POLITECNICO
MILANO 1863

MOX

# Health Analytics: how to exploit complex data to inform Precision Medicine and support Clinical Decision Making

Francesca Ieva

MOX – Department of Mathematics, Politecnico di Milano, Italy

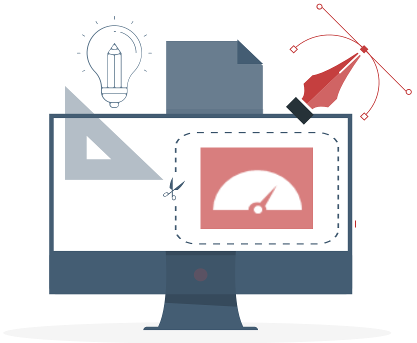
# About me

## *Biosketch*

**Associate Professor of Statistics at MOX (2020-today)**
**Associate Head of the Center for Health Data Science @ Human Technopole (2021-today)**

Senior Researcher in Statistics at MOX (2016-2020)
Junior Researcher in Probability and Statistics at Dept. of Mathematics, Università degli Studi di Milano (2013-2016)
Visiting at MRC Biostatistic Unit @ Cambridge (2013)
PhD in Mathematical Models and Methods for Engineering (2012)
MD in Mathematical Engineering (2008)

francesca.ieva@polimi.it
https://sites.google.com/view/francesca-ieva/home



## *Research interests*

My research is mainly focused on **statistical learning in biomedical context**, from a methodological and applied point of view.
In particular, I deal with **health analytics for complex data in medicine**.

The most part of my activity is concerned with modelling data coming from integration of clinical surveys and administrative databanks. This data drove her scientific interest towards the study of **frailty Multi State Models and Stochastic processes** for *disease progression*, as well as Mathematical Modeling (**Multilevel models and Bayesian nonparametric hierarchical models**) for *Evaluation of Healthcare Processes*.
Moreover, I deepened the study of **depth measures for (multivariate) Functional Data and Functional Data Analysis** applied to *Pharmacoepidemiological* setting for addressing research issues concerning the analysis of complex data like *vital signs* or time varying covariates describing *drugs intake or biomarker evaluation within personalized predictive models*.
In the last years, I enlarged her interests to the study of **Machine Learning and Representation Learning** techniques, aimed at including fingerprints that patients provide in terms of *genomic or medical imaging data into predictive models for personalized medicine.*

# Outline

- Background & Setting
  - From the one-fits-all paradigm to Precision Medicine
  - Two ways for supporting decision medicine
  - Data sources and Health Analytics on Real World Data

- Case studies
  - Clinical Registries & Administrative Data
  - Genomic Data
  - Medical Imaging

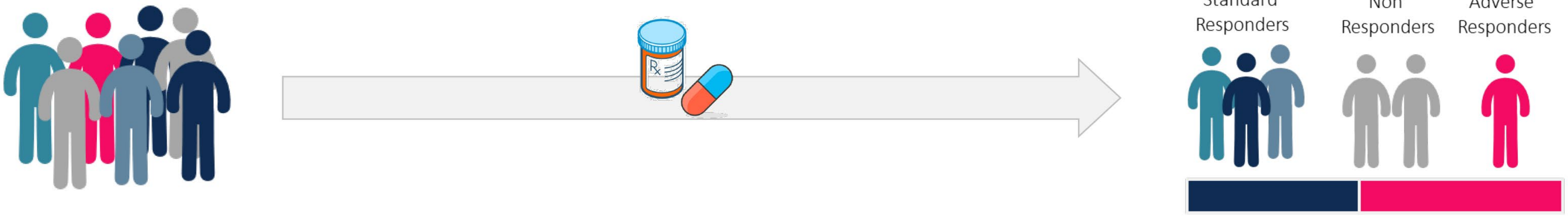
- Take home messages

- References

# Background: Precision Medicine to support clinical decisions

- The medical practice is currently undergoing a transformative era, shifting the paradigm from the primarily reactive medicine of the past to a more proactive and predictive medicine, and trying to outdo the traditional **one-size-fits-all approach** designed for the **average patient**.

