SYNTHEMA + ERN-EuroBloodNet

Joint Training Programme on Synthetic Data Generation in SCD and AML

Funded by the European Union

Opportunities of Synthetic Data Generation in Rare Hematological Diseases: Acute Myeloid Leukemia

Eleonora Iascone – IRCCS Humanitas

#1

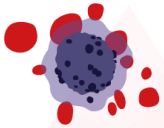
May 29th, 2026

Acute Myeloid Leukemia (AML)

AML synthetic data opportunities in SYNTHEMA federated context

AML Disease Context

- **What is AML?** Aggressive hematological malignancy (blood cancer) caused by the uncontrollable growth of immature myeloid cells (blasts) in the bone marrow, interfering with normal blood cells production.
  - **Rapid progression**: AML requires prompt diagnosis and treatment
  - **Biological heterogeneity**: AML includes multiple subtypes defined by different genetic changes
- **How is it diagnosed?** Through blood tests, bone marrow biopsy and genetic analysis
- **How is it treated?** Chemotherapy, targeted drugs and in some cases a stem cell transplant, with choice depending on age, fitness, comorbidities and specific genetic profile



1%	Of all new cancer diagnoses
68 yrs	Median age at diagnosis
~28% (50% in younger patients, <10% in >60 yrs)	Estimated 5-year overall survival (OS)

Clinical Challenge

What makes AML research so difficult?

Data Scarcity

Rare diseases as AML suffer from small cohorts limiting statistical power of single-centre analyses and AI model training and validation

Privacy Barriers

GDPR and ethical constraints prevent cross-istitutional sharing of real patients due to the intrinsic sensitive nature of health data. Data cannot be freely shared across institutions.

Fragmented and Heterogeneoeus Data

Data are often spread across different hospitals and come from heterogeneous sources

Research needs large and diverse datasets

To understand rare subtypes and develop better treatments, researchers need data from many patients across many centers

ERN | EuroBloodNet

SYNTHEMA

Synthetic Data Opportunities

Synthetic Data for AML

Privacy-preserving data sharing

SD enables **cross-institutional research** within legal and ethical boundaries and **aggregation of multiple datasets** from different sources (improved generalizability).

Rare subtype augmentation

Generative models can produce additional synthetic patients with specific molecular profiles, improving **dataset balance** and enabling more robust analyses of **underrepresented subtypes**.

Accelerate clinical research

Synthetic AML cohorts have been shown to reliably reproduce survival curves, relationships and molecular distributions from real trial data, enabling **exploratory analyses** and even **synthetic control arms**.

AML Data in SYNTHEMA

Data modalities

Tabular Data

Structured patient records collected at each participating centre, including:

- **Clinical data**: demographics, laboratory parameters, disease classification (WHO), risk stratification (ELN), treatment details, disease relapse and survival outcomes
- **Genomic data**: mutation status, cytogenetic abnormalities

Dataset: 2500+ patients with AML, 157 variables (55 required)

Imaging Data

- **Histopathological slides**: Hematoxilin & Eosin (H&E), Gomori staining
- **Cytological slides**: May-Grunwald Giemsa (MGG)

Dataset: 979 patients with 2103 H&E, 1155 Gomori and 1409 MGG

AML Data in SYNTHEMA

Data Challenges

		Technical challenges for SD generation
1	**Multi-modal complexity:** AML data spans three different modalities, each with different structure, scale and technical requirements	Each modality requires a dedicated generative model and validation pipeline
2	**Biological heterogeneity and small sample size per subtype:** AML comprises multiple genetically distinct subtypes, each with different prognosis and treatment response, making pattern recognition difficult.	Rare subtypes are underrepresented (model collapse)
3	**High dimensionality and genomic sparsity:** the dataset includes a large number of variables, including mutation status. The resulting matrix is highly sparse (most entries are 0).	Difficulties of the models in capturing rare genomic features

Synthetic Data Generation

Data quality & Preprocessing

Multi-centre data harmonisation

Patients data collected across 5 clinical centres (ICH, UNIPD, UMCU, CHA, GLSMED LH) were aligned into a single coherent dataset

