PSI 2025 CONFERENCE

8 - 11 JUNE 2025

WEMBLEY STADIUM
LONDON

AI Generated Synthetic Control Arms to optimize Clinical Trials

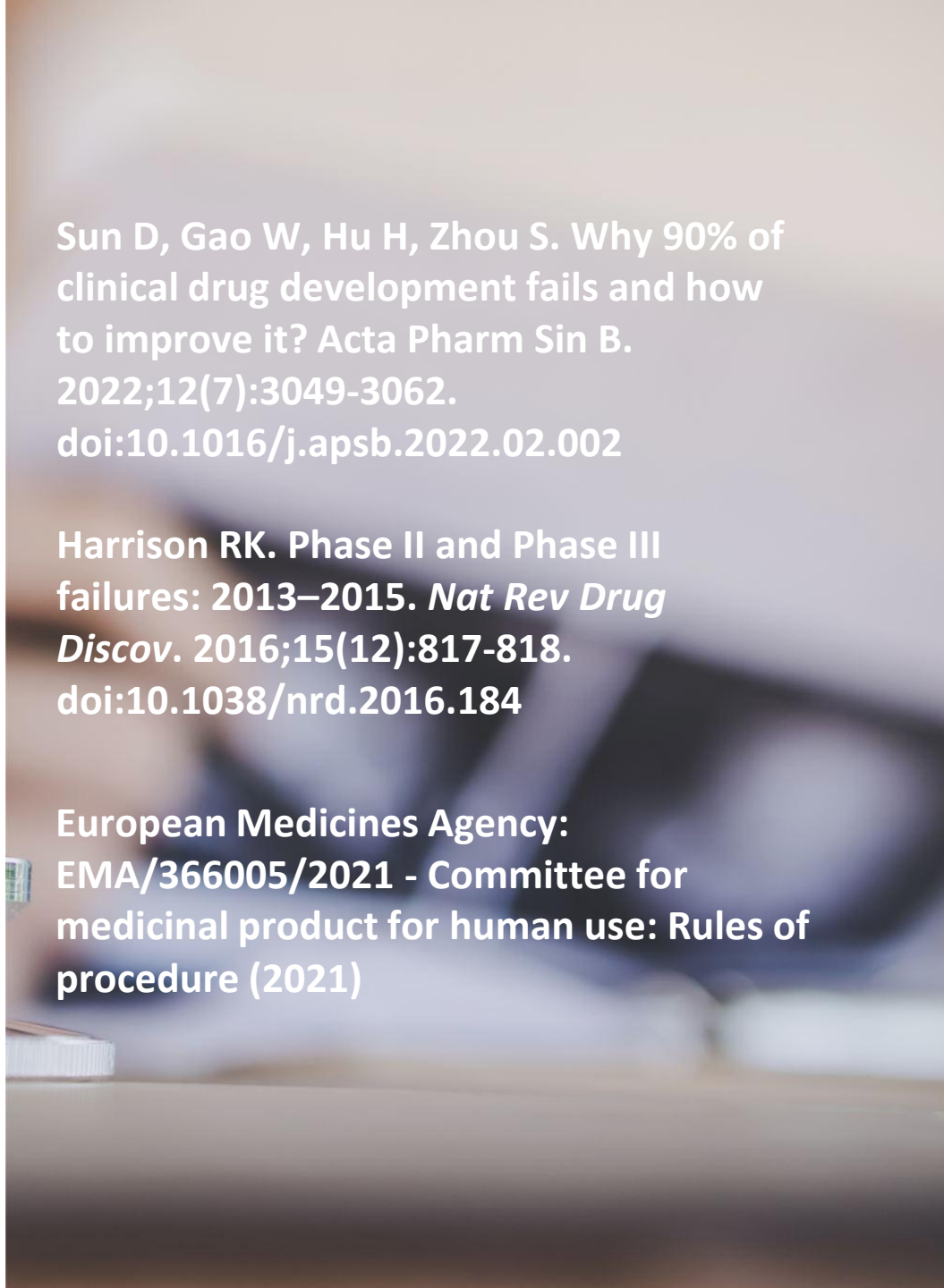
Paola Berchialla, paola.berchialla@unito.it University of Torino
Danila Azzolina, danila.azzolina@unina.it University of Naples

UNIVERSITÀ DI TORINO

CENTRE FOR BIOSTATISTICS, EPIDEMIOLOGY AND PUBLIC HEALTH

Challenges in conducting clinical trials

- The reluctance of patients to participate in RCTs where they might not receive the experimental treatment
- Complexities of collecting sensitive patient data, has led to escalating costs and extended trial durations
    - The assessment of an application for a new medicine alone, which is but a small part of the process, takes up to 210 active days
    - expenses of up to $2.7 billion are needed for a new medical product release
- Failures in confirmatory Phase III trials have the greatest impact on research investment
    - the probability of success in Phase III trials is about 60%, and among failures, between 40% and 50% are due to lack of statistical significance in the primary endpoint
- Patient data is sensitive
    - General Data Protection Regulation, introduced in 2018
    - collected patient information cannot just be reused in or combined with other trials, resulting in a less than optimal use of its potential
    - sharing original data with external researchers is often prohibited



Sun D, Gao W, Hu H, Zhou S. Why 90% of clinical drug development fails and how to improve it? Acta Pharm Sin B. 2022;12(7):3049-3062. doi:10.1016/j.apsb.2022.02.002

Harrison RK. Phase II and Phase III failures: 2013–2015. *Nat Rev Drug Discov*. 2016;15(12):817-818. doi:10.1038/nrd.2016.184

European Medicines Agency: EMA/366005/2021 - Committee for medicinal product for human use: Rules of procedure (2021)

Synthetic Data to Augment Control Arm in RCT


Journal of Clinical Oncology

FREE ACCESS | COMMENTS AND CONTROVERSIES | February 23, 2024

Use of External Control Cohorts in Pediatric Brain Tumor Clinical Trials

Authors: Ashley S. Margol, MD, MS; Annette M. Molinaro, PhD; Arzu Onar-Thomas, PhD; Adam Resnick, PhD; Derek Hanson, MD; Mark Kieran, MD, PhD; Pallavi Mishra-Kalyani, PhD; Donna Rivera, PharmD; Amy Barone, MD; David Arons, JD; Clair Meehan, MEM; and Michael Prados, MD

AUTHORS INFO & AFFILIATIONS

Publication: Journal of Clinical Oncology • Volume 42, Number 12 • https://doi.org/10.1200/JCO.23.01084

2,448 / 2 | 28 | PDF

Introduction

CNS tumors are the most common source of cancer-related death in children and are heterogeneous with varied molecular pathogeneses, resulting in a small population of over 70 subtypes.[1-5] They are also heterogeneous with respect to outcomes, with some having very poor survival rates because of lack of effective therapies.[6-11] Prospective randomized clinical trials are

Minimizing Placebo Use: Fewer patients are exposed to placebo or less effective treatments.

Safety: Reduces exposure to potentially harmful interventions.

Patient-Centric Trials: Focus on maximizing therapeutic benefit while reducing unnecessary interventions.

Rare Diseases: Synthetic data supports studies where recruiting large control groups is difficult or unethical.

Faster Access to Innovative Therapies: Accelerates timelines for potentially life-saving treatments.

Promoting Equity in Clinical Trials: Ensures fair access to experimental therapies for all patients by reducing the need for large control groups.

UNIVERSITÀ DI TORINO

BEPH

Synthetic population

- **Synthetic population** is artificial population data that fits the distribution of people and their relevant characteristics living in a specified area as according to the demographics from census data

