1

Disclosures of _____ _____

Company name	Research support	Employee	Consultant	Stockholder	Speakers bureau	Advisory board	Other

2

Introduction and motivation

- Even within HARMONY (biggest world DB of HMs) we can have :
  - Presence of **missing data**
  - **Data heterogeneity** (data are collected from multiple sources)
  - **Classes under-representation** (age, gender, ther. resp., relapse, etc)
  - **Data scarcity** (Deep Learning requires millions of labelled data)
- Classical solutions: **imputation, regularization, data augmentation** etc.
- Data sharing is slowed by **GDPR strict rules** (EU state dependent)
- New approaches (especially in the Medical field) for mitigate these effects:
- **Federated Learning**, **Swarm Learning**, and **Synthetic Data Generation**

3

Synthetic Data (SD) and Original Data (OD)

- In HMs, including MM there is a **growing demand for large amount of high quality data** to build **Clinical Decision Support Systems** (to improve diagnosis, prognosis and personalized treatment), with the great challenge of **preserving patient privacy**.
- **SD can be a solution**, by capturing the complex statistical properties of the OD.
- The capability of SD to accelerate translational research is tested by a **SD Validation Framework** to evaluate their **quality and privacy preservability**
- So, what are SD and how they can be generated?

4

#data points >10×(# parameters in the model)

"In God we trust.
All others must bring data."

- *Dr. W. Edwards Deming*

5

EDPS EUROPEAN DATA PROTECTION SUPERVISOR

SD are artificial data **generated from OD** and a **model that is trained to reproduce the characteristics and structure of the OD**.

SD and OD should deliver very similar results when undergoing the same statistical analysis.

The **degree to which SD are accurate proxy for the OD** is a measure of the **utility** of the method and the model.

6

Positive foreseen impacts of SD on data protection

•**Enhancing privacy : data protection by design**, SD could provide, upon a **privacy assurance assessment**, an excellent method for not disclose personal data.

•**Improved fairness**: SD might contribute to mitigate bias by using **fair synthetic datasets to train AI models, in order to have a better representativeness of the world** (e.g. without gender-based or racial discrimination).

•**Data quality improvements** (missingnes, statistical harmonization, cohorts balancing, etc)

7

Negative foreseen impacts on data protection

•**Output control could be complex:** the best way to ensure output accuracy is to **compare SD with OD, hence access to OD is required**.

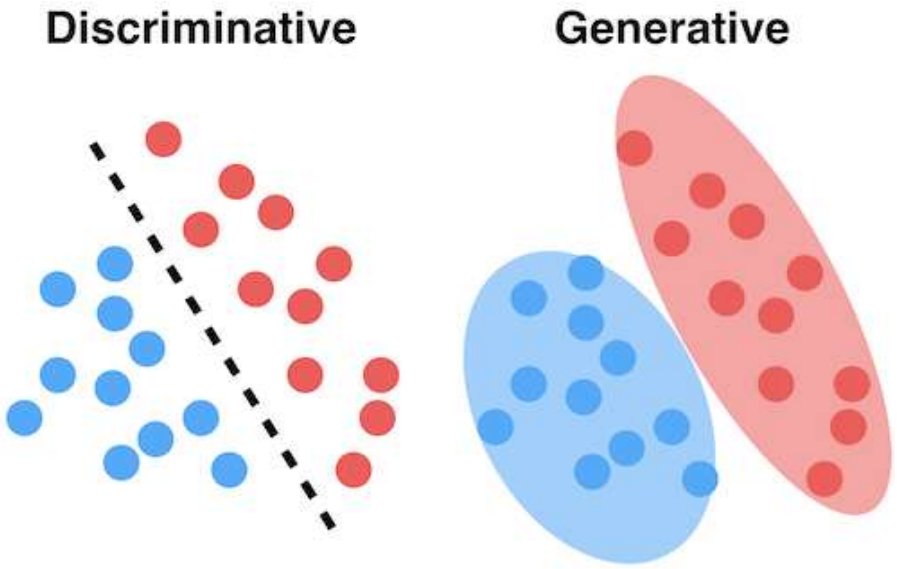
•**Difficulty to map outliers: SD may not cover some outliers that OD has** (data patients outliers can be very important).

•**Quality of the model depends on the data source**: the **quality of SD are highly correlated with the OD quality**.

•SD may reflect biases of OD, and **the creation of fair SD might result in inaccurate data.**

8

Generation of SD: Discriminative vs Generative AI



9

Generation of SD (SDG)

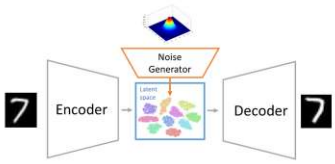
- The **SDG**, also called **synthesis**, can be performed using different techniques, such as deep learning algorithms (GANN, CGANN, VAE, Stable Diffusion, etc).
- SD can be classified with respect to their relation with OD
  - the first type employs **real datasets**,
  - the second employs knowledge gathered by the analysts instead (**classical and Bayesian simulation with a priori clinical knowledge**),
  - and the third type is a **combination of these two**.