REDCap
Research Electronic Data Capture

Variable Selection

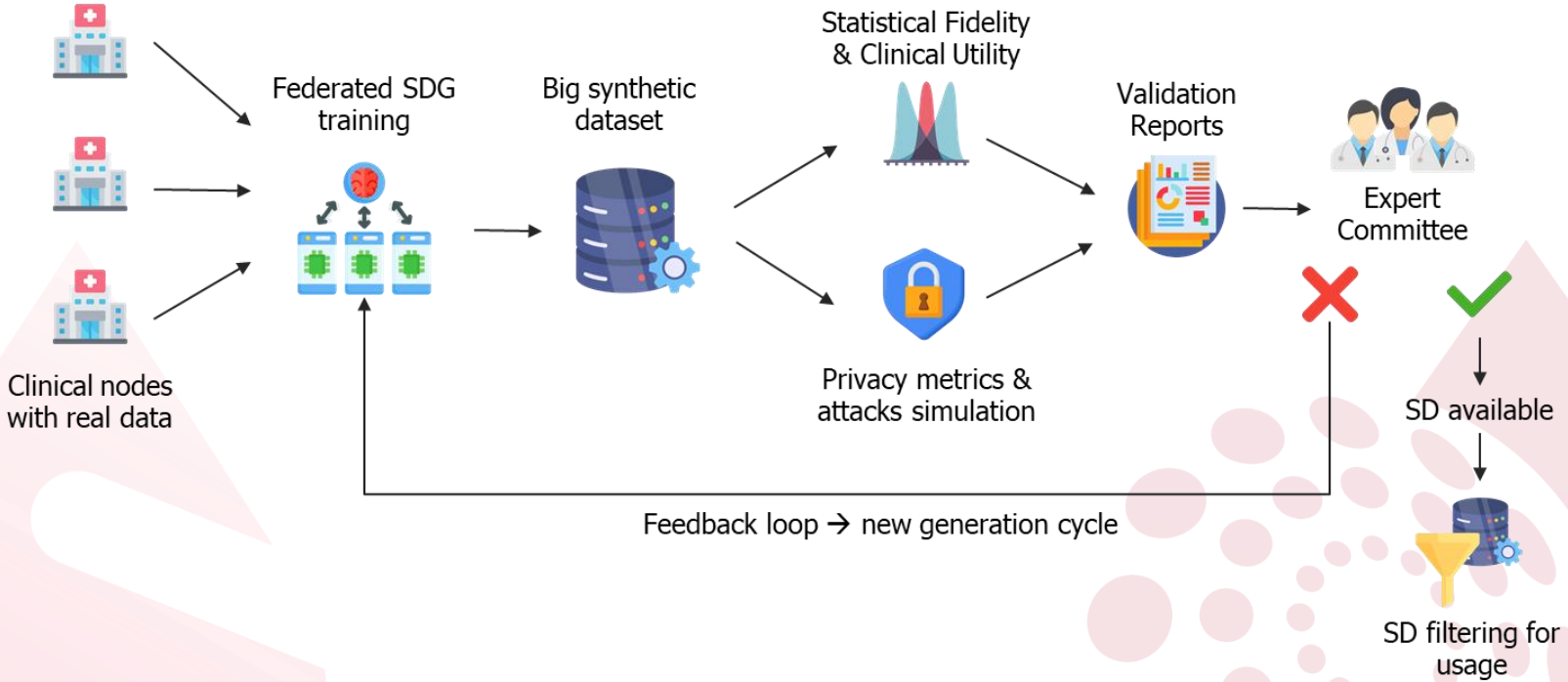
158 variables → 51 retained

Variable selection guided by clinical relevance and AI training requirements. Variables with excessive missingness or redundancy excluded

