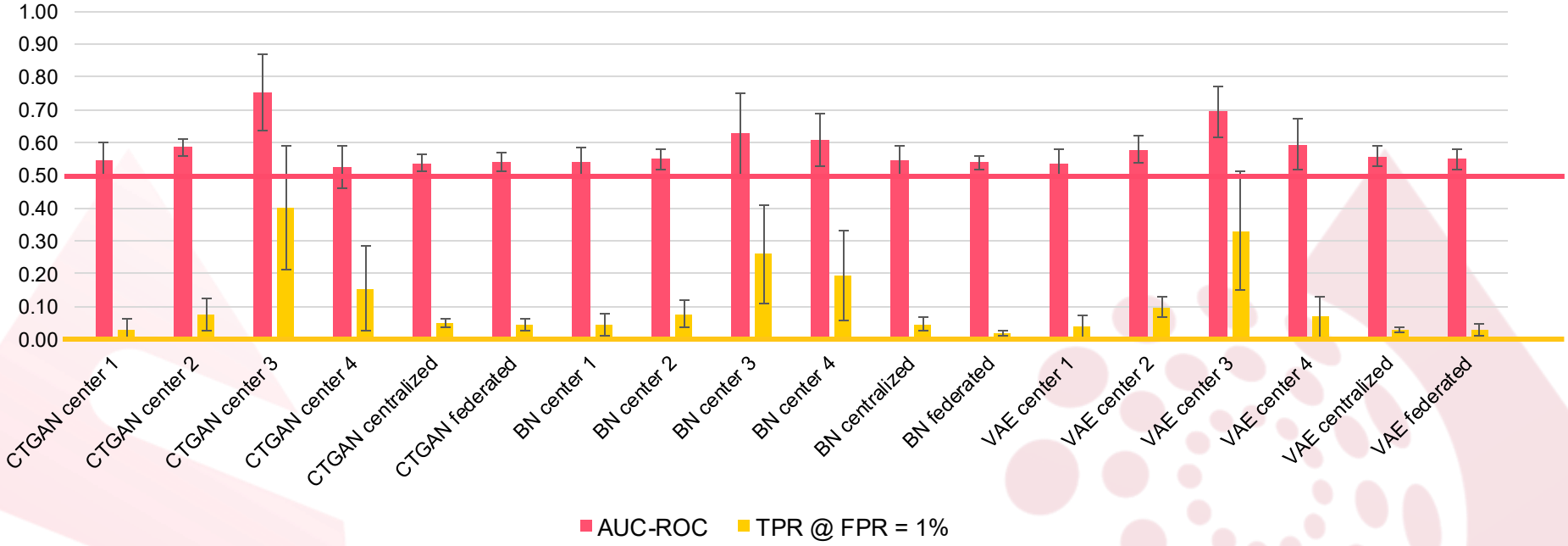


Synthetic data generation and anonymization

Real impact: SYNTHEMA as a use case

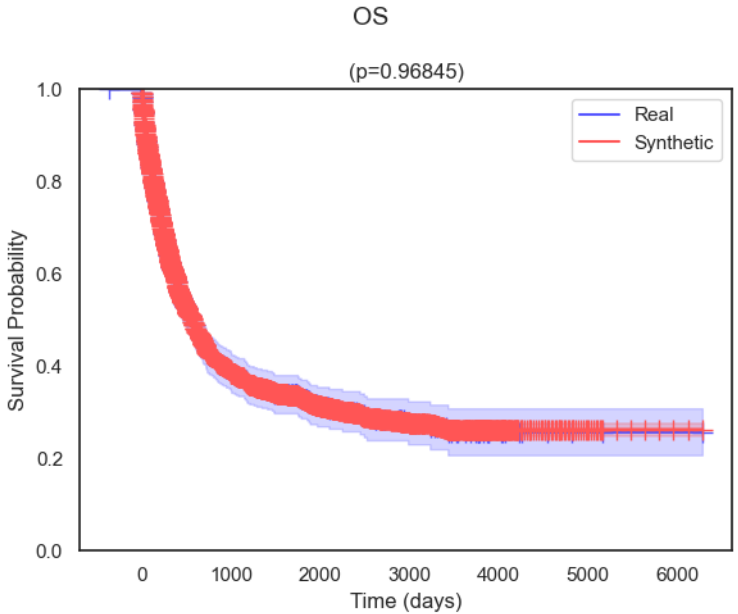
Best AUC-ROC = 0.5
Best TPR = 0

Privacy – Acute Myeloid Leukemia



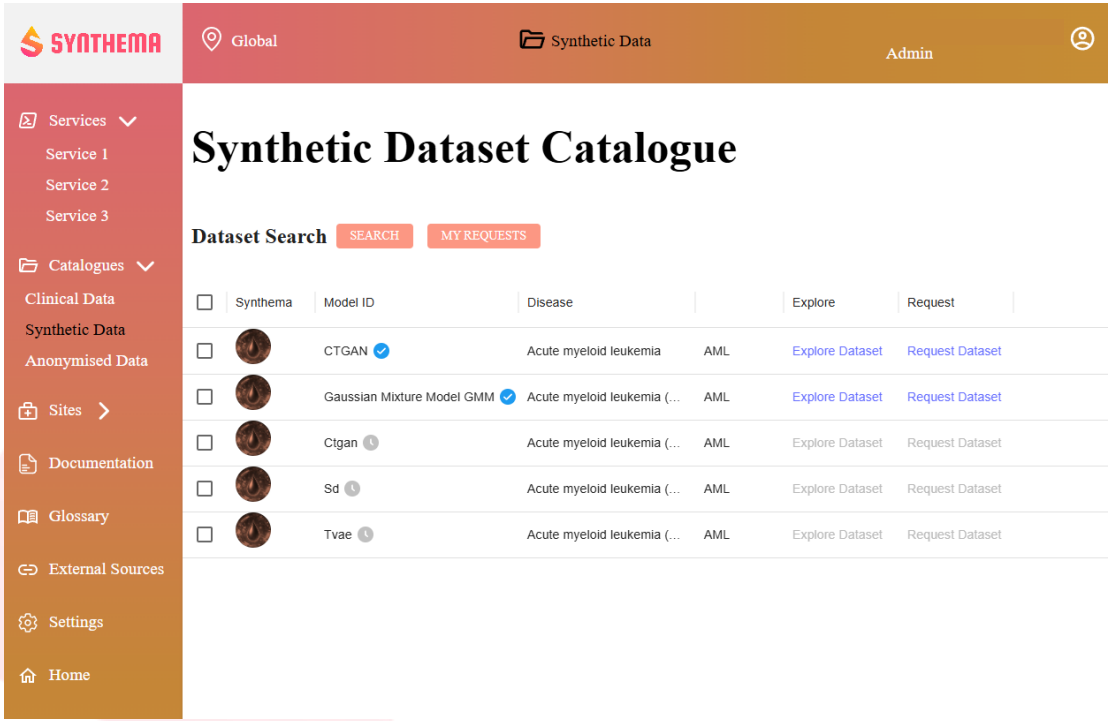
Synthetic data generation and anonymization

Real impact: SYNTHEMA as a use case



Synthetic data generation and anonymization

Real impact: SYNTHEMA as a use case



- Access to synthetic data generated using generative models trained on federated settings.
- Easy-to-access catalogue

Synthetic data generation and anonymization

Opportunities

- Stronger privacy preservation
- Collaborative learning – Improved access to data
- Improved AI robustness
- Better regulatory compliance

and challenges

- Heterogeneity
- Complex evaluation methodology
- Communicative and computational cost
- Generative model federation

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.