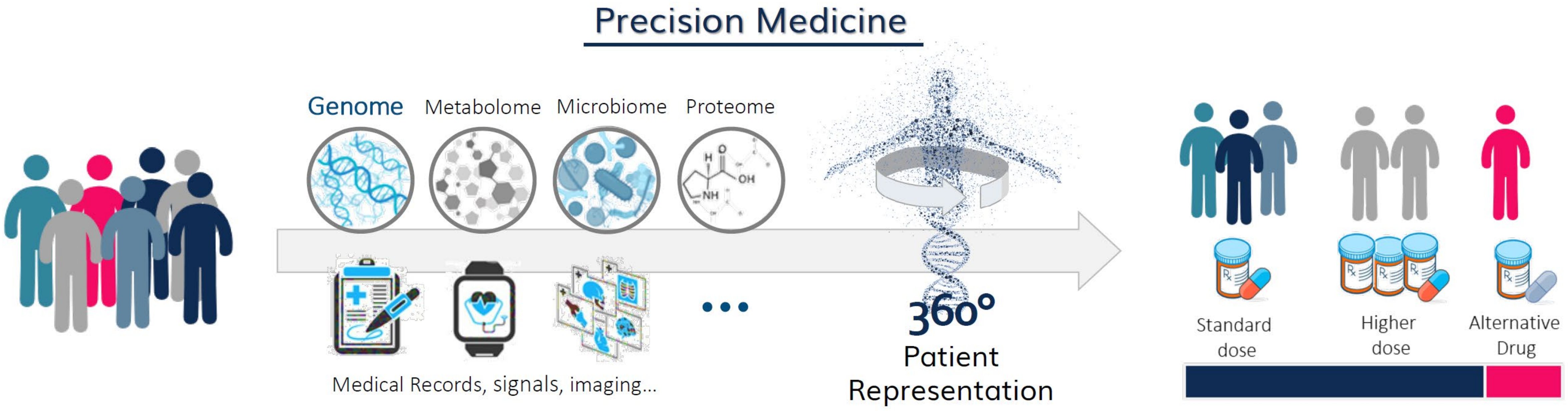
## Standard Framework



- This **new paradigm** takes the name of **Precision Medicine**.

  Rather than treating a disease, the attention now is moving toward streating the individual patients.
  In other words, this methodological framework seeks to include a range of personal data in order to build a *Patient Representation*, that ***combined with a tailored modelling*** can answer relevant clinical research questions and **support clinical decisions**.

# Background: Precision Medicine to support clinical decisions



- ➢ The power of precision medicine lies in its ability to **guide healthcare decisions** toward the most effective treatment for each individual, and thus, improve care quality, while reducing the need for unnecessary diagnostic testing and therapies.

# Background: Precision Medicine to support clinical decisions

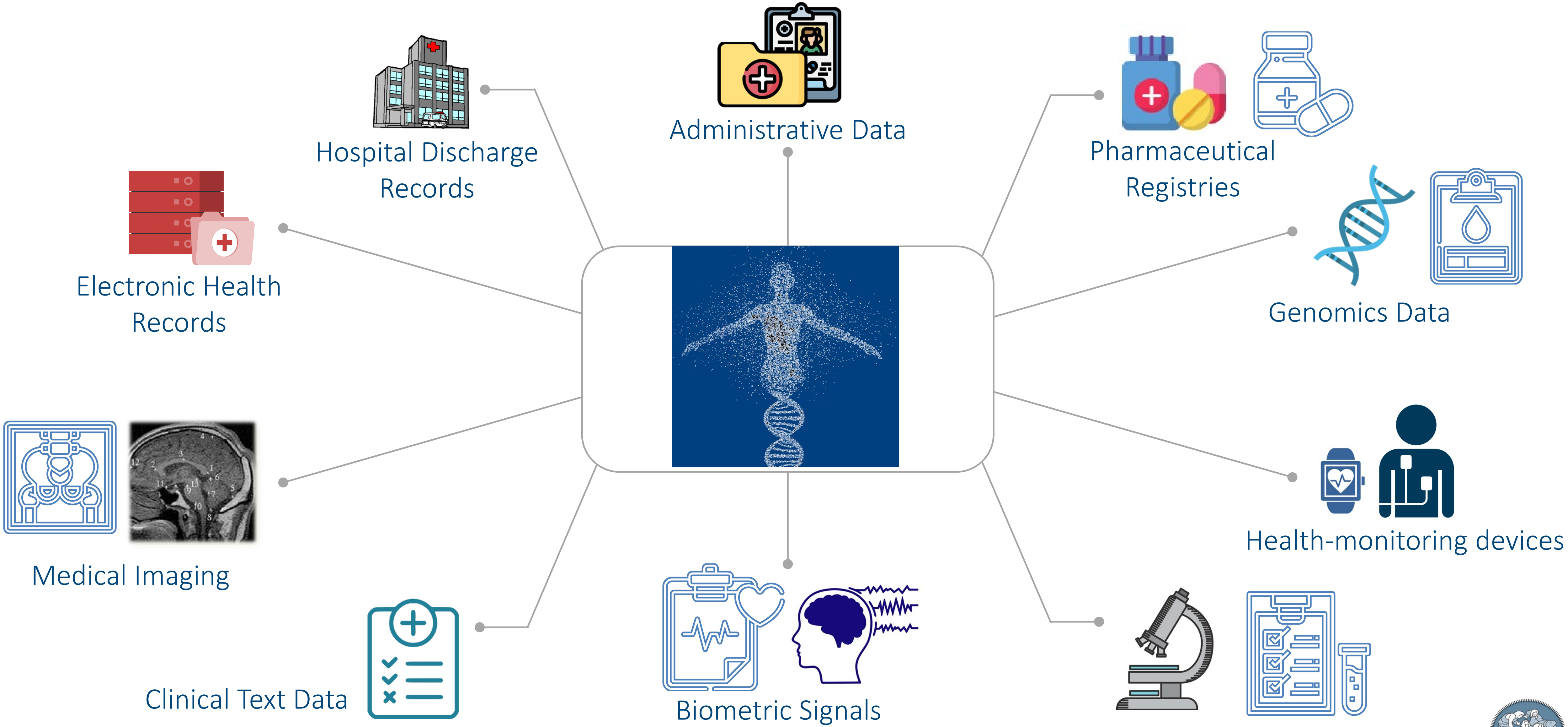
- There are essentially **two ways for supporting decision making in healthcare**:
  - Supporting the Policy Making with Real World Evidence
  - Supporting Clinicians with advanced analytics to exploit the potential of AI in medicine ★