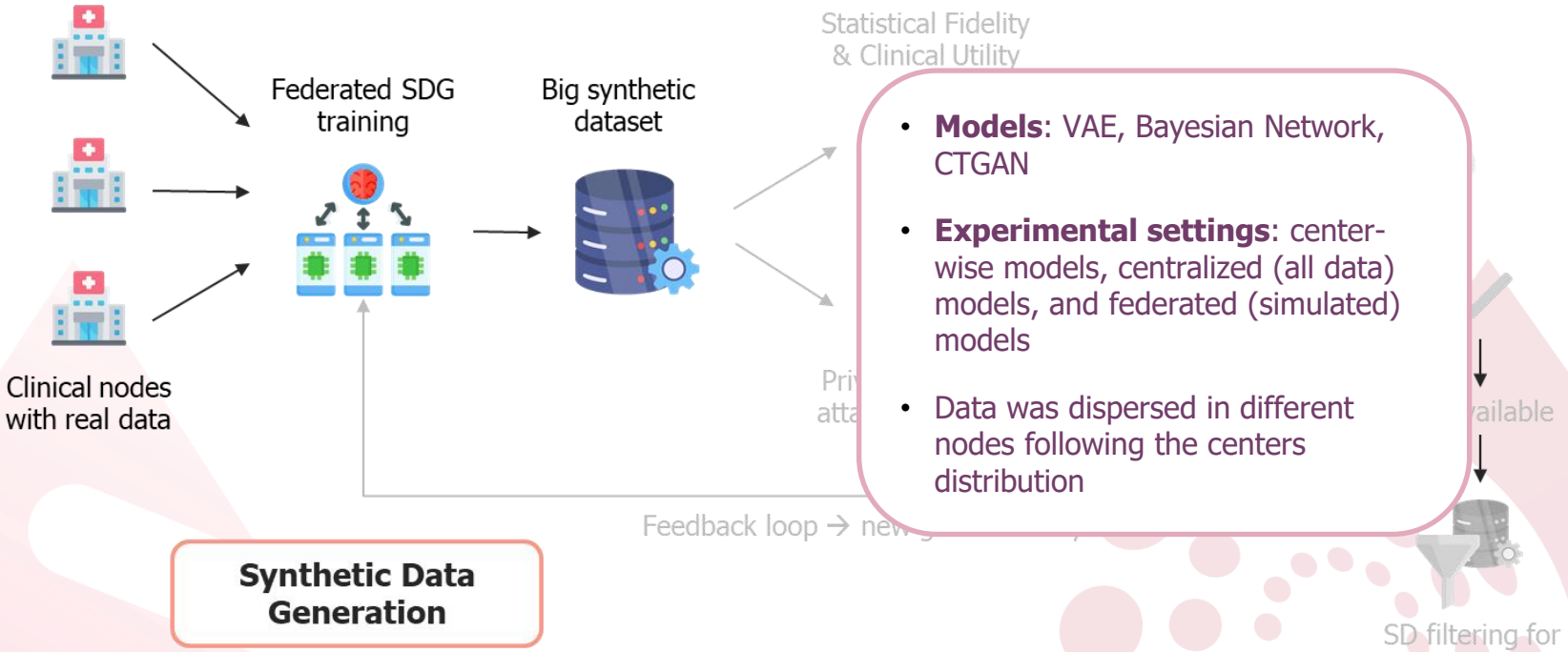
Missing Data Imputation

MICE (Multiple Imputation by Chained Equations) for continuous clinical variables;
KNN + rule-based imputation for genomic and categorical variables

Synthetic Data Generation & Validation Pipeline

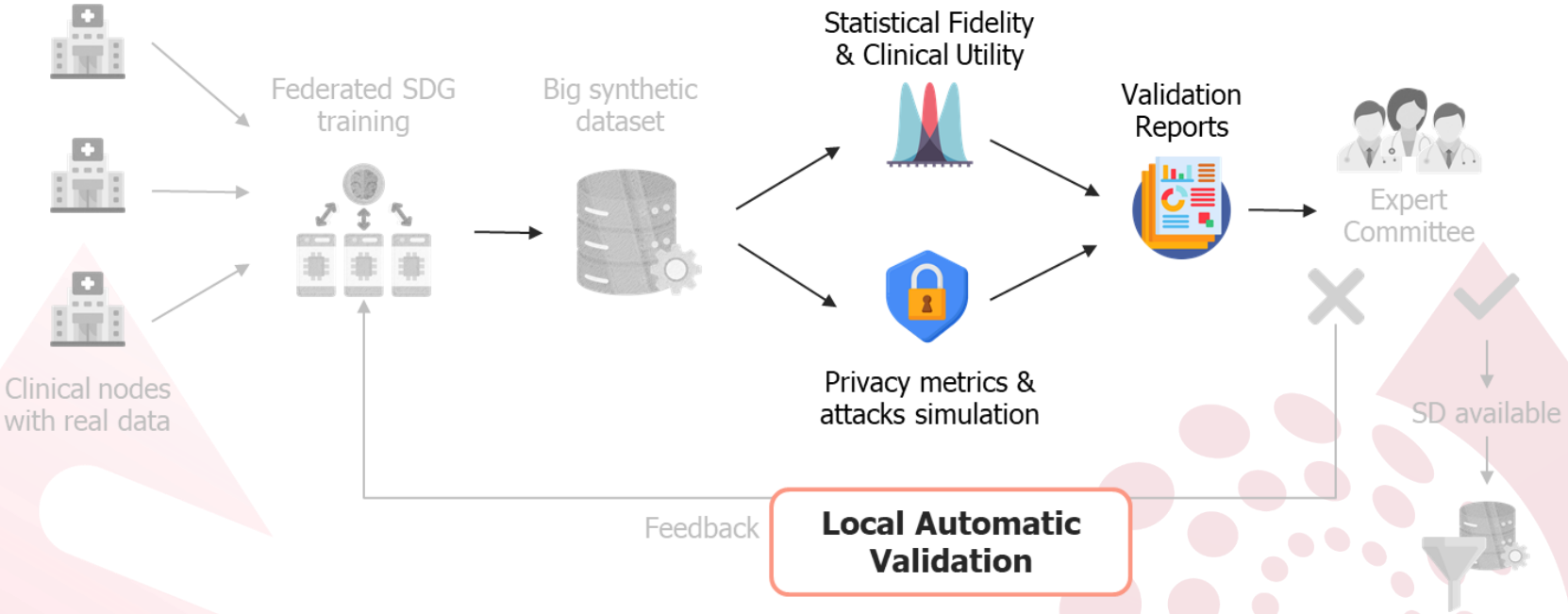


Synthetic Data Generation & Validation Pipeline



- **Models**: VAE, Bayesian Network, CTGAN
- **Experimental settings**: center-wise models, centralized (all data) models, and federated (simulated) models
- Data was dispersed in different nodes following the centers distribution

Synthetic Data Generation & Validation Pipeline



AML Validation Framework

Validation Dimensions

Statistical Fidelity

*Do synthetic data **replicate real distributions** and statistical properties?*

Clinical Utility

*Are synthetic data capable of reproducing the same **clinical outcomes** and conclusions?*

*Are synthetic data **useful** for downstream **clinical applications**?*

Privacy Preservation

*Do synthetic data **expose real patients**?*

*Are **sensitive attributes** sufficiently protected?*

Validation is not a quality check, but the **prerequisite for any downstream use** of synthetic data

ebn ERN | EuroBloodNet SYNTHEMA

AML Validation Framework

Statistical Fidelity

Statistical Fidelity

*Do synthetic data **replicate real distributions** and statistical properties?*

Clinical Fidelity

- Distribution analysis
- Correlation analysis
- Principal Component Analysis (PCA)
- Statistical tests: KS, Wilcoxon, Chi-Square, Fisher