Project in Europe

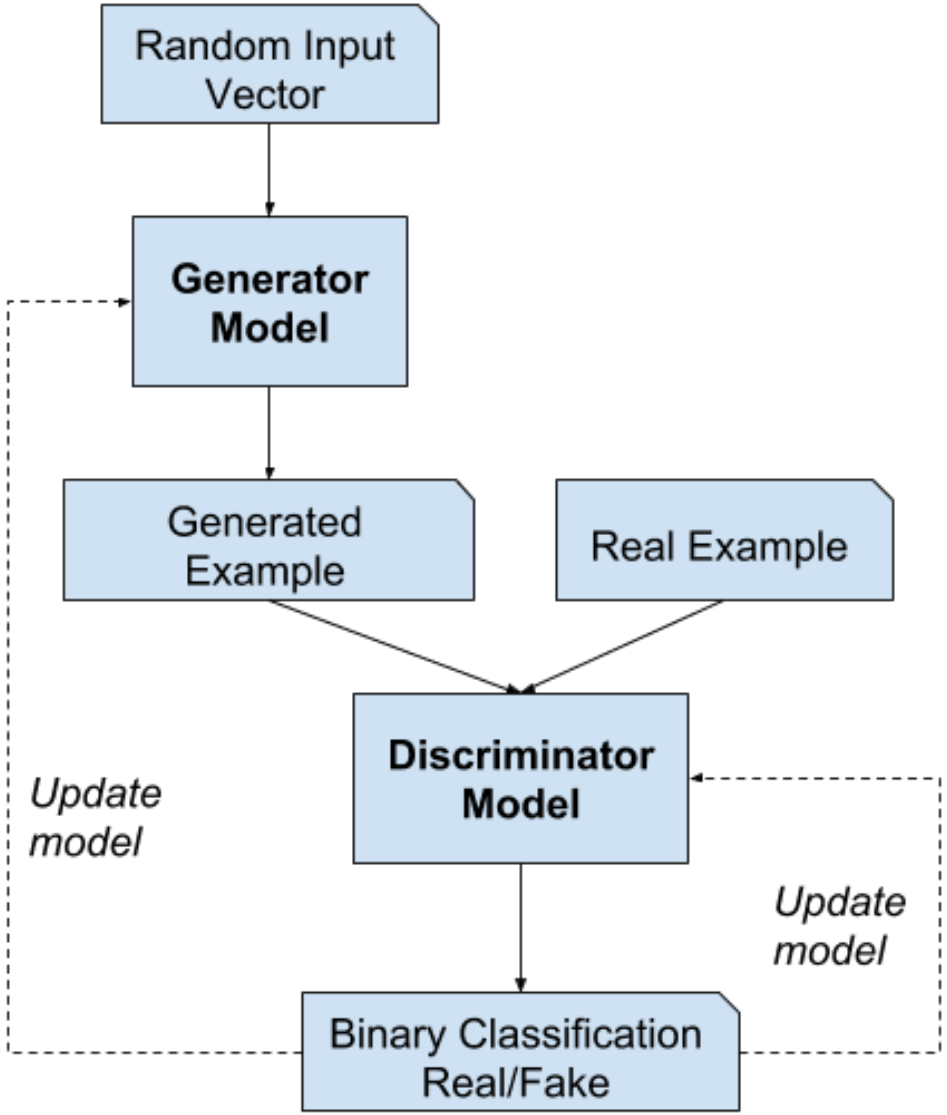
Project	Type of synthetic data	Aim of using synthetic data
PRECISE4Q **Personalized Medicine by Predictive Modeling in Stroke for Better Quality of Life**	Perfusion images from existing acute stroke imaging modalities	Using perfusion images, synthetically generated from existing acute stroke imaging, to simplify perfusion imaging - which requires contrast agents, advanced processing, and specialized skills
CINECA Synthetic Cohort Dataset	- EUROPE UK1: 76 synthetic subject attributes and phenotypic data derived from UKBiobank that match - EUROPE CH SIB: 6733 synthetic samples containing both phenotypic and genotypic information derived from CoLaus and PsyColaus cohorts	To increase accessibility to cohort data for standards development while mitigating ethical and legal privacy concerns that arise with cohort data sharing, including pseudonymized data
CPRD - Clinical Practice Research Datalink	- Cardiovascular disease synthetic dataset - COVID-19 symptoms and risk factors synthetic dataset	Training purposes; improving algorithms; machine learning workflows
SIMULACRUM	Oncological patient records	To support study hypotheses generations and formulation of research questions
INVEST	Virtual patients	Rare disese

UNIVERSITÀ DI TORINO

BEPH

Generative Adversarial Network

- A GAN comprises 2 models trained using an adversarial process
  - the **generator** generates synthetic data
  - the **discriminator** distinguishes between real and synthetic data
- The **discriminator** is updated to get better at discriminating real and fake samples in the next round
- The **generator** is updated based on how well the generated samples fooled the discriminator



GAN VAE

A VAE is a probabilistic model that learns a compressed (latent) representation of data.

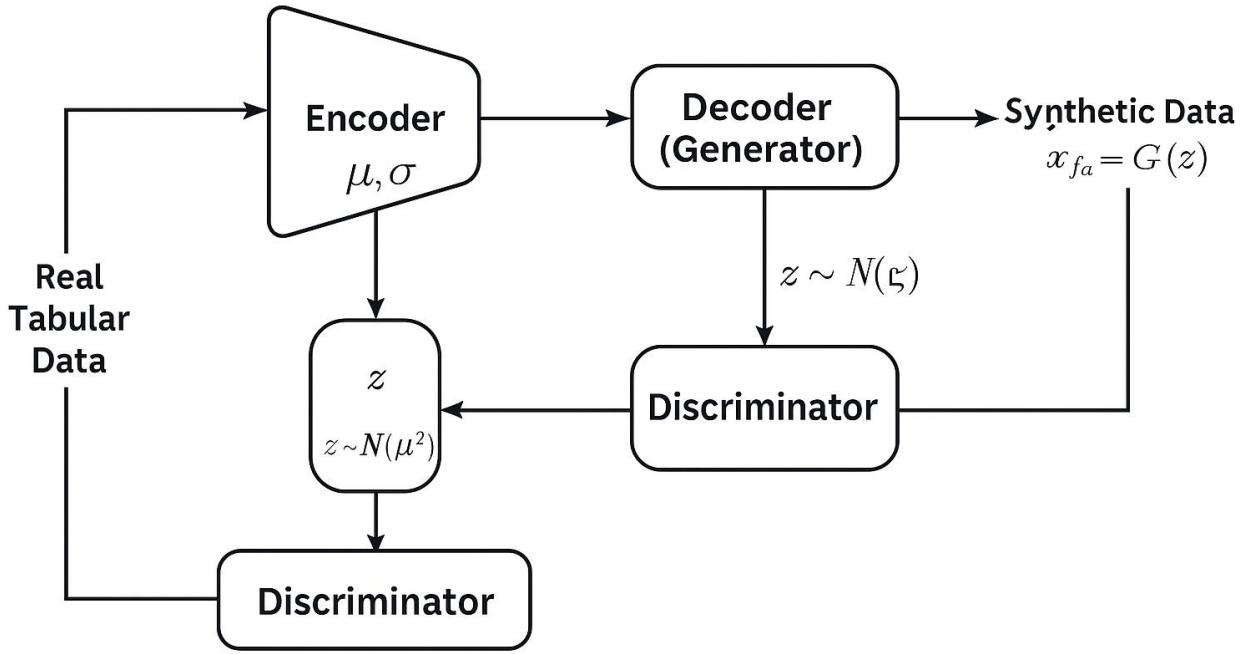
It consists of:

Encoder: Transforms real data into a latent distribution $z \sim N(\mu, \sigma^2)$

Decoder: Reconstructs data from latent representation

Trained to minimize:

1. Reconstruction error between orginal and generated data.
2. KL divergence between the latent distribution and a standard Gaussian



Anatomy of clinical data

- Clinical data represents information of one individual patient
  - age, gender, blood pressure or BMI etc.
- Trials are usually distinguished based on the *endpoint* of interest
  - a quantitative endpoint represents a certain continuous variable measured in multiple patients at one point in time
  - A qualitative endpoint usually is a binary target variable with two possible outcomes
    - “treatment success” “yes” vs. “no
  - a time-to-event endpoint captures values of one or more variables per patient over time, which are considered to have specific relations with each other

A first experiment

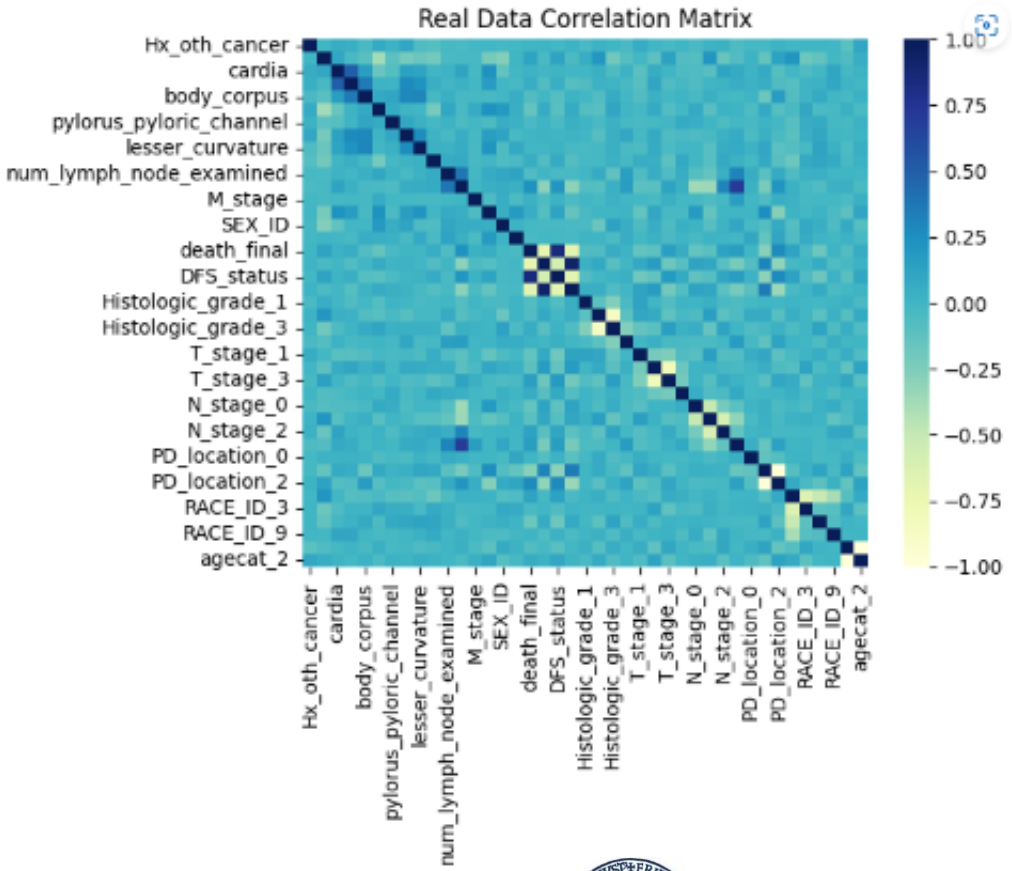
- Randomized phase III trial to compare two different chemotherapy and radiation therapy regimens in treating patients who have undergone surgery for stomach or esophageal cancer
- **Primary endpoint**: Overall survival
- **Secondary endpoint**: disease-free survival

OUTLINE: This is a randomized, multicenter study. Patients are stratified according to depth of tumor penetration (T1 or T2 vs T3 vs T4), lymph node involvement (0 vs 1-3), and extent of lymphadenectomy (D1 or D2 vs D0 or unknown). Patients are randomized to 1 of 2 treatment arms.