10

State of the art in SDG

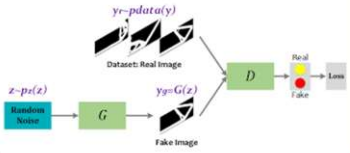
VAE

- Encoder + Latens Space + Decoder
- Sample the latent space to generate data
- Produce blurred images



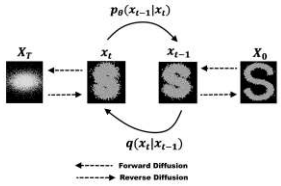
GAN

- Generator + Discriminator
- Generator: generate images
- Discriminator: distinguish between real and generated images



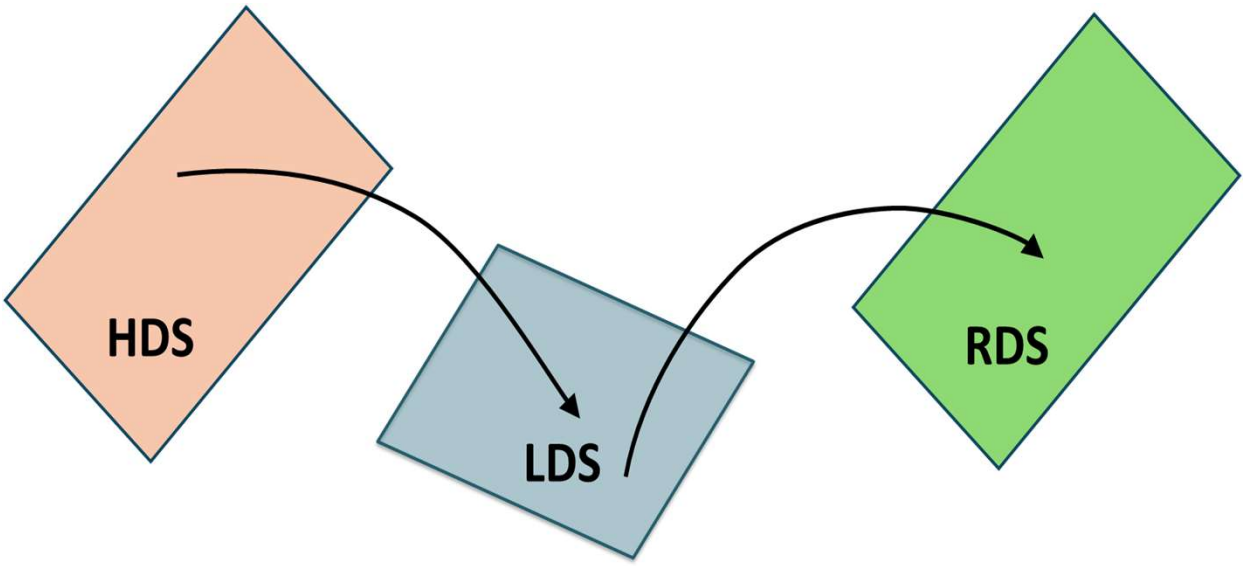
Diffusion

- Iterative refinement
- Diffusion + denoising
- map noisy to clean data
- Generate running denoising