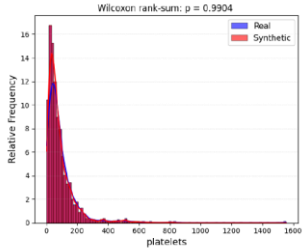
Genomic Fidelity

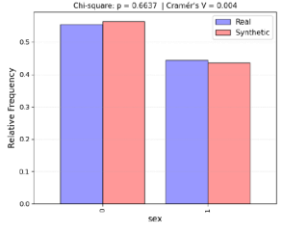
- Number of mutations per patient
- Mutation frequencies
- Pairwise association: co-mutations, odds ratios

AML Validation Framework

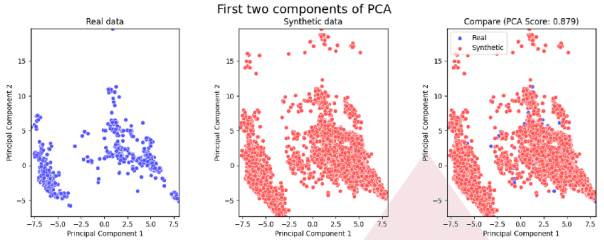
Statistical Fidelity – Clinical Data

Distribution Analysis

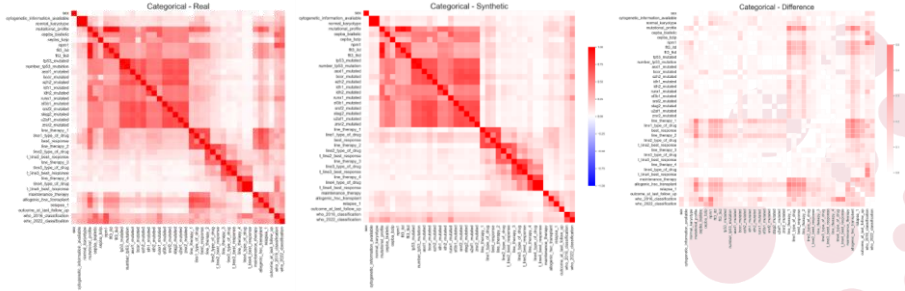




Principal Component Analysis (PCA)

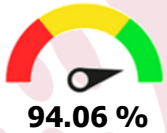


Correlation Analysis



Real Data | Synthetic Data | Absolute Difference

Clinical Synthetic Fidelity (CSF)



94.06 %

AML Validation Framework

Statistical Fidelity – Genomic Data

1.2 Genomic Analysis

Co-occurrences Score: 0.9966

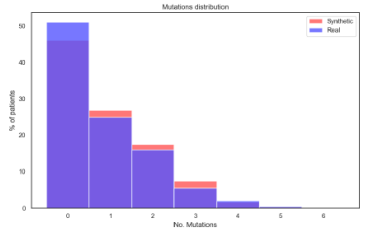
Odds Score: 0.5779

Frequency Score: 0.9736

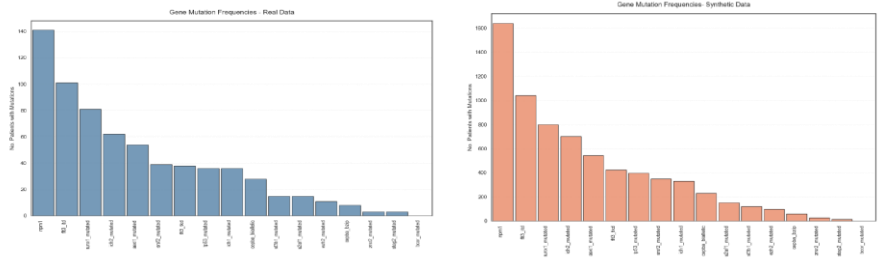
Mutation Score: 0.9376

GSF (Genomic Synthetic Fidelity): 0.8714

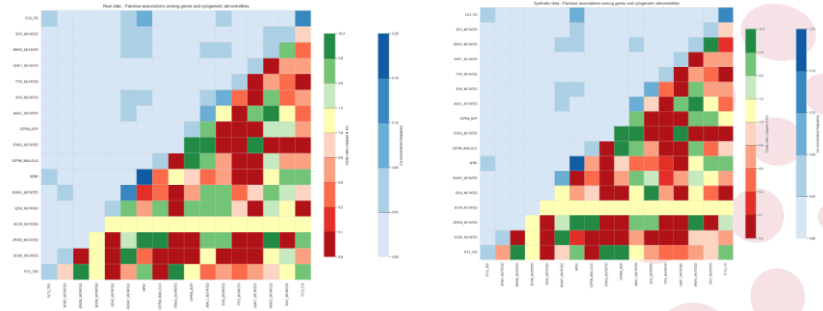
Mutation Distribution Plot



Mutation Frequencies



Pairwise Associations



Genomic Synthetic Fidelity (GSF)