- The **synthesis of the two is still far to come**, but represents the main challenge of the healthcare research.

- Today: focus on ★ and on challenges related to dealing with complex high dimensional data coming from modern clinical practice => explore situations where the use of advanced analytics designed on complex, multi modal and multi omics data allows for an effective support of clinical decision making in the oncological setting.

- **Examples**:
  1. Joint use of Functional Data Analysis within a time-to-event framework as a tool for risk stratification and personalized prediction, motivated by a real problem where the overall survival of patients affected by chronic conditions, in a pharmacoepidemiological setting.
  2. Use of Machine Learning techniques for predicting the development of toxicity adverse events due to radiotherapy in prostate cancer patients, starting from genomic information.
  3. Assessment of the potential of the virtualy biopsy in predictive the treatment response of the patients.

# Healthcare Real World Data & Patient Representation



Hospital Discharge Records

Administrative Data

Pharmaceutical Registries

Electronic Health Records

Genomics Data

Medical Imaging

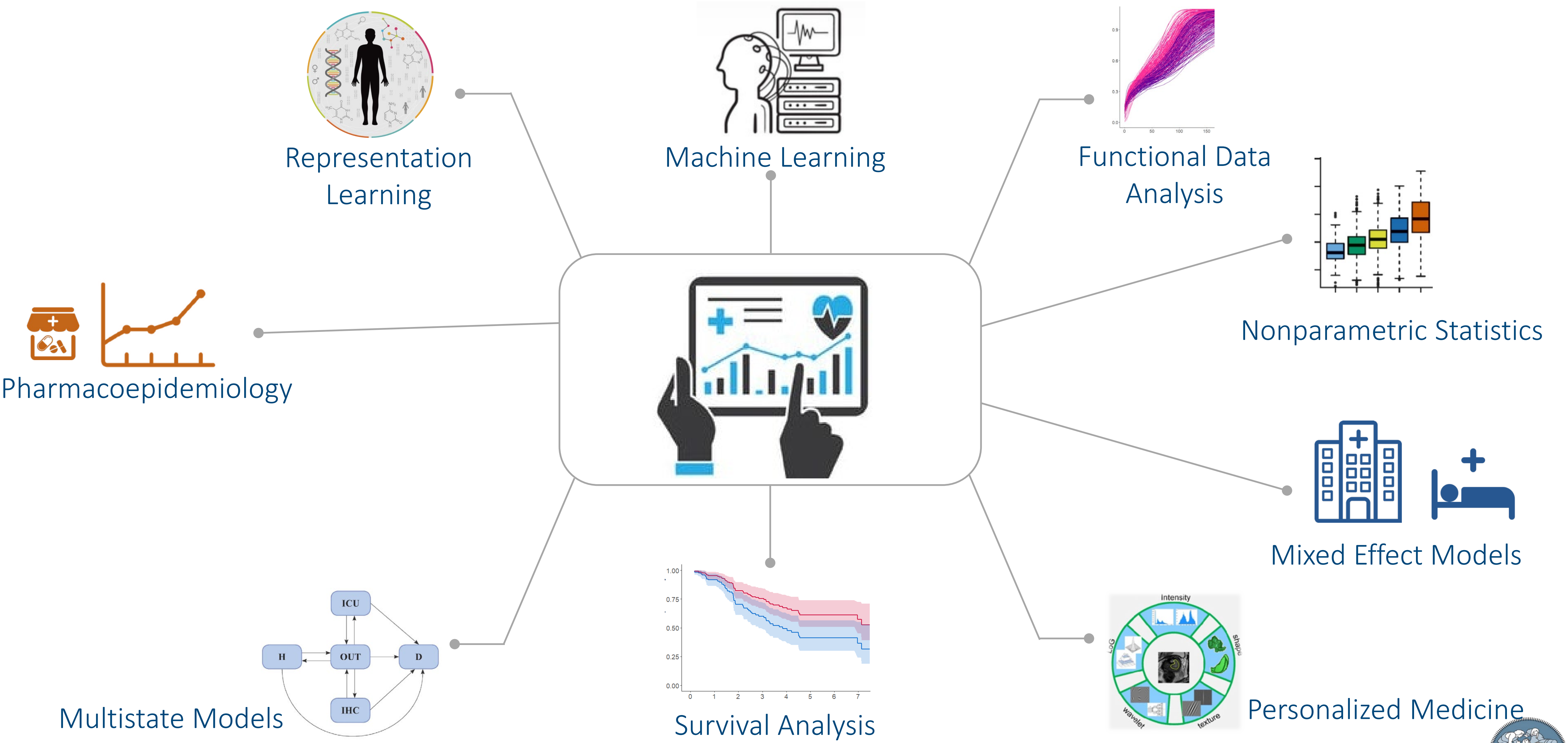
Health-monitoring devices

Clinical Text Data

Biometric Signals

Laboratory Data

# Health Analytics @MOX



Representation Learning

Machine Learning

Functional Data Analysis

Nonparametric Statistics

Pharmacoepidemiology

Mixed Effect Models

Multistate Models

Survival Analysis

Personalized Medicine

# Block I

Data sources: Clinical Registries and Administrative Data

Methods: Functional Data Analysis – Stochastic Process Theory – Survival Modelling

# Why Heart Failure

Regione Lombardia

- Heart Failure (HF) is **widespread** all over the world (especially for > 65 years)
- HF is **chronic** disease characterized by a **high morbidity** and **mortality**
- Advances in **therapy** are changing the prognosis and **improving survival** with

✓*reduction in symptoms* ✓*decrease in the rate of hospitalizations* ✓ *prevention of premature death*

- Two key characteristics in HF treatment:

*Re-hospitalizations*

*Drugs consumption*

*Angiotensin-Converting Enzyme (**ACE**) inhibitors*
*Beta-Blocking (**BB**) agents*
*Anti-Aldosterone (**AA**) agents*