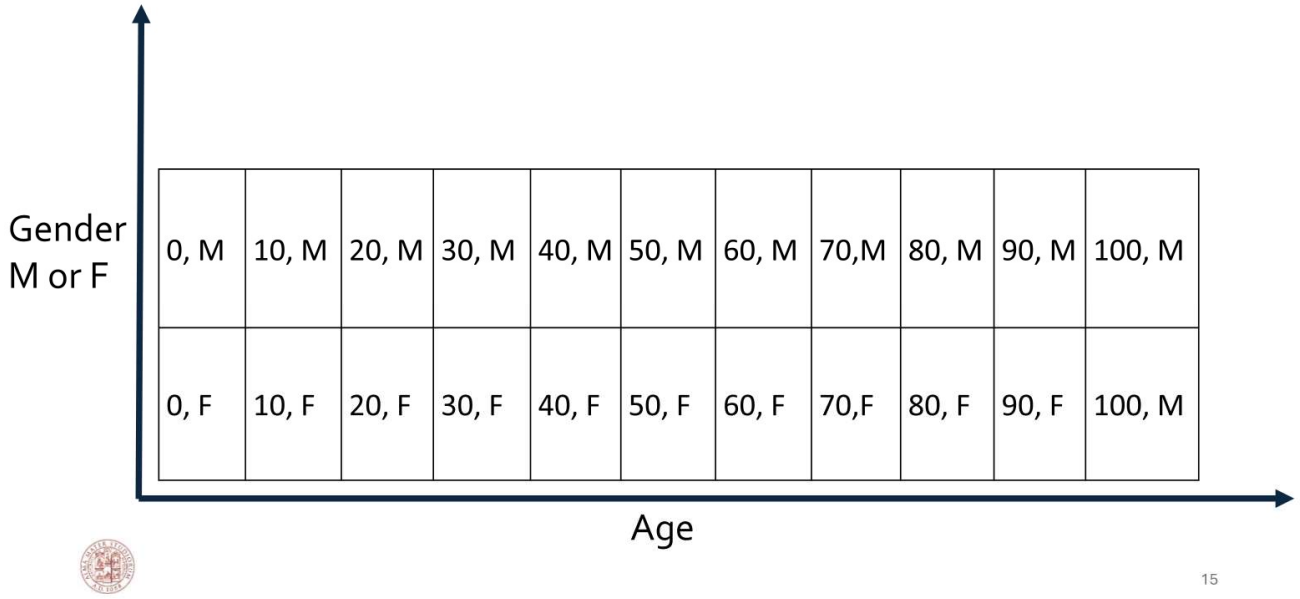
11

Generation of SD



12

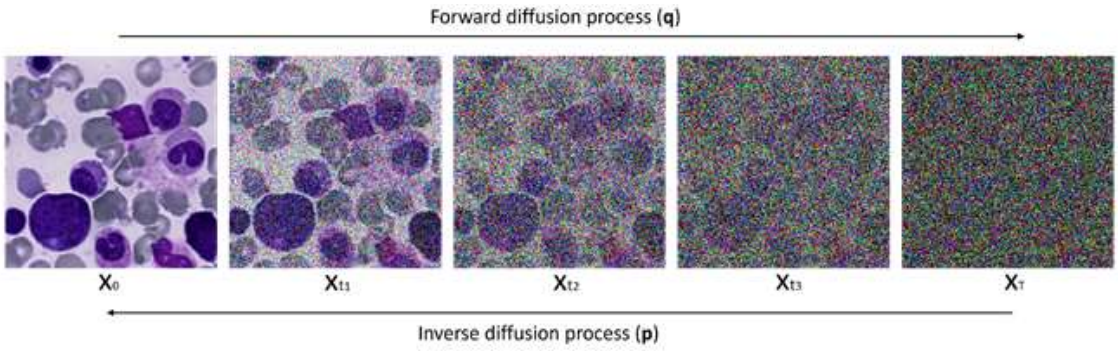
Latent space factorization for cohort balancing



15

13

Generation of SD with diffusion model



The forward (or inference) diffusion process converts any complex data distribution into a simple one, and then learn a finite-time reversal of this diffusion process which defines the generative model distribution.

14

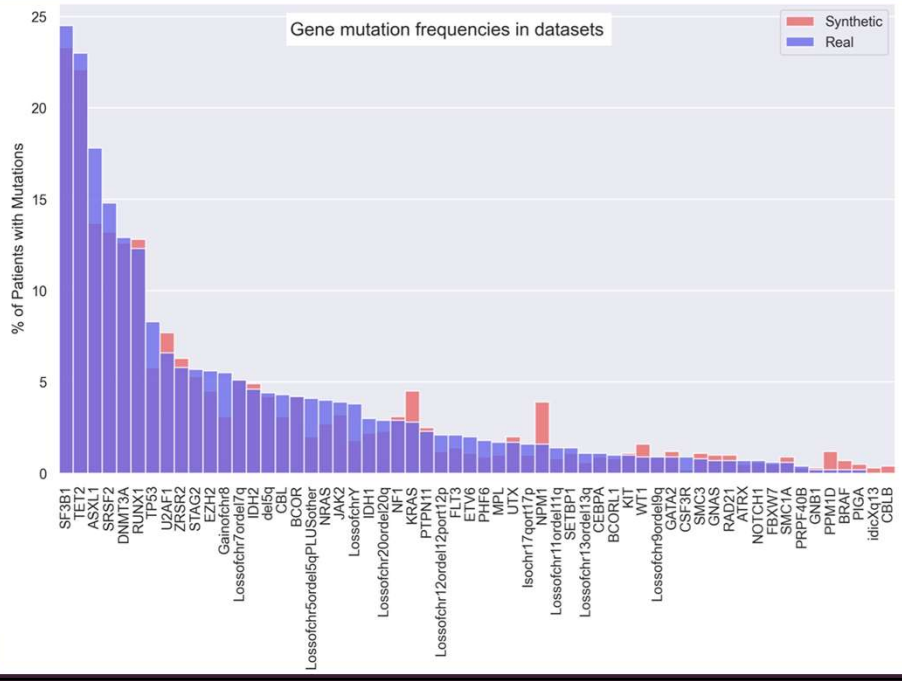
Data sources

Clinical Data	Omics Data	Imaging Data
Structured clinical data	Genomic variants (GWAS)	Radiomics (3D-T1, FLAIR, DWI, MRIs, PET and CT scan)
Laboratory	Transcriptomics (TWAS)	
Comorbilities	Cytogenetics	WSI histopathological images
Treatment information	Metabolomics	
Survival times		