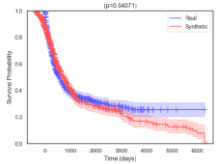
96.4 %

AML Validation Framework

Clinical Utility

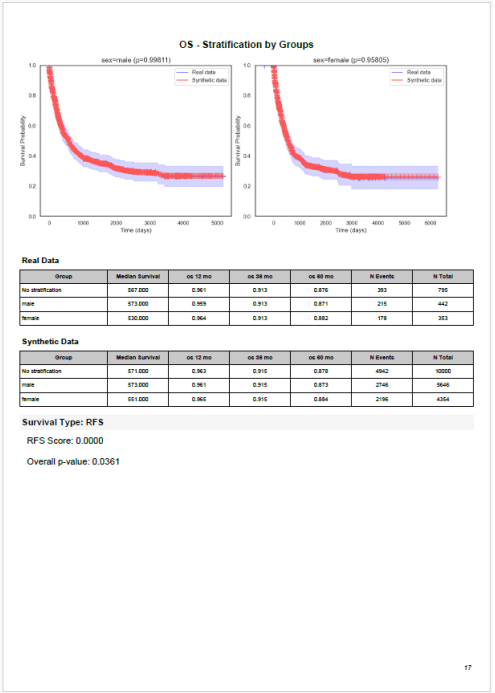
Clinical Utility

*Are synthetic data capable of reproducing the same **clinical outcomes** and conclusions?*
*Are synthetic data **useful** for downstream **clinical applications**?*

Replication of Clinical Studies	Survival Analysis	Train on Synthetic/ Test on Real (TSTR)	Human-in-the-loop
Re-run published analyses on synthetic data. Verify **consistency of clinical endpoints**, effect directions and statistical significance.	**Compare Kaplan-Meier curves** (real vs. synthetic) via log-rank test. 	Train **predictive models** on synthetic only and evaluate on real held-out data. Performance parity demonstrates downstream clinical utility.	**Clinicians** and **experts** are directly involved in the validation process, defining validation criteria upfront and review results downstream.

AML Validation Framework

Clinical Utility – Survival Analysis



Real Data

Group	Median Survival	os 12 mo	os 36 mo	os 60 mo	N Events	N Total
No stratification	567.000	0.961	0.913	0.876	393	795
male	573.000	0.959	0.913	0.871	215	442
female	530.000	0.964	0.913	0.882	178	353

Synthetic Data

Group	Median Survival	os 12 mo	os 36 mo	os 60 mo	N Events	N Total
No stratification	571.000	0.963	0.915	0.878	4942	10000
male	573.000	0.961	0.915	0.873	2746	5646
female	551.000	0.965	0.915	0.884	2196	4354

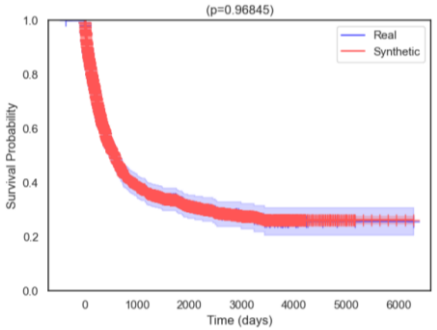
Survival Type: RFS

RFS Score: 0.0000

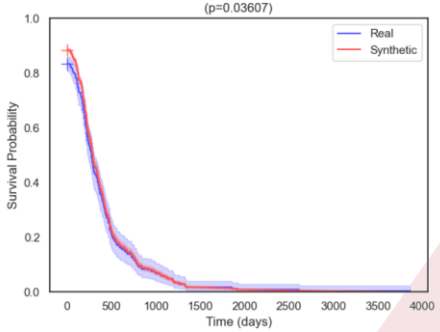
Overall p-value: 0.0361

17

Overall Survival (OS)



Relapse Free Survival (RFS)



Real Data

Group	Median Survival	os 12 mo	os 36 mo	os 60 mo	N Events	N Total
No stratification	567.000	0.961	0.913	0.876	393	795
male	573.000	0.959	0.913	0.871	215	442
female	530.000	0.964	0.913	0.882	178	353

Synthetic Data

Group	Median Survival	os 12 mo	os 36 mo	os 60 mo	N Events	N Total
No stratification	571.000	0.963	0.915	0.878	4942	10000
male	573.000	0.961	0.915	0.873	2746	5646
female	551.000	0.965	0.915	0.884	2196	4354

AML Validation Framework

Privacy

Privacy Preservation

*Do synthetic data **expose real patients**?*
*Are **sensitive attributes** sufficiently protected?*