- Arm I: Patients receive leucovorin calcium IV and fluorouracil (5-FU) IV on days 1-5 of courses 1, 3, and 4. Courses repeat every 28 days. During course 2, patients undergo radiotherapy 5 days a week and receive 5-FU IV continuously for 5 weeks. Patients rest for 28-35 days between course 2 and 3.
- Arm II: Patients receive epirubicin IV over 3-15 minutes and cisplatin IV over 1 hour on day 1 and 5-FU IV continuously on days 1-21 during course 1. Beginning 1 week later, patients undergo radiotherapy 5 days a week and receive 5-FU IV continuously for 5 weeks. Patients rest for 28-35 days before beginning course 2 of chemotherapy. Patients then receive epirubicin, cisplatin, and 5-FU as in course 1. Treatment repeats every 21 days for 2 courses.

UNIVERSITÀ DI TORINO BEPH

A first experiment

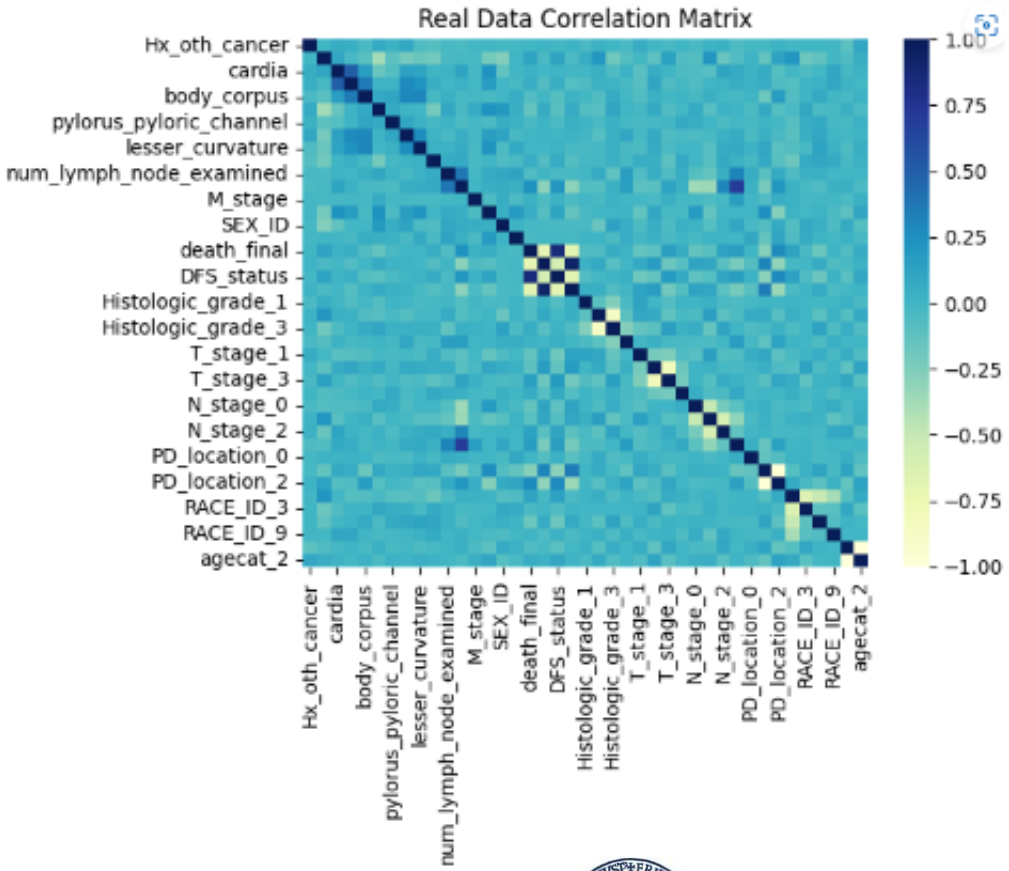


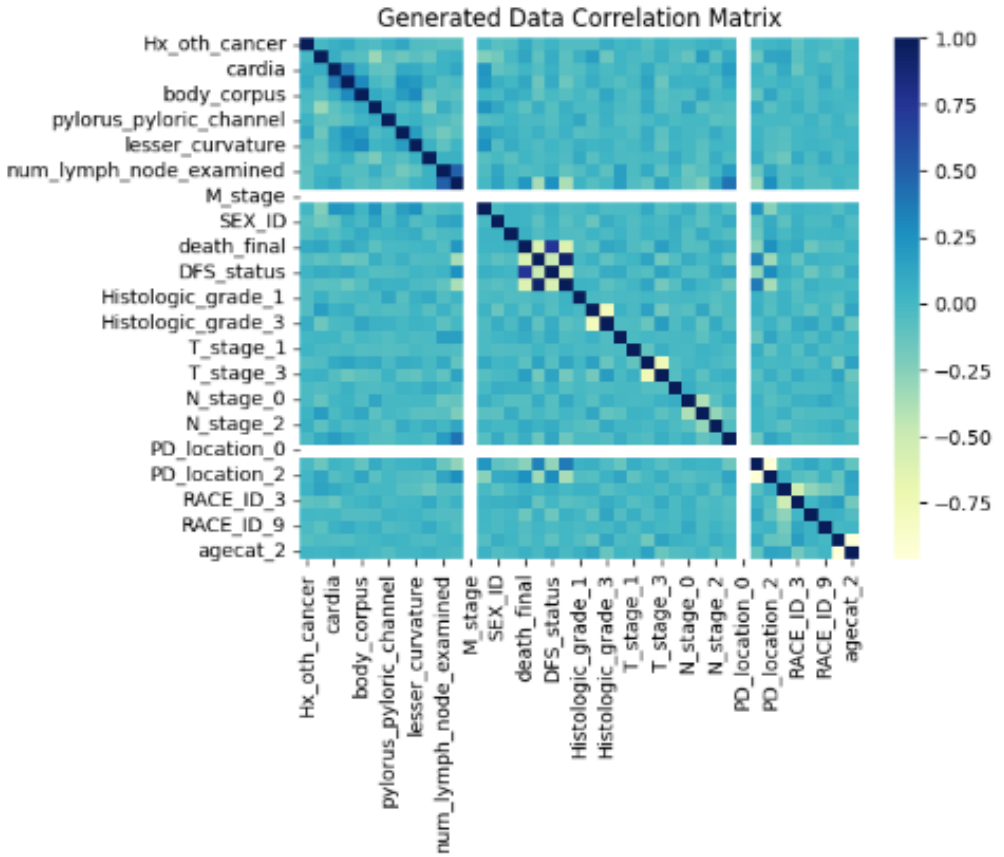
UNIVERSITÀ DI TORINO

BEPH

A first xperiment

simulation of Box's M statistics: p-value of 0.74 through the empirical cumulative distribution



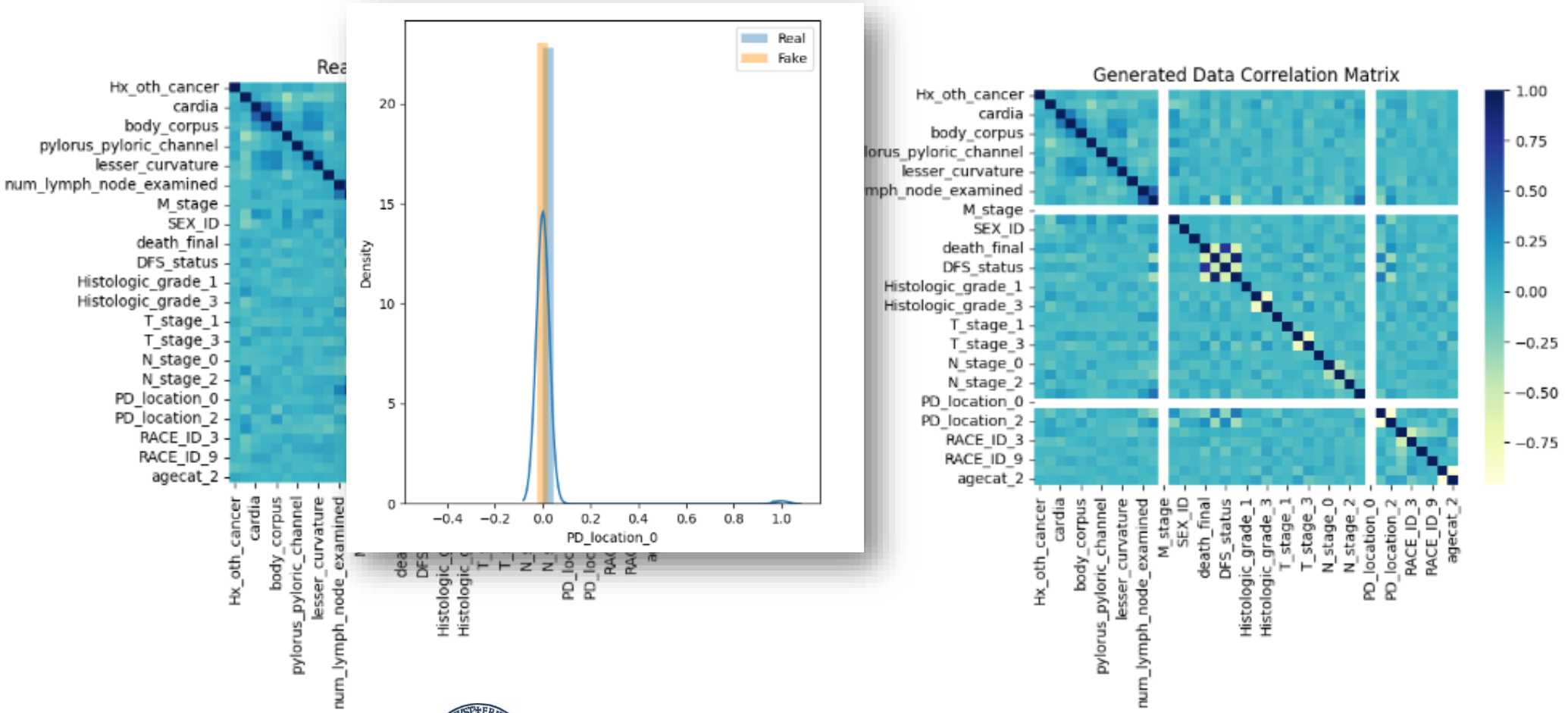


UNIVERSITÀ DI TORINO

BEPH

A first experiment

Location of progression:
unknown 5/546 (.9%)