15

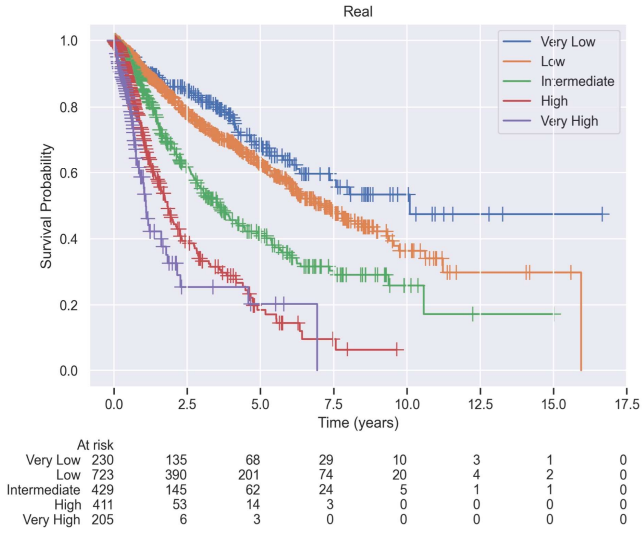
Frequency of mutated genes in real and SD

Bersanelli M, …, Castellani G, Della Porta MG. Classification and Personalized Prognostic Assessment on the Basis of Clinical and Genomic Features in Myelodysplastic Syndromes. J Clin Oncol. 2021

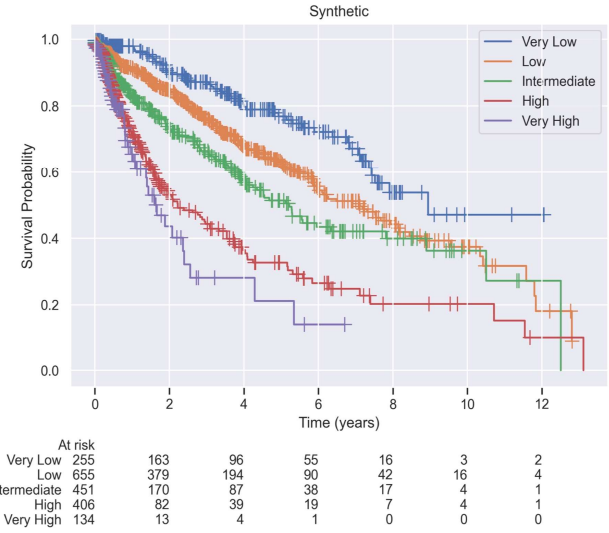


16

Validation protocol: OS comparison in AML and MDS



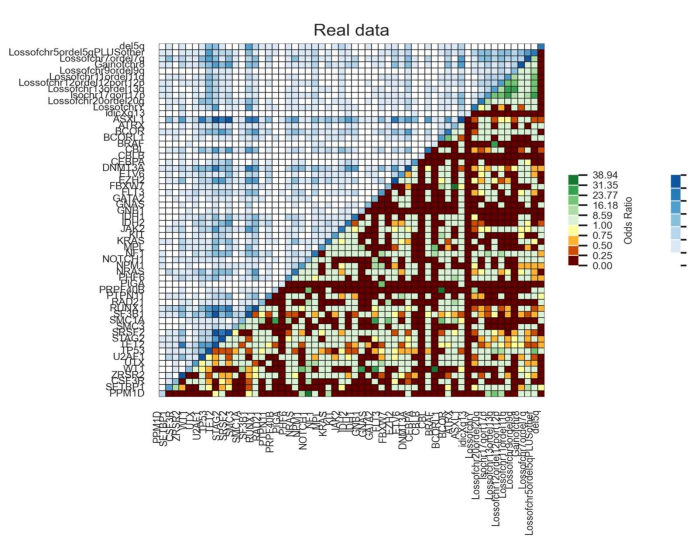
At risk							
Very Low	230	135	68	29	10	3	1	0
Low	723	390	201	74	20	4	2	0
Intermediate	429	145	62	24	5	1	1	0
High	411	53	14	3	0	0	0	0
Very High	205	6	3	0	0	0	0	0

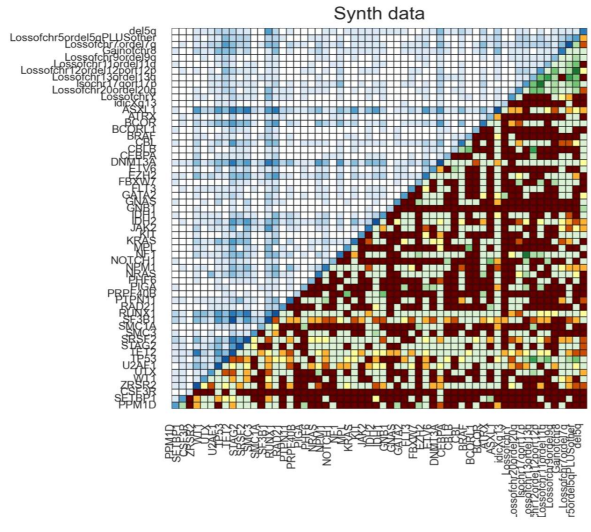


At risk							
Very Low	255	163	96	55	16	3	2
Low	655	379	194	90	42	16	4
Intermediate	451	170	87	38	17	4	1
High	406	82	39	19	7	4	1
Very High	134	13	4	1	0	0	0