Distance and Similarity metrics

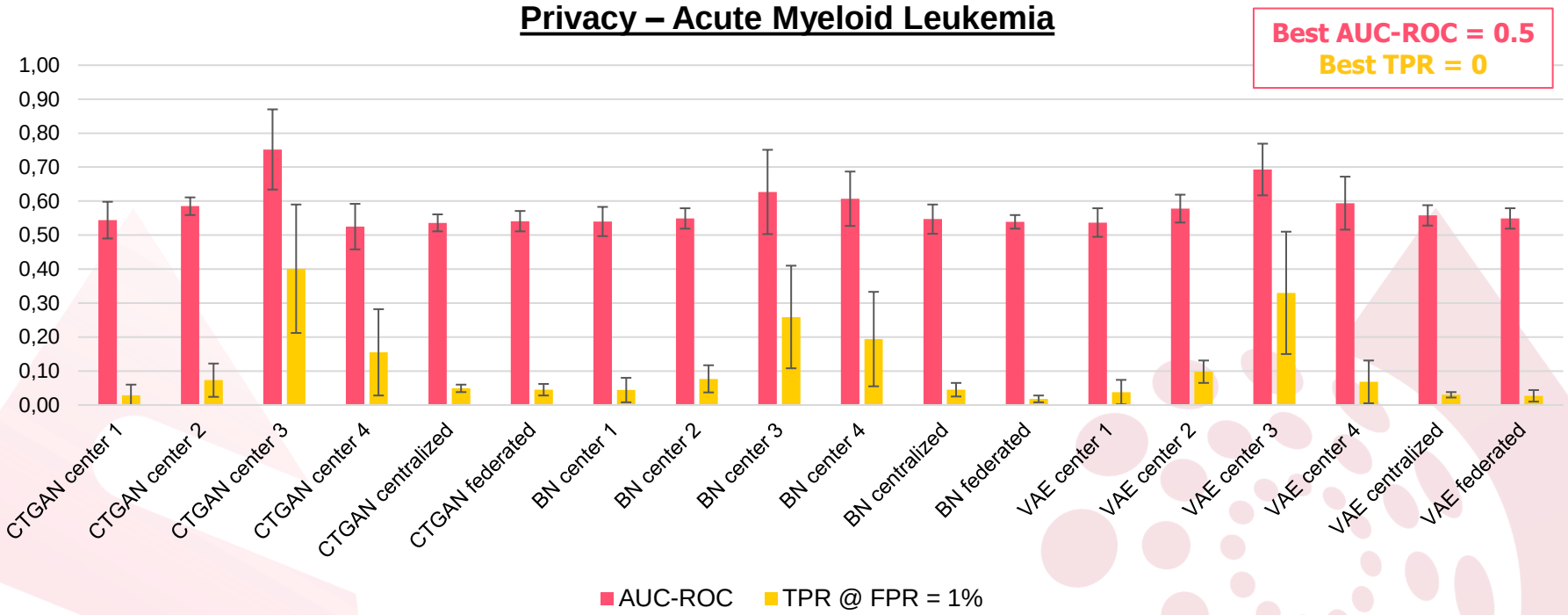
- Distance to Closest Record (DCR)
- Nearest Neighbor Distance Ratio (NNDR)
- Outliers Similarity (OS)
- Cosine Similarity (COS)
- Hausdorff Distance (HD)

Attack-based metrics

- Membership Inference Attack (MIA), AUC-ROC
- Attribute Disclosure (AD)

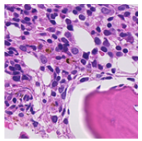
AML Validation Framework

Privacy



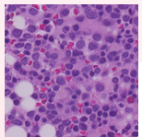
AML Validation Framework

Imaging Data – Histopathological Slides



REAL

Generative Model

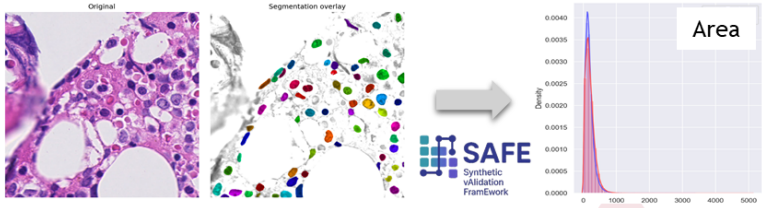


SYNTH

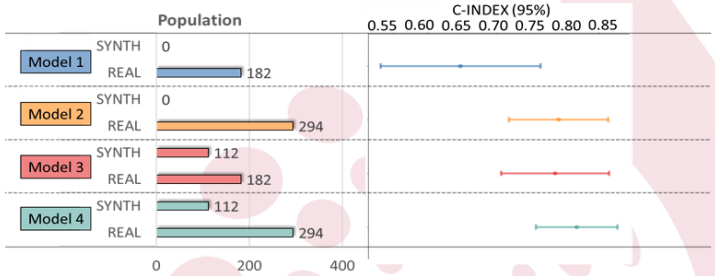
Validation Pipeline

1. **Statistical Metrics for images** (MS-SSIM, FID)
2. **Morphological features extraction and SAFE comparison**
3. **Prognostic Models (Cox Proportional Hazard)**
   Clinical data + Imaging-extracted features
4. **Blinded classification** by experts

2



3



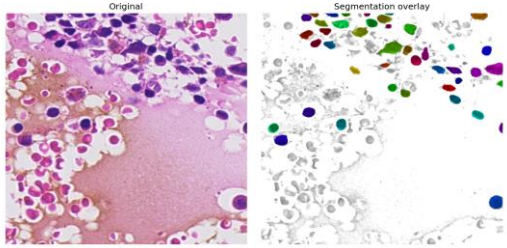
Hybrid approach combining **automated metrics** with **expert judgement** (Human-in-the-loop)