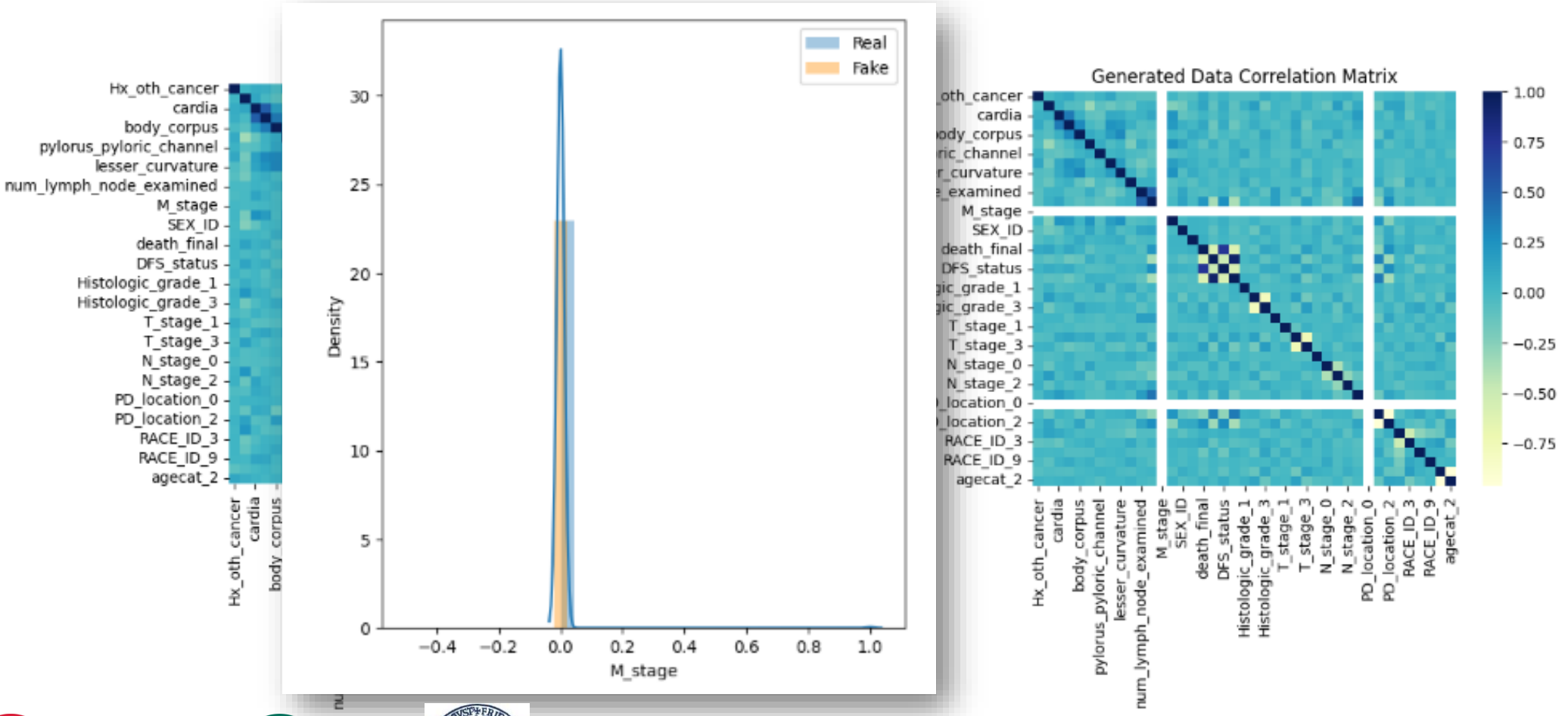


UNIVERSITÀ DI TORINO

BEPH

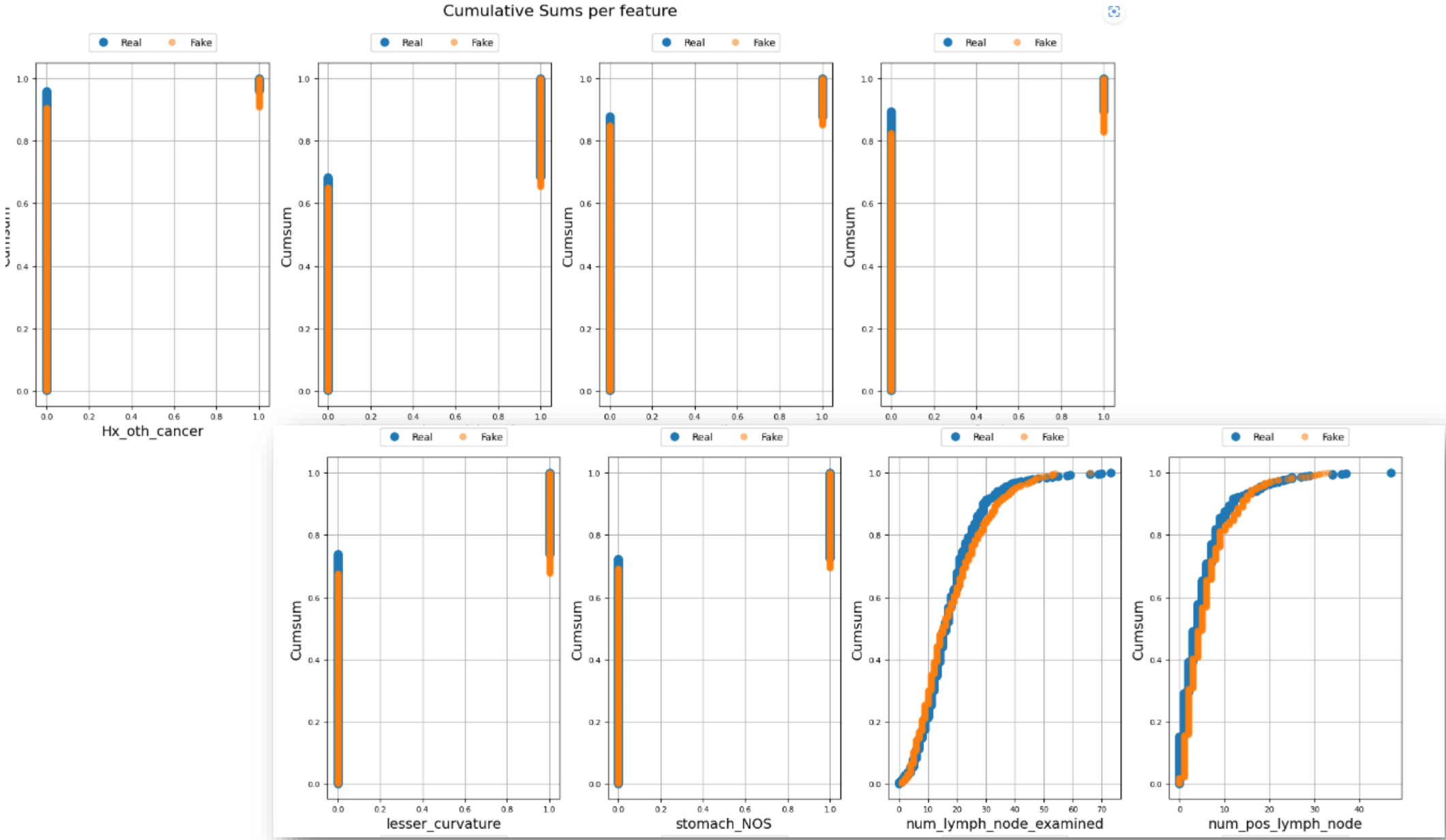
A first experiment

M stage 0: 1/546 (.2%)