17

Validation protocol: Correlation and Co-occurrence matrix

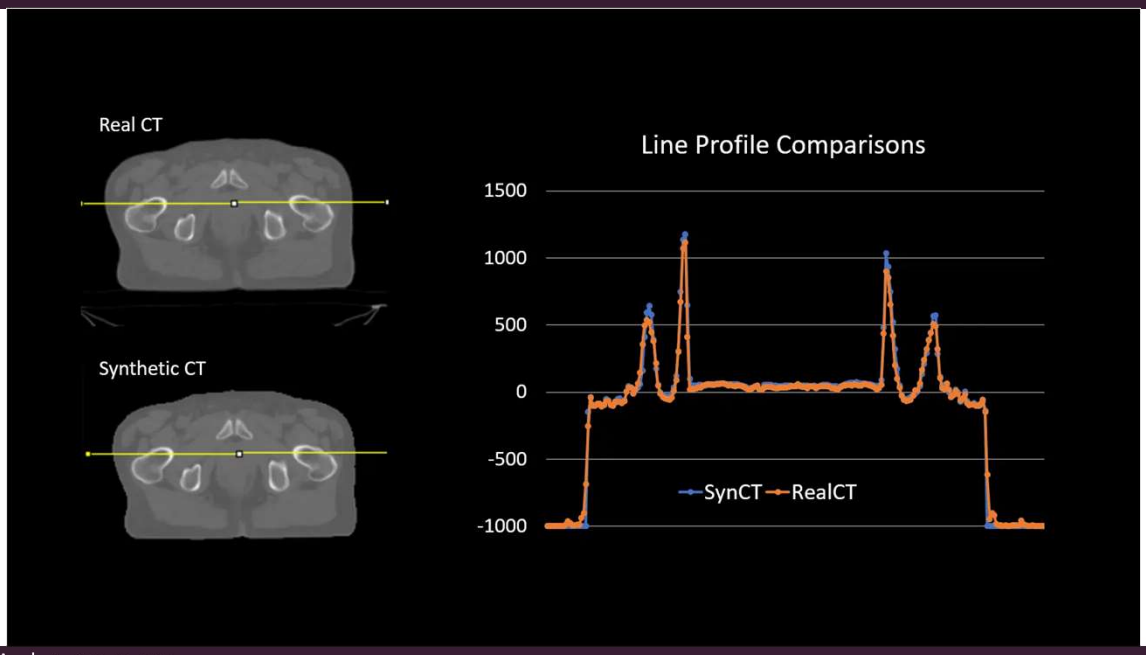




D'Amico S, ..., Castellani G, Della Porta MG. Synthetic Data Generation by Artificial Intelligence to Accelerate Research and Precision Medicine in Hematology. JCO Clin Cancer Inform. 2023