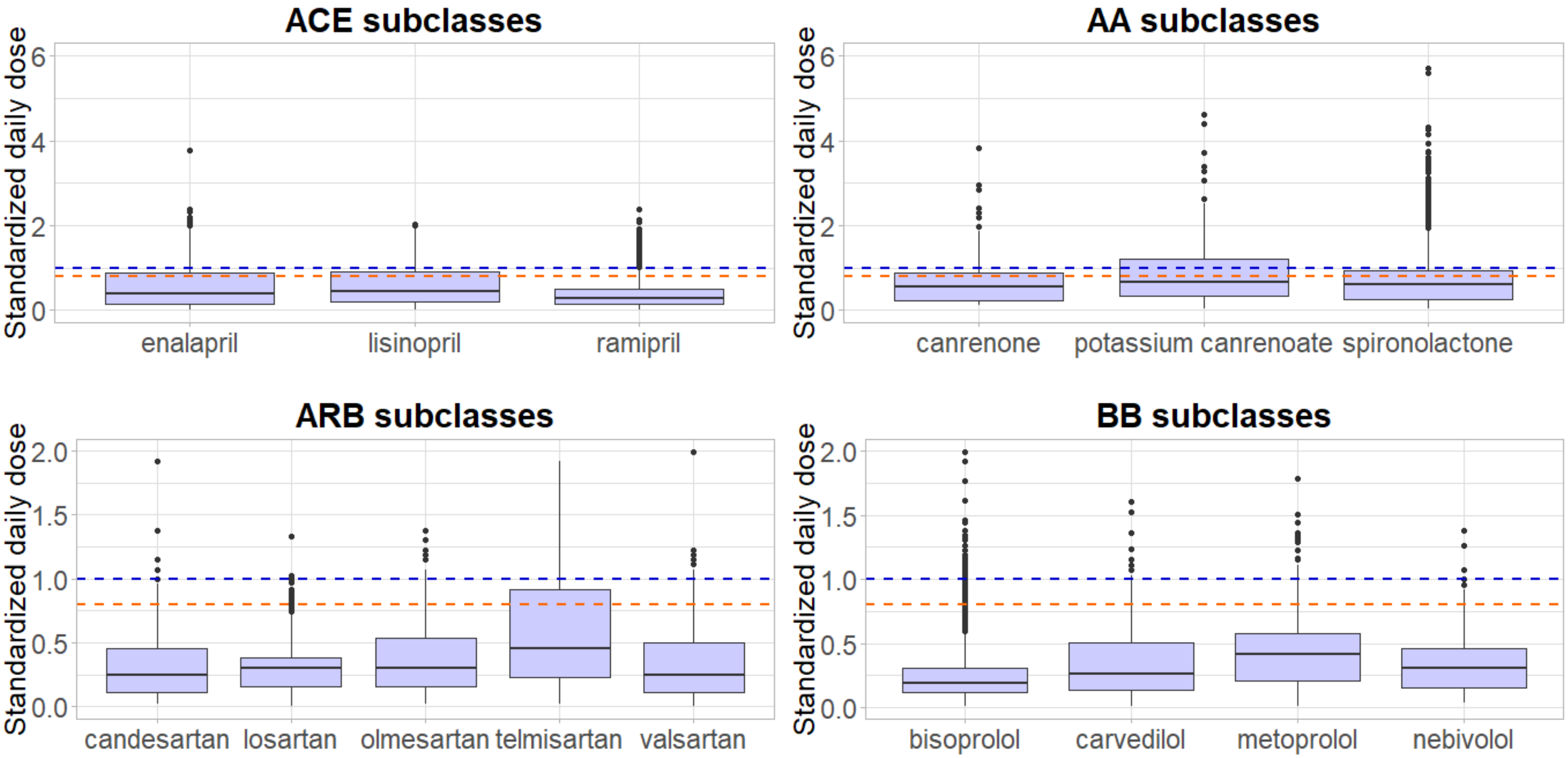
Domande che traducono supporto alle decisioni:
Gestione «ottimale» del pz cronico (tanti) passa da

- Sua capacità di essere aderente a terapia
- Comprensione di come questo influisce su endpoint primari e secondari (ie sopravvivenza e riospedalizzazioi)

=> Quantificazione consente valutazioni economiche, costo/efficacia e quindi informa le policies in sanità

*How does **proper/improper adherence to medication** affect survival in Heart Failure?*
*What is the impact of **re-hospitalizations and subsequent drugs consumption** on survival?*

# Why drugs?



Spreafico *et al.* (2020)
Adherence to disease-modifying therapy in patients hospitalized for Heart Failure: findings from a community-based study.
*American Journal of Cardiovascular Drugs*, 20: 179–190

# Data & Information retrieval

Heart Failure project

**Complex data integration** among different administrative databases: *anagraphic, hospital discharge cards (SDO), pharmacological registries*

Regione Lombardia

## Personal characteristics

Date of birth
Gender
Date of death

## Hospitalizations

Date of admission
Length of stay
Comorbidities

## Pharmaceutical purchases

Date of purchase
ATC code
Covered days

Dataset

Anonymous information from Lombardy Region administrative RWD about **4,541 new-incident patients** hospitalized in 2006-2012 with primary diagnosis of HF.

- Four time-varying processes: **Re-hospitalizations and drugs purchases** [Angiotensin-Converting Enzyme (ACE) inhibitors + Angiotensin Receptor Blockers (ARB), Beta Blockers (BB), Anti-Aldosterone (AA) agents]
- Time-to-event outcome: **Long-term survival**

# Data & Information retrieval

*How can we **model the processes of re-hospitalizations** and **subsequent drugs consumption** over time in HF patients? What is their **impact on long-term survival**?*

- Processes of re-hospitalizations and drug purchases → stochastic process with **recurrent events**.