UNIVERSITÀ DI TORINO

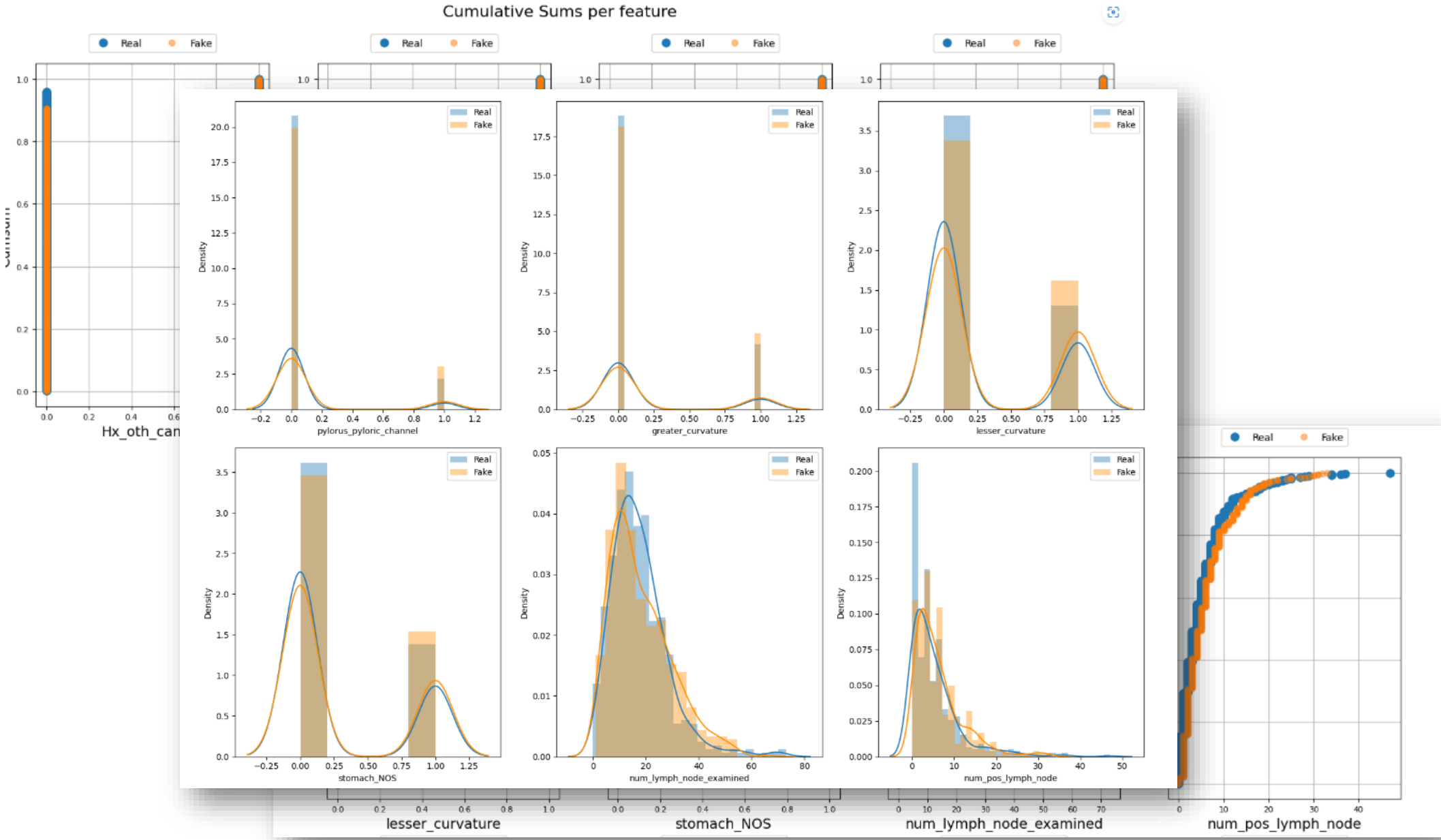
BEPH



UNIVERSITÀ DI TORINO

BEPH

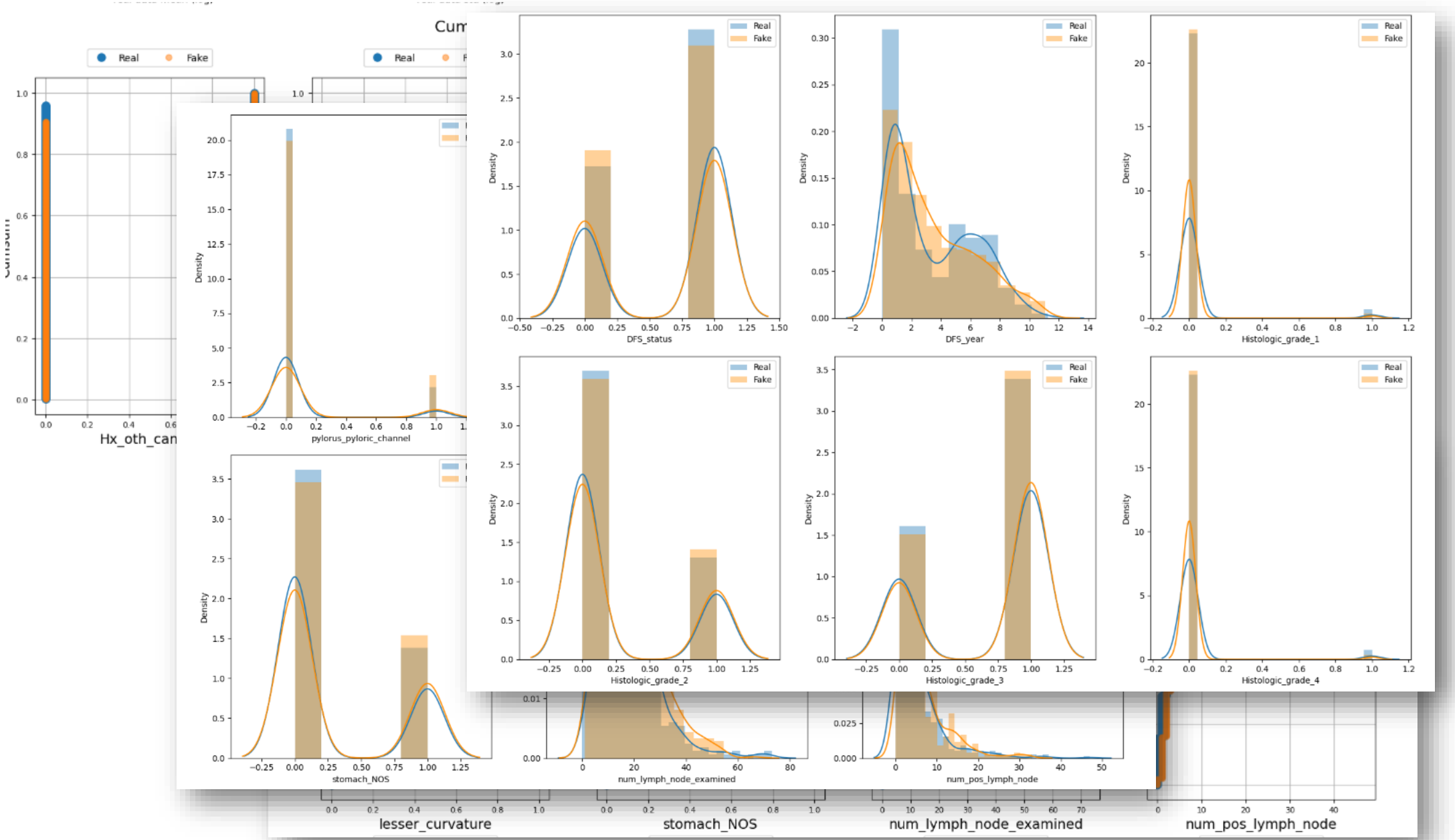
FERRARIAE UNIVERSITAS 13 91 EX LABORE FRUCTUS