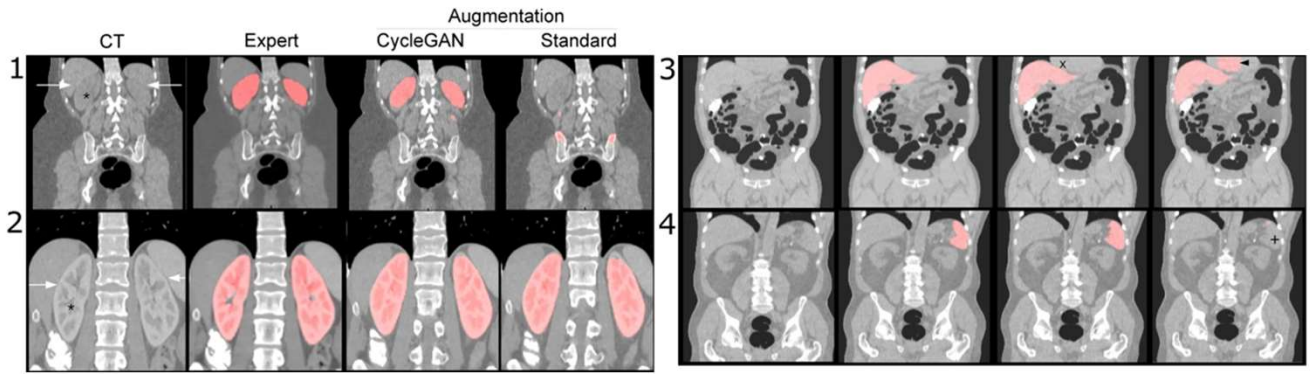
18

Validation protocol: CT segmentation



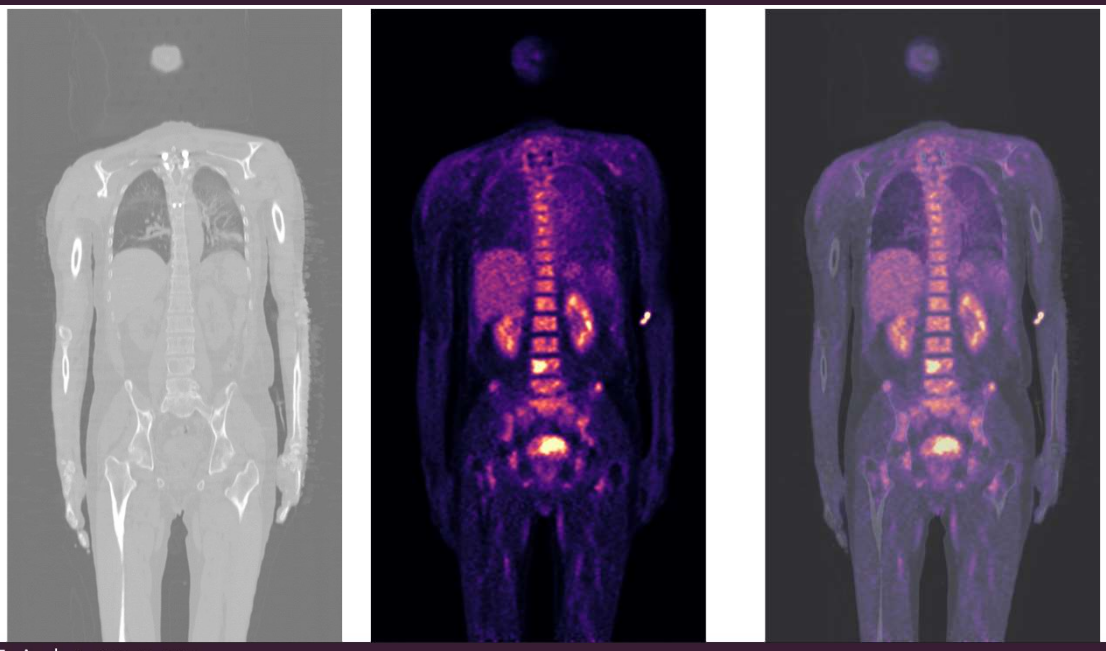
19

Data augmentation using (CycleGAN) to improve CT segmentation



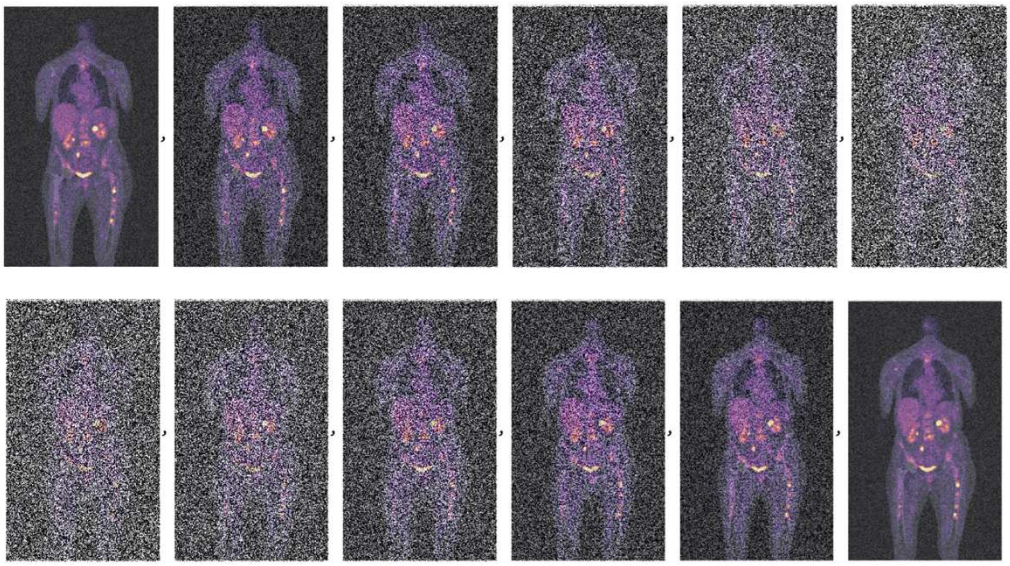
20

PET and CT in MM patients



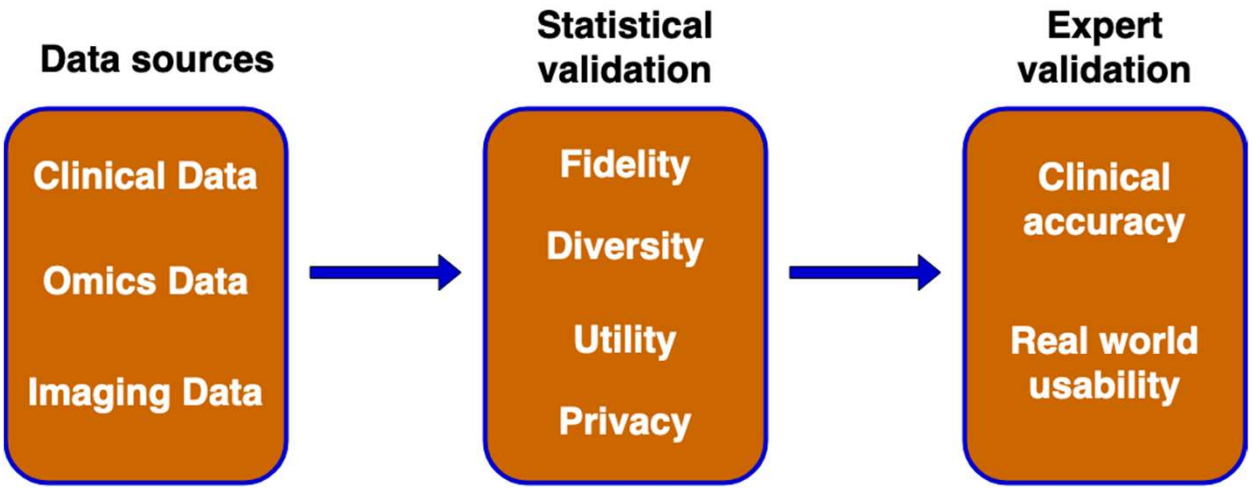
21

SD generation pf PET/CT in the case of MM



22

Validation framework



23

Validation framework

Fidelity

- **Marginal and joint distributions** (Wasserstein distance, Kolmogorov-Smirnov test, Kullback–Leibler divergence, Hellinger Distance).
- **Correlation structures** (Pairwise correlations, Kendall's tau, Levine's test)