ERN | EuroBloodNet

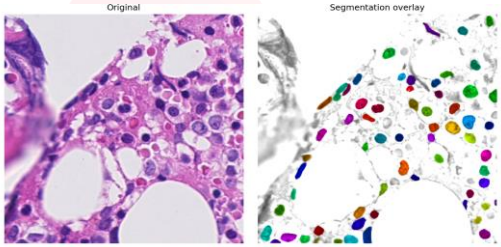
SYNTHEMA

AML Validation Framework

Imaging Data – Histopathological Slides



REAL

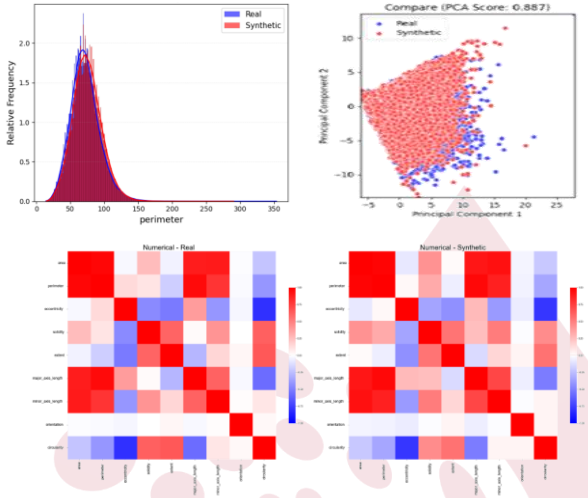


SYNTH

9 morphological features
extracted per nucleus:

- Perimeter
- Eccentricity
- Solidity
- Extent
- Major/minor axis length
- Orientation
- Circularity

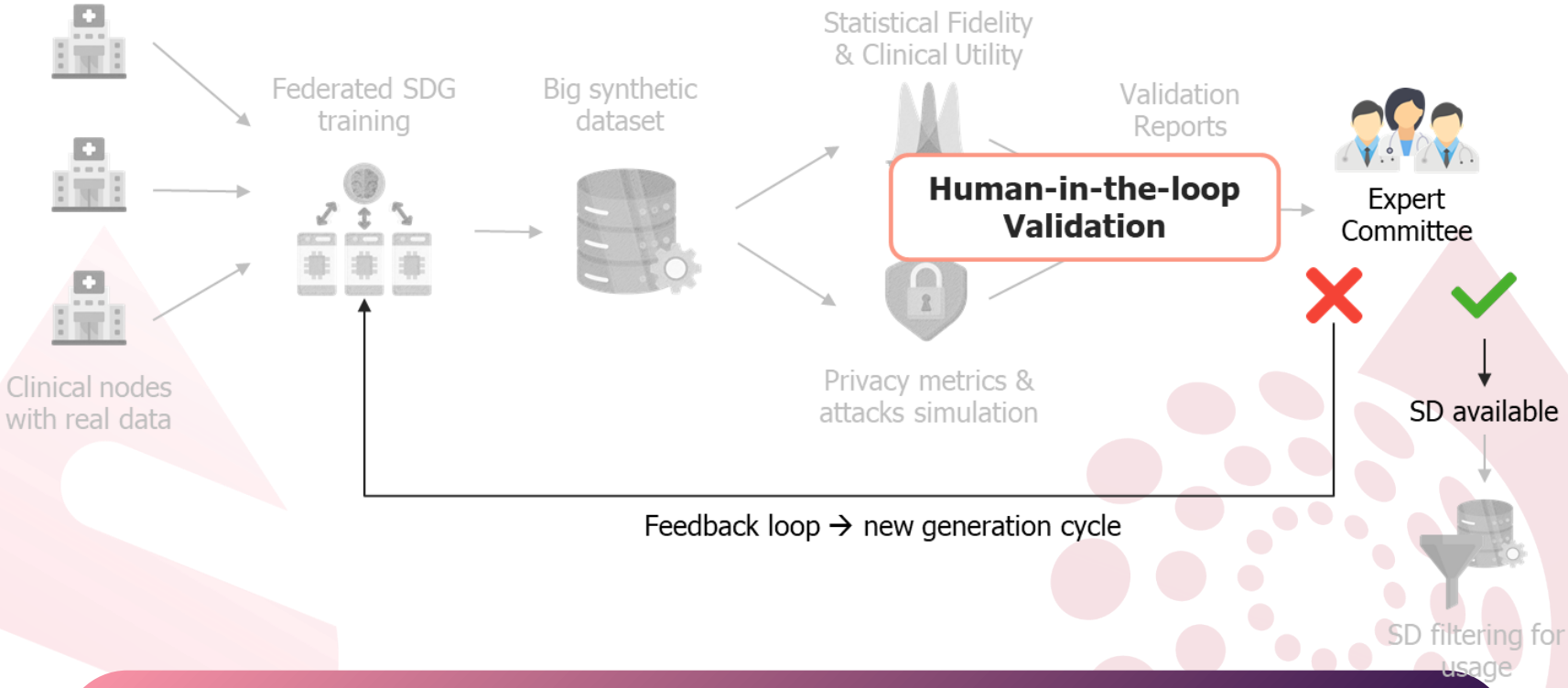
Nucleus-level



CSF: 0.898

Distance Score (numerical): 0.8724
PCA Score: 0.8868
Correlation Score: 0.9357

Synthetic Data Generation & Validation Pipeline



Federated Scenario

Why validation in SYNTHEMA is complex?