UNIVERSITÀ DI TORINO

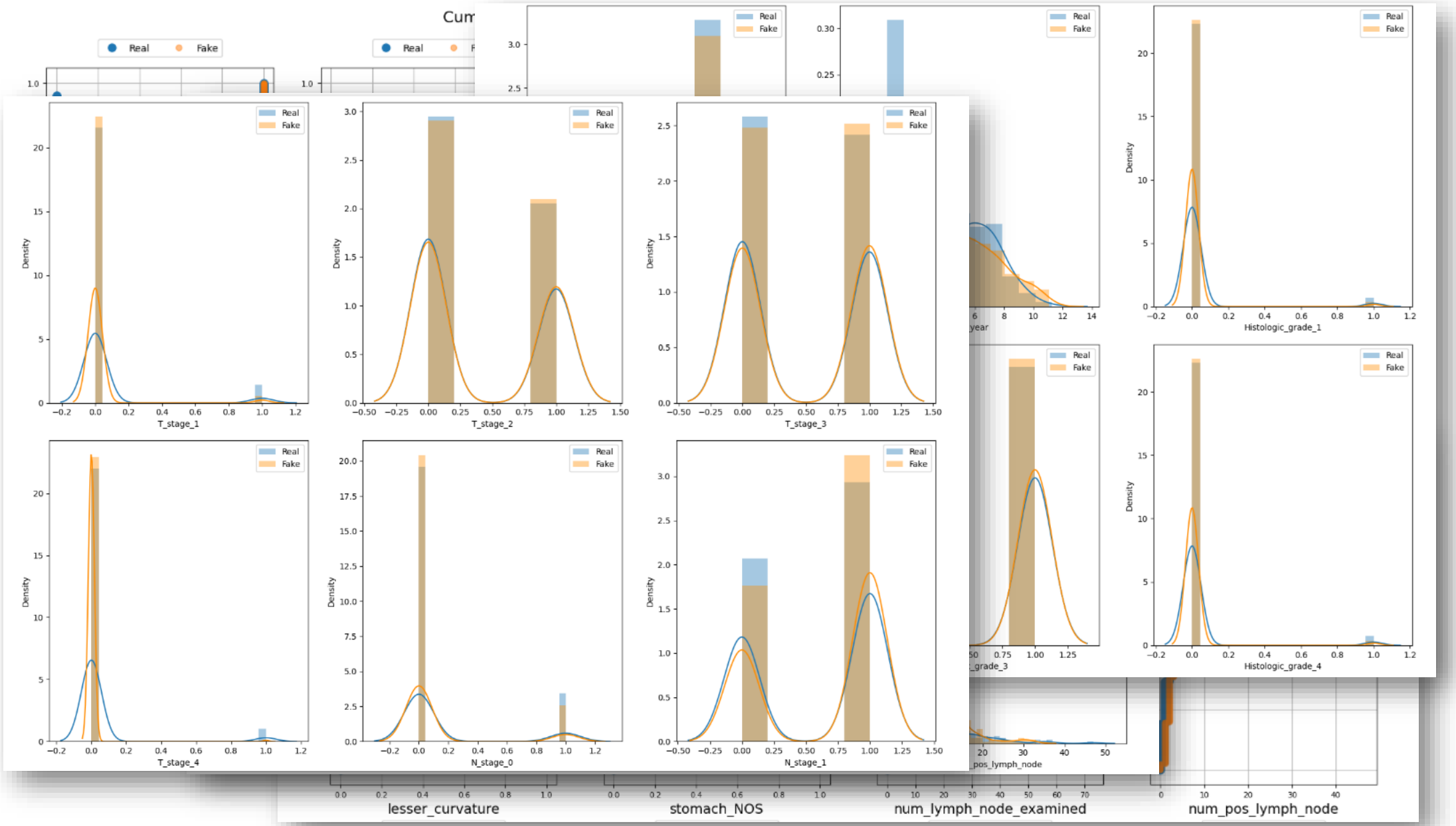
BEPH

FERRARIAE UNIVERSITAS 13 91 EX LABORE FRUCTUS



UNIVERSITÀ DI TORINO

BEPH

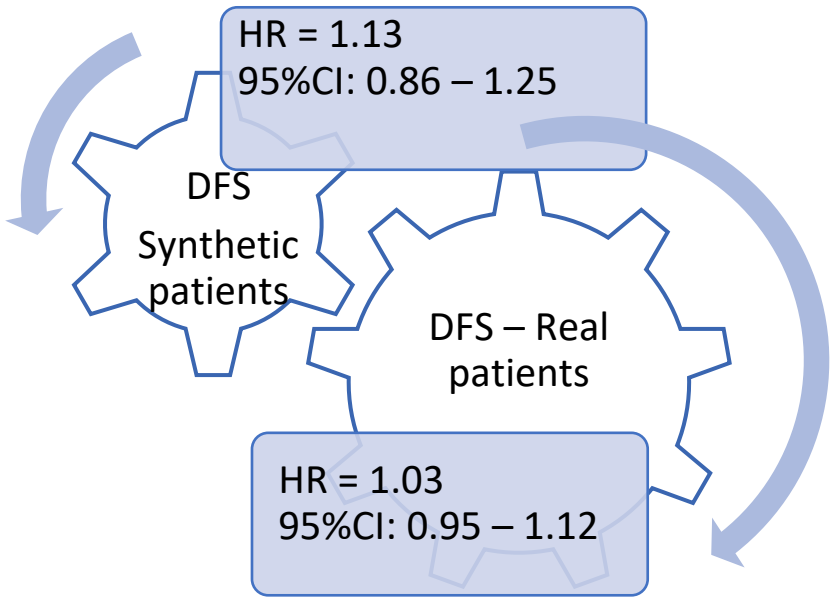


UNIVERSITÀ DI TORINO

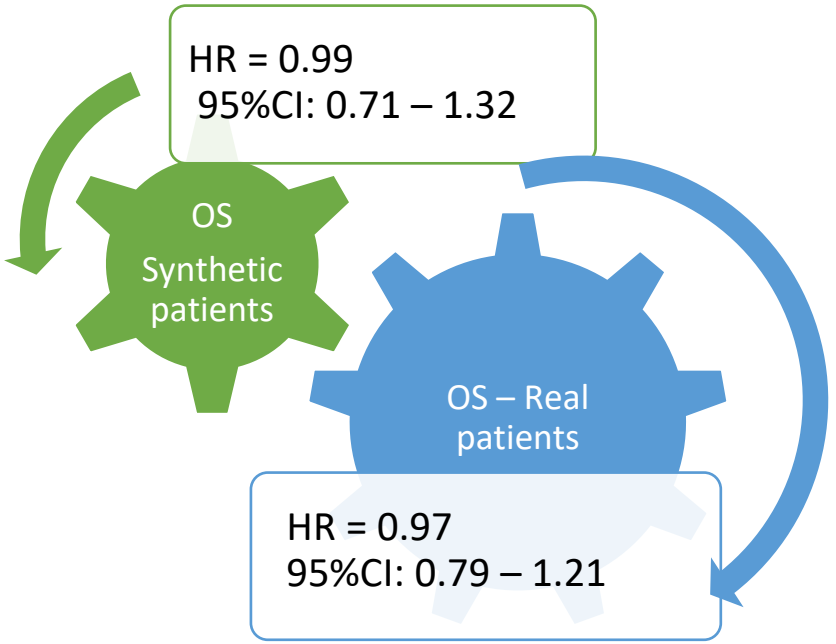
BEPH

Treatment estimate

DISEASE FREE SURVIVAL



OVERALL SURVIVAL



Chemotherapy plus or minus bevacizumab for platinum-sensitive ovarian cancer patients recurring after a bevacizumab containing first line treatment: The randomized phase 3 trial MITO16B-MaNGO OV2B-ENGOT OV17.

Authors: Sandro Pignata, Domenica Lorusso, Florence Joly, Ciro Gallo, Nicolett

behalf of MITO, GINECO, MaNGO, SAKK and HeCOG groups | AUTHORS INFO

VOLUME 29 · NUMBER 27 · SEPTEMBER 20 2011

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Carboplatin Plus Paclitaxel Versus Carboplatin Plus Pegylated Liposomal Doxorubicin As First-Line Treatment for Patients With Ovarian Cancer: The MITO-2 Randomized Phase III Trial

CrossMark

Carboplatin plus paclitaxel once a week versus every 3 weeks in patients with advanced ovarian cancer (MITO-7): a randomised, multicentre, open-label, phase 3 trial

Sandro Pignata, Giovanni Scambia, Dionyssios Katsaros, Ciro Gallo, Eric Pujade-Lauraine, Sabino De Placido, Alessandra Bologna, Beatrice Weber, Francesco Raspagliesi, Pierluigi Benedetti Panici, Gennaro Cormio, Roberto Sorio, Maria Giovanna Cavazzini, Gabriella Ferrandina, Enrico Breda, Viviana Muraia, Cosimo Sacco, Saverio Cinieri, Vanda Salutari, Caterina Ricci, Carmela Pisano, Stefano Greggi, Rossella Lauria, Domenica Lorusso

UNIVERSITÀ DI TORINO

BEPH

FERRARIAE UNIVERSITAS 13 91 EX LABORE FRUCTUS

Chemotherapy plus or minus bevacizumab for platinum-sensitive ovarian cancer patients recurring after a bevacizumab containing first line treatment: The randomized phase 3 trial MITO16B-MaNGO OV2B-ENGOT OV17.

Authors: Sandro Pignata, Domenica Lorusso, Florence Joly, Ciro Gallo, Nicoletta C
behalf of MITO, GINECO, MaNGO, SAKK and HeCOG groups | AUTHORS INFO & A

VOLUME 29 · NUMBER 27 · SEPTEMBER 20 2011

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Carboplatin Plus Paclitaxel Versus Carboplatin Plus Pegylated Liposomal Doxorubicin As First-Line Treatment for Patients With Ovarian Cancer: The MITO-2 Randomized Phase III Trial

CrossMark

Carboplatin plus
in patients with a
a randomised, mu

Sandro Pignata, Giovanni Scambia, Diony
Francesco Raspagliesi, Pierluigi Benedetti
Viviana Muraia, Cosimo Sacco, Saverio Ci

Study	Treatment	n
MITO 2	Carboplatin + paclitaxel	332
MITO 7	Carboplatin + paclitaxel	701
MITO 16	Carboplatin + paclitaxel + bevacizumab	398

UNIVERSITÀ DI TORINO

BEPH

GAN VAE

A VAE is a probabilistic model that learns a compressed (latent) representation of data.

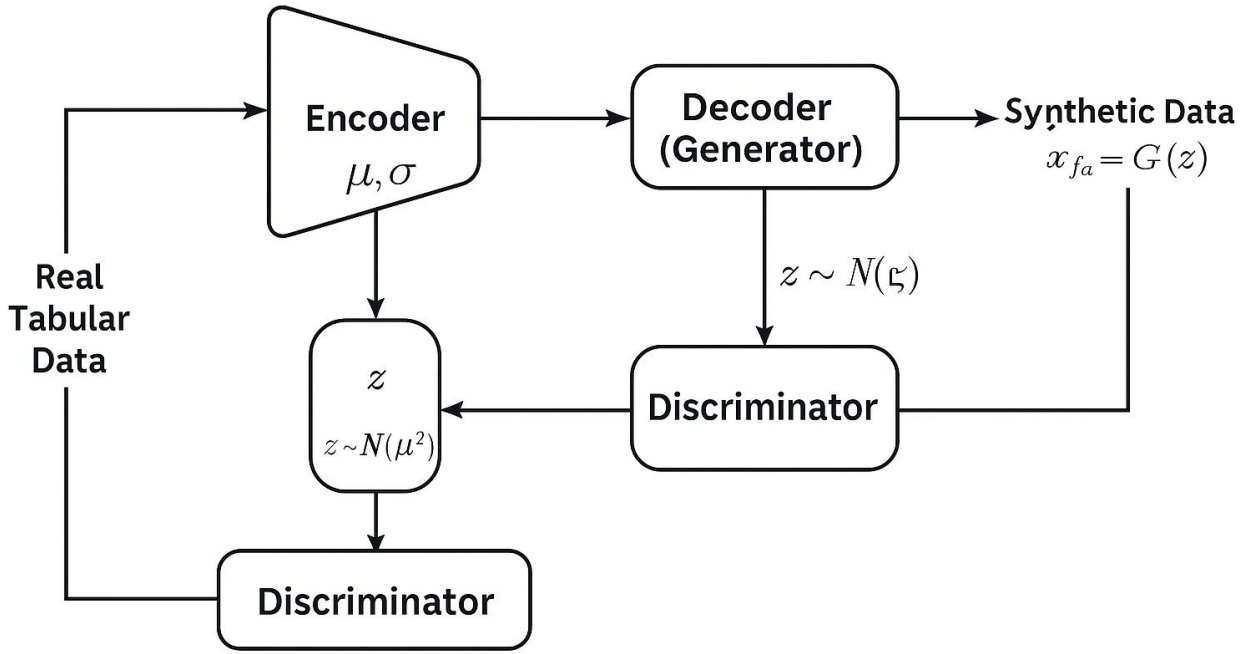
It consists of:

Encoder: Transforms real data into a latent distribution $z \sim N(\mu, \sigma^2)$

Decoder: Reconstructs data from latent representation

Trained to minimize:

1. Reconstruction error between orginal and generated data.
2. KL divergence between the latent distribution and a standard Gaussian.