=> Need to model the **trajectories of the *compensators*** underlying the processes.

- **Administrative Real-World Data** (RWD) → Real-time monitoring of population-based records
- Patients' **clinical history** of **hospitalizations** or **drugs consumption** could be reconstructed using:
  i. administrative data related to admission to hospital (Hospital Discharge Charts);
  ii. pharmaceutical purchases registries.

➢ **Develop methodologies** able **to extract** from RWD **additional information** related to these events in a novel and **tailored** way, properly taking into account their possibly **time-varying** nature.

Mazzali *et al.* (2016). **Methodological issues on the use of administrative data in healthcare research: the case of heart failure hospitalizations in Lombardy Region, 2000 to 2012**. *BMC Health Services Research*, 16 (1): 234

# Research Questions

Characterizing the association between **time-varying** covariates and **time-to-event outcomes** (e.g. death) is a **challenging problem** in the actual clinical/healthcare setting

Drugs intake, treatments, biomarkers
are all processes that change over time

↓

Time-varying covariates

↓

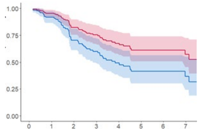
Modulated effect on patient's health status and disease progression

**Idea: representation** of dynamic covariates in terms of **functional data** + dimensionality reduction to plug them into **Cox type regression models**.