Statistical similarity is not enough

- Synthetic data (SD) can replicate distributions while failing to preserve clinical meaning
- Clinicians need to **trust** the data

Data cannot leave the nodes

- SD is generated via **Federated Learning** (FL): each center trains locally the generative algorithms and no real data is centralised
- Validation must happen under the same constraint: **real data must not leave the clinical site**

How can we quantify the clinical utility in synthetic data? Who is in charge of this decision?

How do you compare synthetic data to real data you cannot access centrally?

Federated Scenario

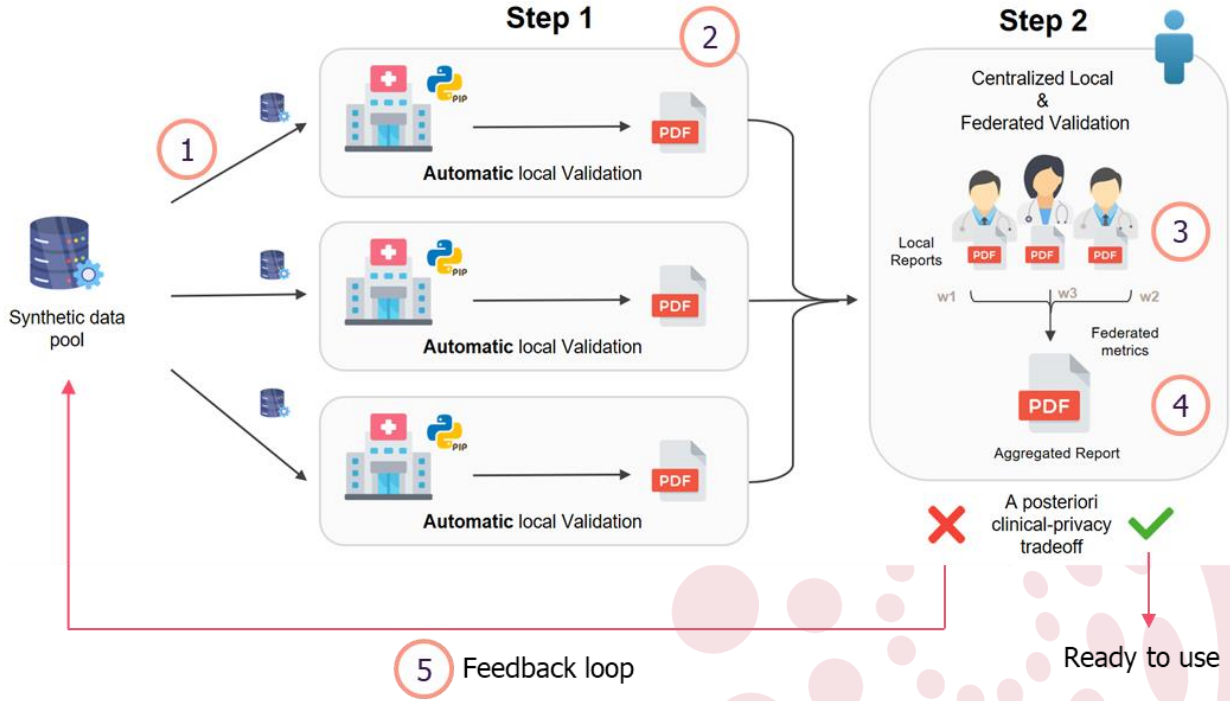
Validation in federated setting

Step 1 – Local Automatic Validation

1. SD pool distribution
2. Validation package (pip) runs locally generating per-centre HTML report

Step 2 – Centralized Federated Validation

3. Centralized validation of local reports (human-in-the-loop)
4. Validation of the aggregated (federated) report
5. SD generation refinement through feedback loop



Practical Applications

What we can do with validated SD for AML?

Clinical Questions

- Which **prognostic markers** matter most in patients aged 65–75 treated with intensive chemotherapy?
- Which **genetic risk profiles** drive outcomes in patients receiving non-intensive treatment?
- Which **biological markers** can predict **treatment response** in patients over 65?

How Synthetic Data helps

- **SD augments underrepresented age subgroups**, enabling statistically powered analyses that real data alone cannot support
- **Rare genetic subtypes can be augmented synthetically**, allowing survival analysis stratified by ELN risk group with adequate sample sizes
- SD enables the development and validation of predictive models using **information from different institutions**, without ever moving real patient records

ebn ERN | EuroBloodNet SYNTHEMA

Conclusions

Key Takeaways

What we have shown

- **AML is a paradigmatic use case**, since it combines rare disease challenges with multi-modal complexity
- **Synthetic data can be clinically valid**, if built with clinical rigour
- **Validation is essential** to build trust and **expert review** is an integral part of the process
- **Federated learning enables scalability**. The same pipeline can operate across institutions without moving patient data

Looking ahead

- **Regulatory recognition** of synthetic data as a valid evidence source in clinical research and clinical trials
- **Standardisation** of validation thresholds and frameworks across institutions and disease contexts
- **Building clinician trust** through transparency, interpretability and human-in-the-loop design

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.