GAN VAE

A VAE is a probabilistic model that learns a compressed (latent) representation of data.

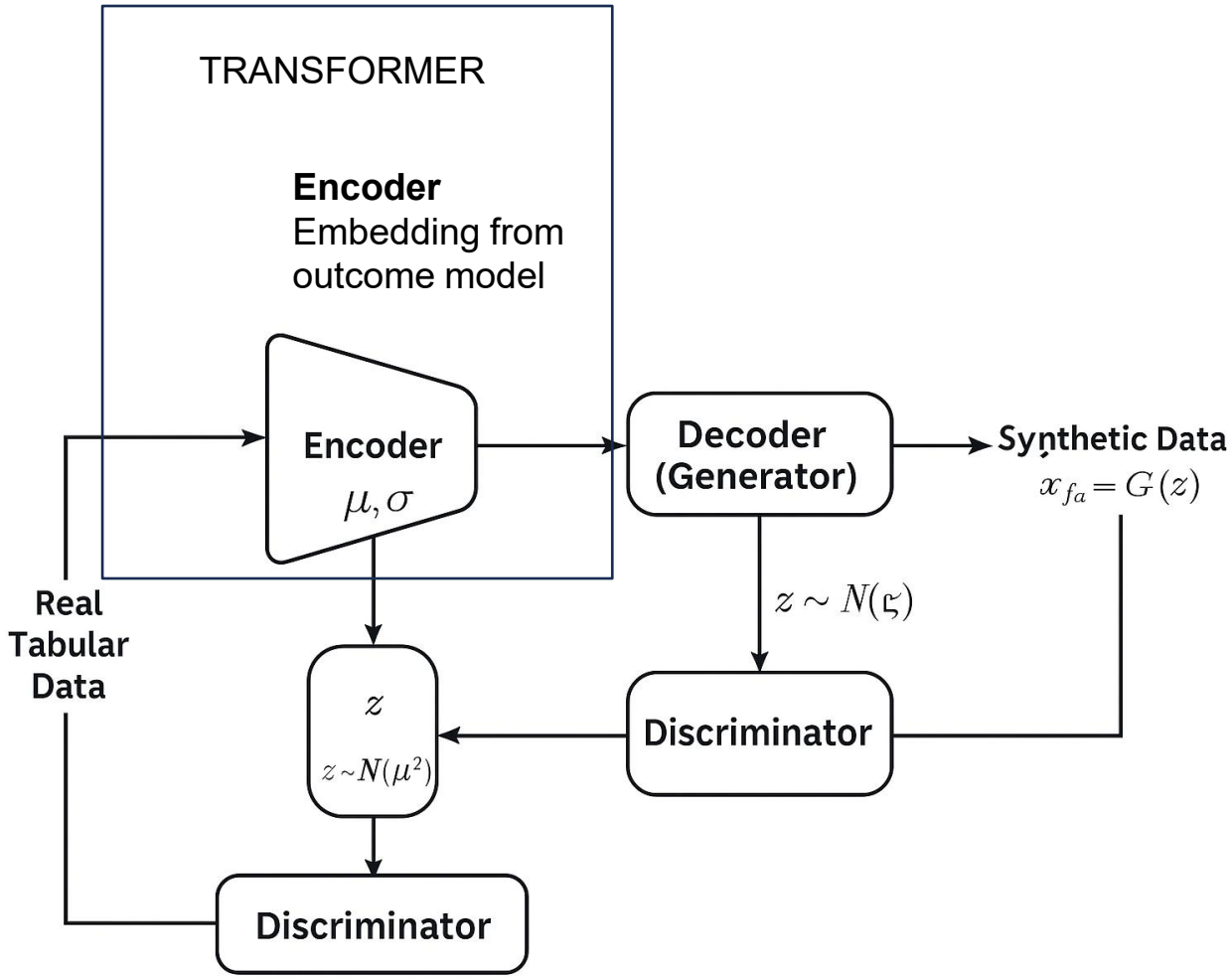
It consists of:

Encoder: Transforms real data into a latent distribution $z \sim N(\mu, \sigma^2)$

Decoder: Reconstructs data from latent representation

Trained to minimize:

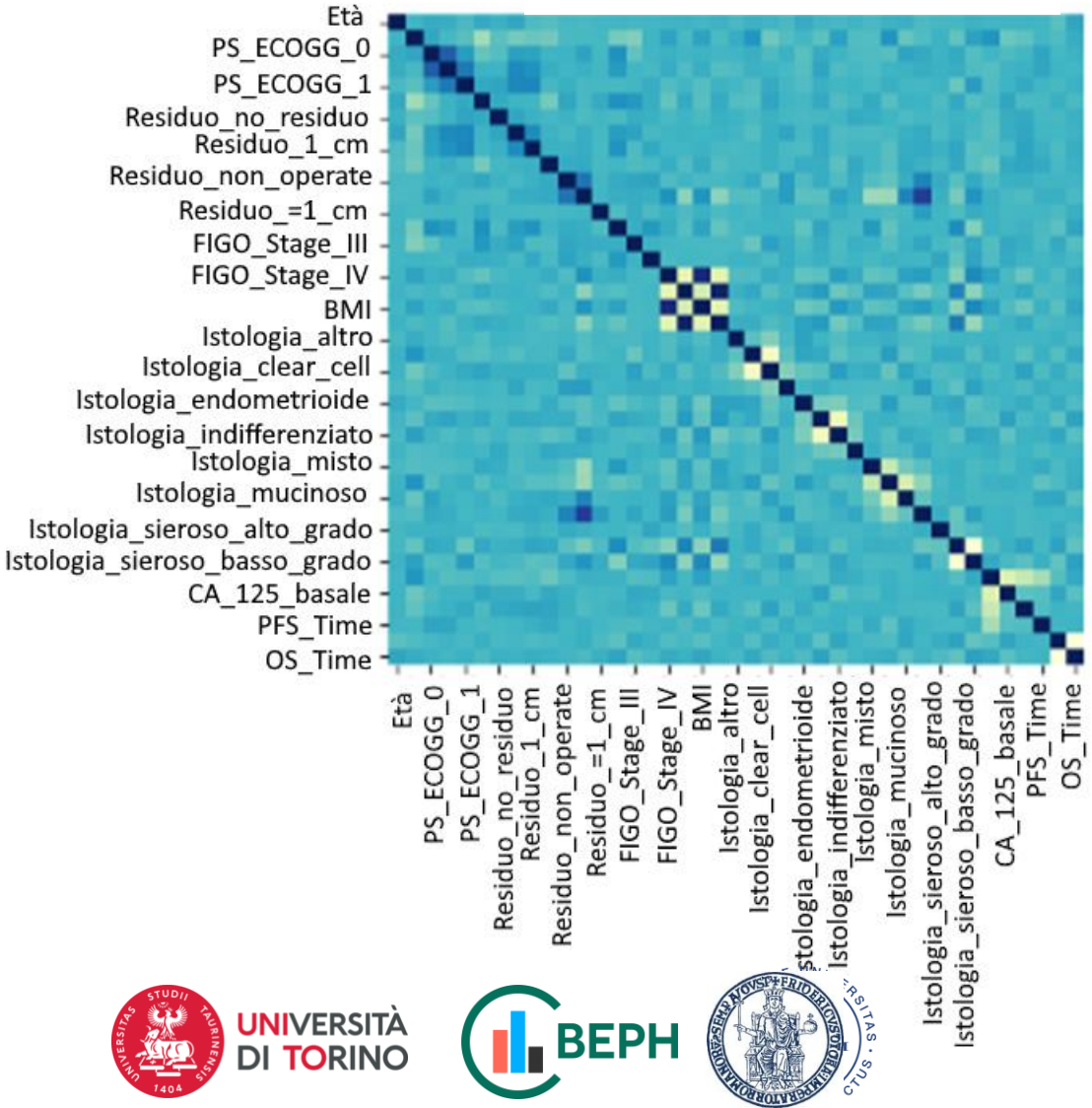
1. Reconstruction error between orginal and generated data.
2. KL divergence between the latent distribution and a standard Gaussian.