Time-varying covariates

*Stochastic processes theory* ⟷ *Functional Data Analysis*

Time-to-event models



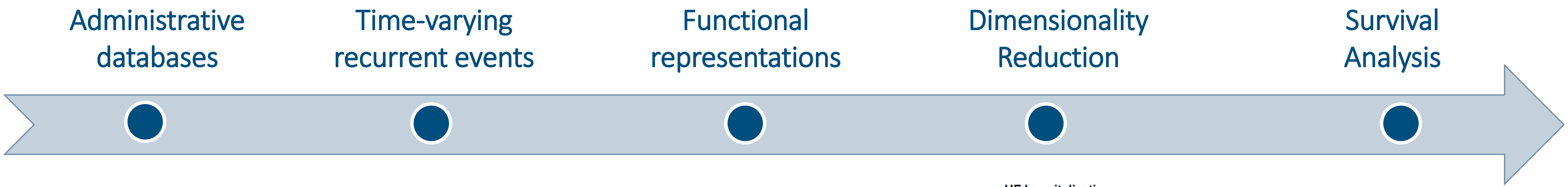
R **Representation**

**Appropriate dynamic characterization of the time-varying processes under study**

**Modelling** M

**Proper modelling of dynamic covariates in a time-to-event framework**

# Functional modeling of recurrent events on time-to-event processes



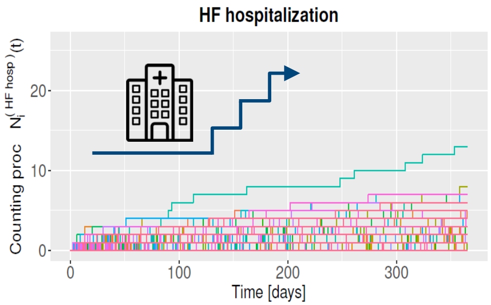
Administrative databases

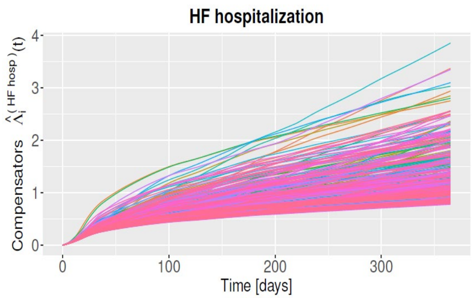
Time-varying recurrent events

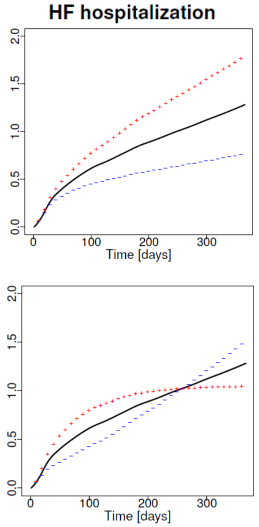
Functional representations

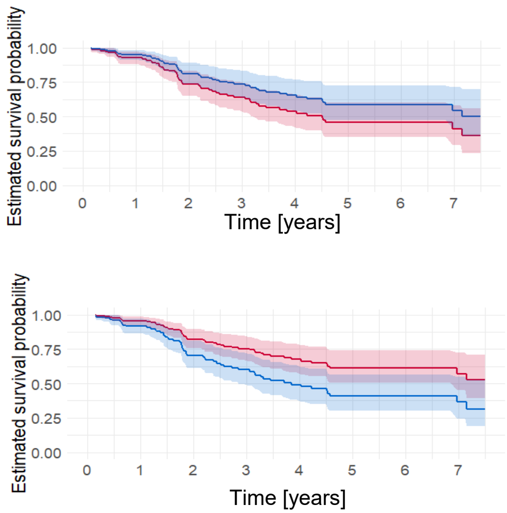
Dimensionality Reduction

Survival Analysis









**R**

*Marked Point Process formulation for Recurrent Events (MPPRE)*

↓

**Retrieve functional trajectories of the compensators of such processes** (which may represent the **rate** at which events happen) by means of Functional Data Analysis theory.

*Multivariate Functional Linear Cox Regression Model (MFLCRM)*

**M**

↓

To quantify the association between the functional compensators and time-to-death.

Spreafico, M., Ieva, F. (2021). Functional modeling of recurrent events on time-to-event processes. *Biometrical Journal*. doi: 10.1002/bimj.202000374

# Counting process formulation

**Concurrent event processes:** *re-hospitalizations, drugs purchases.*

Proper modeling of the concurrent process enables a useful quantification of the effects of the concurrent process on the dynamics of the outcome.