Omics

- Mutation co-occurrence, Mutual exclusivity, Bradley-Terry clonality test

Imaging

- **Image quality** (Kernel Inception Distance)
- **Texture features** (Haralick features)

Diversity

Variability
Total variation distance
Entropy

Coverage
Support coverage
Range coverage

Distribution of extracted features

Utility

Clinical

- **Survival analysis** (Cox, Bayesian models)
- **Clustering** (Clustering,metric, silhouettes)
- **Cross-classification**

Omics

- **Clustering and stratification** (Dirichlet)
- **Feature extraction**
- **Cross-classification**
- **Causal relationships** (Bayesian networks)

Imaging

- **Image segmentation** (Morphological and radiomic features, spatial patterns of ROI)
- **Cross-classification**

24

SDG for EHR and RCT

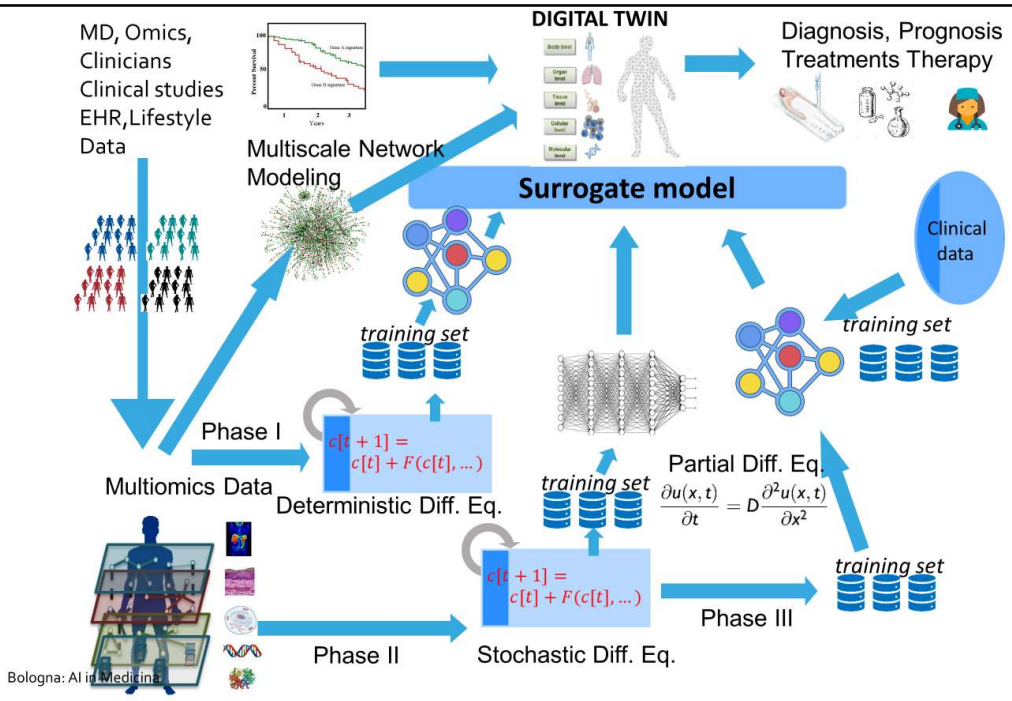
- SD can be used as a **Control Arm in RCT**
- SD together with **Large Language Model and Natural Language Processing** can be used for generating EHR
- A recent evolution of SDG is the capability to generate **images from text and vice-versa**

25

Conclusion

- SD is a new and promising concept in AI
- They can be very useful for ML training, especially for rare diseases
- Methods, based on latent space sampling are very promising
- SD as RWD can accelerate the process required in new drugs approval that is actually based on large and expensive RCTs
- The next future will see a merging between SD and **Physics Informed Neural Networks** for a new generation of predictive tools