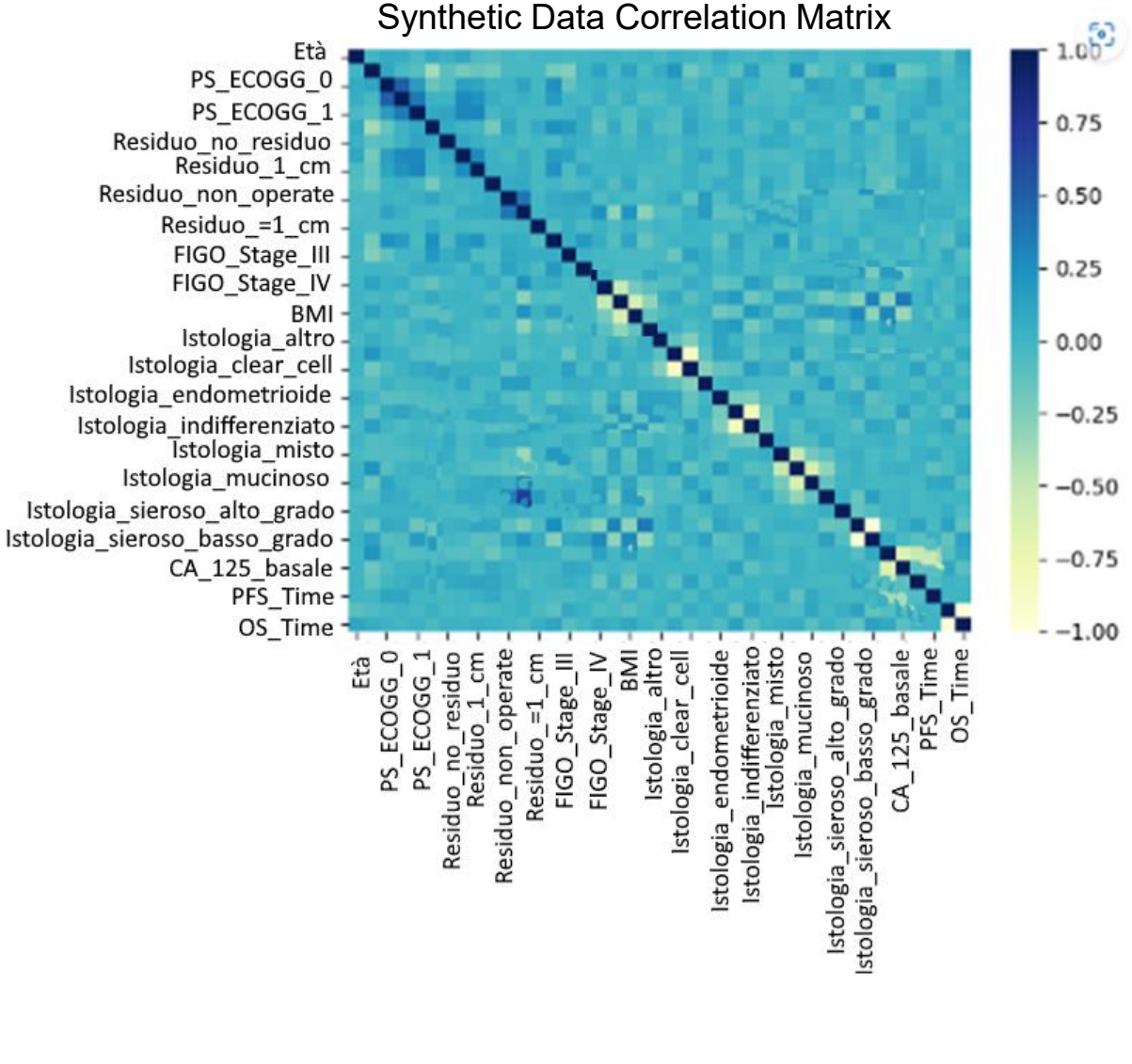
Results

simulation of Box's M statistics: p-value of 0.23 through the empirical cumulative distribution

Real Data Correlation Matrix



Synthetic Data Correlation Matrix



Statistical Validity

	Original Data		Generated Data	
Predictors	**HR**	**95%CI**	**HR**	**95%CI**
Age	1.01	1.00-1.02	1.019	1.01-1.032
PS ECOGG	1.26	1.1-1.43	1.31	1.17-1.50
FIGO Stage	1.73	1.49-2.00	1.68	1.33-2.061
BMI	1.02	1.01-1.04	1.01	1.002-1.033

UNIVERSITÀ DI TORINO

BEPH

Statistical Validity

	Original Data		Generated Data	
Predictors	**HR**	**95%CI**	**HR**	**95%CI**
Age	1.01	1.00-1.02	1.019	1.01-1.032
PS ECOGG	1.26	1.1-1.43	1.31	1.17-1.50
FIGO Stage	1.73	1.49-2.00	1.68	1.33-2.061
BMI	1.02	1.01-1.04	1.01	1.002-1.033
HR	0.51	0.42-0.64	0.61	0.50 – 0.75

UNIVERSITÀ DI TORINO

BEPH

Validity metrics we are working on

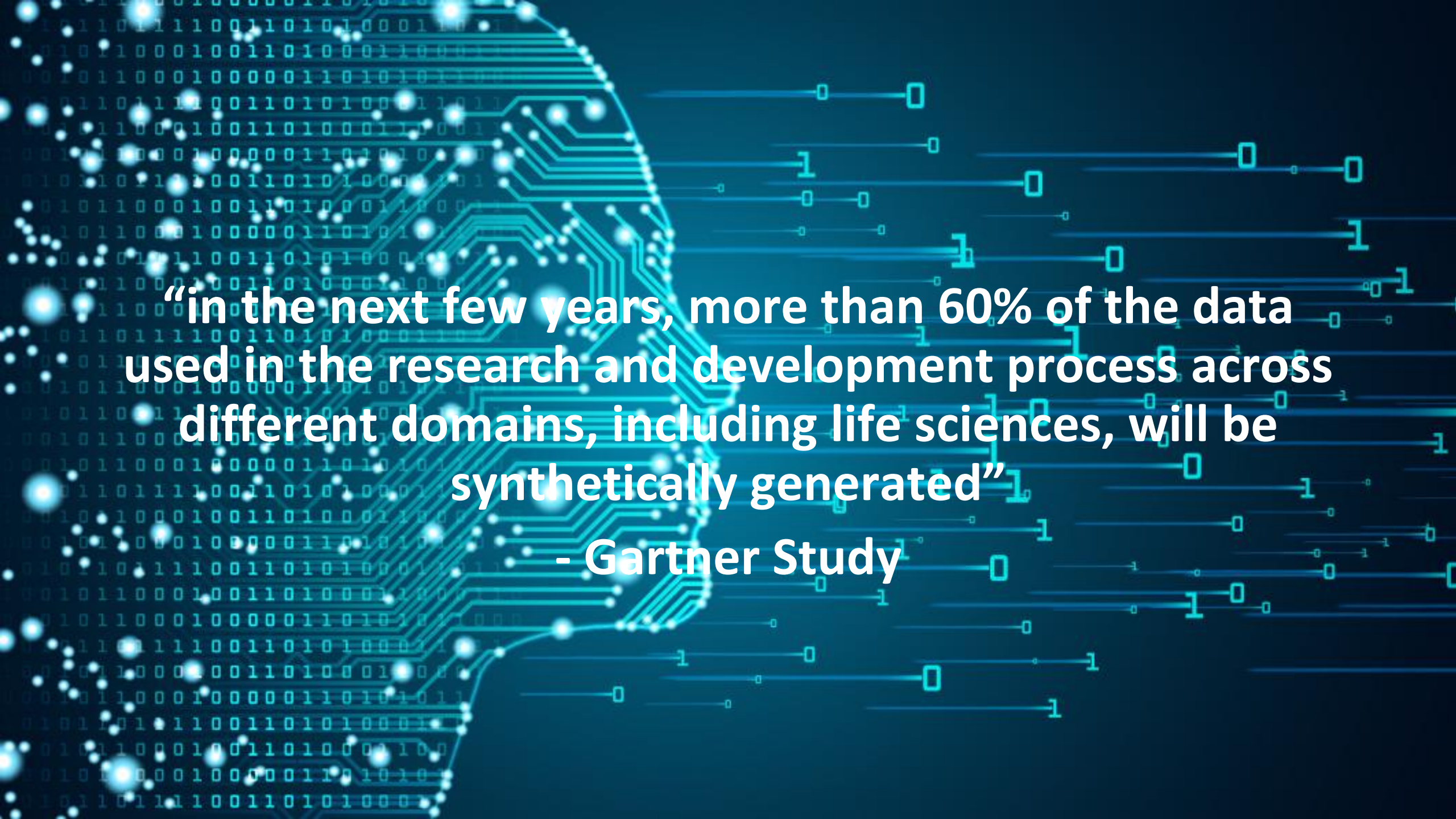
- **Distribution similarity metrics**
  - **Wasserstein Distance -** Measures the "work" needed to align two distributions
  - **Maximum Mean Discrepancy (MMD) -** Kernel-based test for distribution similarity

- How generated data predict as new data in established outcome models?
- Variability in PFS and OS is also due to unboserved confounders

UNIVERSITÀ DI TORINO BEPH

Conclusions

- Using GAN-VAE to generate synthetic patient data to augment control arm in RCT seems promising
  - GAN-VAE mimics real patients
- Unaddressed questions
  - Define proper metrics to assess the validity of synthetic patient data
  - To what extent control arm could be augmented with synthetic patient data
  - In absence of randomization, what statistical outcome analysis?
  - Is the variability in the outcome of synthetic patient data dealt properly?


“in the next few years, more than 60% of the data used in the research and development process across different domains, including life sciences, will be synthetically generated”
- Gartner Study