A ***counting process*** is a stochastic process $\{N(t), t \geq 0\}$ with values that are non-negative, integer, and non-decreasing: $N(t) \geq 0$, with jumps of size +1.

- The stochastic **intensity process** $\lambda(t)$ of the counting process $N(t)$ adapted to a filtration $\{\mathcal{F}_t, t \geq 0\}$ is:

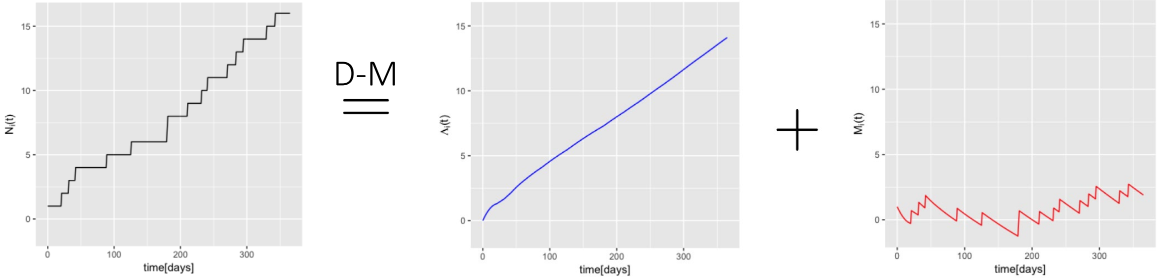
$$\lambda(t) = \lim_{h \to 0} \frac{1}{h} \mathbb{E}\left[N(t+h) - N(t) | \mathcal{F}_t\right]$$

- Counting process formulation for Recurrent Events (**Doob-Meyer decomposition**):

Counting process

Compensator

Martingale residual



$$N(t) \quad = \quad \Lambda(t) = \int_0^t \lambda(s)ds \quad + \quad M(t)$$

→ A counting process where **jumps may have different size** can be modeled as a point process, assuming that a given distribution regulates the size of the jumps.

# Cumulative hazard function of MPP intensity

- **Marked Point Process (MPP):** couple of processes describing the behavior of jumps and marks, whose intensity for individual $i$ related to process $h$ may be modeled as follows

$$\underbrace{\lambda_i^{(h)}\left(t,\mathbf{m}_i^{(h)}\right)}_{\textit{Conditional intensity function}} = \underbrace{\lambda_{ig}^{(h)}(t)}_{\textit{Ground Intensity}}\underbrace{f_i^{(h)}\left(\mathbf{m}_i^{(h)}\right)}_{\textit{Multivariate density of the marks } \mathbf{m}_i^{(h)}}$$