26



27

Related project **ongoing**

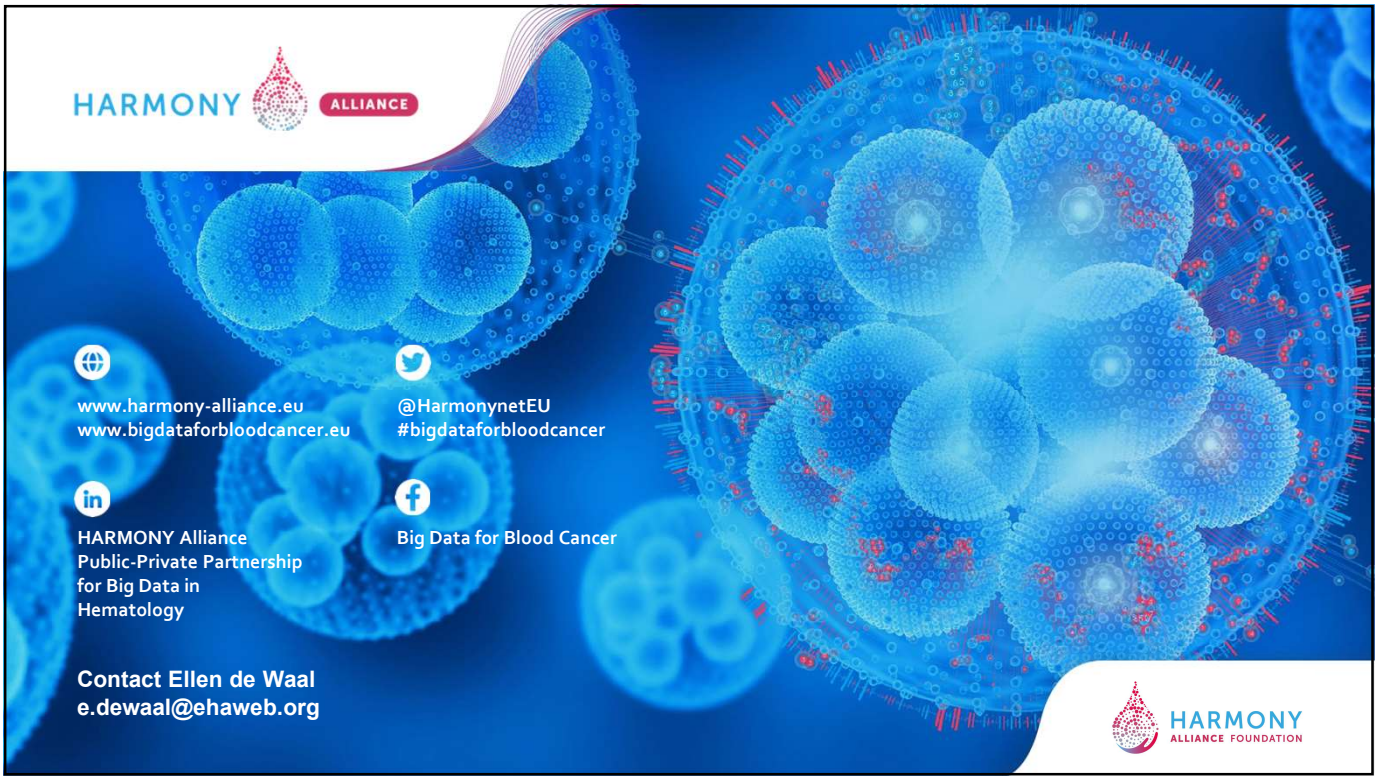
- **2024** Synthetic data generation framework for integrated validation of use cases and AI healthcare applications
- **2023 PRIN-**Personalized Medicine In Myeloid Neoplasms: Explainable Artificial Intelligence Solutions For Next-Generation Classification And Management Of The Patients
- **2023 MAECI** Science and Technology Cooperation Italy-South Korea Grant Years 2023–2025 by the Italian Ministry of Foreign Affairs and International Cooperation.
- **2023 PNRR** on Antimicrobial Resistance (1M€)
- **2022 EU SYNTHEMA** Synthetic generation of haematological data over federated computing frameworks 500 k€ (whole project 6 M€)
- **2022- AIRC Individual Grant** - IG 2021 Artificial intelligence for genomics and personalized medicine in myelodysplastic syndromes (MDS) 700 k€
- **2021 H2020 GENOMED4ALL** Genomics and Personalized Medicine for all though Artificial Intelligence in Haematological Diseases . Federated Learning. 800 k€ (the whole project is 10M€)
- **ISW: (H2020)**In Silico World Lowering the barriers to a universal adoption of In Silico Trials 200 k€ (the whole project is 6M€)

28

Related completed projects

- **2020 EU project HARMONY-PLUS: HEALTHCARE ALLIANCE FOR RESOURCEFUL MEDICINES OFFENSIVE AGAINST NEOPLASMS IN HEMATOLOGY – PLUS (HARMONY PLUS).** 36 months + 6 months extension. Data Analytics and Big Biomedical data integration for hematological malignancies, including the set-up of a pan-European computing facility. Role WP Co-Leader. EU contribution to UNIBO 339.000 € (the whole project was 12 M€)
- **2019 EU Project Versatile Emerging infectious disease Observatory (VEO)** 60 months Data analytics and modeling. Data Analytics and modeling. EU contribution to UNIBO 341378 € (the whole project was 15M€)
- **2017 EU project HARMONY: Alliance for Resourceful Medicines Offensive against Neoplasms in HematologY.** 60 months + 1,5 year extension. Data Analytics and Big Biomedical data integration for hematological malignancies, including the set-up of a pan European computing facility. Role WP Leader. EU contribution to UNIBO 800.000 € (the whole project was 40 M€)

29



30