*HP: conditional independence of jump times and marks*

- Let $N_i^{(h)}(t)$ be the stochastic process which counts the observed events of type $h \in \mathcal{H}$ for the $i$-th individual ($i = 1, \dots, n$) with *possibly censored observations* of multiple events. The following distribution for the **conditional intensity function** is assumed:

$$\lambda_i^{(h)}\left(t,\mathbf{m}_i^{(h)}\right) = Y_i^{(h)}(t)\lambda_0^{(h)}(t)\exp\left\{\boldsymbol{\beta}^{(h)^T}\mathbf{x}_i^{(h)}(t) + \boldsymbol{\gamma}^{(h)^T}\mathbf{z}_i^{(h)}(t)\right\} = \lambda_i^{(h)}(t)$$

*marks* ↔ *covariates*

$\mathbf{m}_i^{(h)} \leftrightarrow \mathbf{z}_i^{(h)}(t)$

Idea: reconstruction of the hazard function of the marked counting process (i.e., the compensator) that describes the time-varying event of interest

## Cumulative hazard function or Compensator

$$\boxed{\Lambda_i^{(h)}(t) = \int_0^t \lambda_i^{(h)}(s)ds = \sum_{j=1}^{N_i^{(h)}(t)} \exp\left\{\boldsymbol{\beta}^{(h)^T}\mathbf{x}_i^{(h)}(t_{i,j-1}) + \boldsymbol{\gamma}^{(h)^T}\mathbf{z}_i^{(h)}(t_{i,j-1})\right\}\left[\Lambda_0^{(h)}\left(min\left(t_{i,j}^{(h)},t\right)\right) - \Lambda_0^{(h)}\left(t_{i,j-1}^{(h)}\right)\right]}$$

- $h$ = type of recurrent event process
- $\mathbf{x}_i^{(h)}(t)$ = covariates of the with coefficients $\boldsymbol{\beta}^{(h)}$
- $\mathbf{z}_i^{(h)}(t)$ = covariates related to the marks $\mathbf{m}_i^{(h)}$ with coefficients $\boldsymbol{\gamma}^{(h)}$

- $i$ = individual index
- $0 = t_{i,0}^{(h)} < t_{i,1}^{(h)} < \dots < t_{i,N_i^{(h)}(\tau)}^{(h)}$ sequence of jump times
- $\Lambda_0^{(h)}(t) = \int_0^t \lambda_0^{(h)}(s)ds$ is the cumulative baseline hazard function

# Realizations of each compensator $\Lambda_i^{(h)}(t)$ and relative estimate $\widehat{\Lambda}_i^{(h)}(t)$

For each recurrent event process $h \in \mathcal{H}$, let $0 = t_{i,0}^{(h)} < t_{i,1}^{(h)} < \ldots < t_{i,N_i^{(h)}(\tau)}^{(h)}$ be the sequence of jump times related to process $N_i^{(h)}(t)$ for individual $i$, with $\tau$ being the censoring time (possibly equal for all individuals or not)

$$\Lambda_i^{(h)}(t) = \int_0^t \lambda_i^{(h)}(s)ds = \int_0^t Y_i^{(h)}(s)\lambda_0^{(h)}(s)\exp\left\{\boldsymbol{\beta}^{(h)^T}\mathbf{x}_i^{(h)}(s)+\boldsymbol{\gamma}^{(h)^T}\mathbf{z}_i^{(h)}(s)\right\}ds$$

$$= \sum_{j=1}^{N_i^{(h)}(t)} \int_{t_{i,j-1}^{(h)}}^{min\left(t_{i,j}^{(h)},t\right)} \lambda_0(s)\exp\left\{\boldsymbol{\beta}^{(h)^T}\mathbf{x}_i^{(h)}(t_{i,j-1})+\boldsymbol{\gamma}^{(h)^T}\mathbf{z}_i^{(h)}(t_{i,j-1})\right\}ds$$

$$= \sum_{j=1}^{N_i^{(h)}(t)} \underbrace{\exp\left\{\boldsymbol{\beta}^{(h)^T}\mathbf{x}_i^{(h)}(t_{i,j-1})+\boldsymbol{\gamma}^{(h)^T}\mathbf{z}_i^{(h)}(t_{i,j-1})\right\}}_{\substack{\text{Partial likelihood estimation}\\ \left(\widehat{\boldsymbol{\beta}}^{(h)},\widehat{\boldsymbol{\gamma}}^{(h)}\right)}} \underbrace{\left[\Lambda_0^{(h)}\left(min\left(t_{i,j}^{(h)},t\right)\right)-\Lambda_0^{(h)}\left(t_{i,j-1}^{(h)}\right)\right]}_{\substack{\text{Breslow estimators}\\ \downarrow \\ \text{Constrained smoothing}\\ \widetilde{\Lambda}_0^{(h)}(t)}}$$

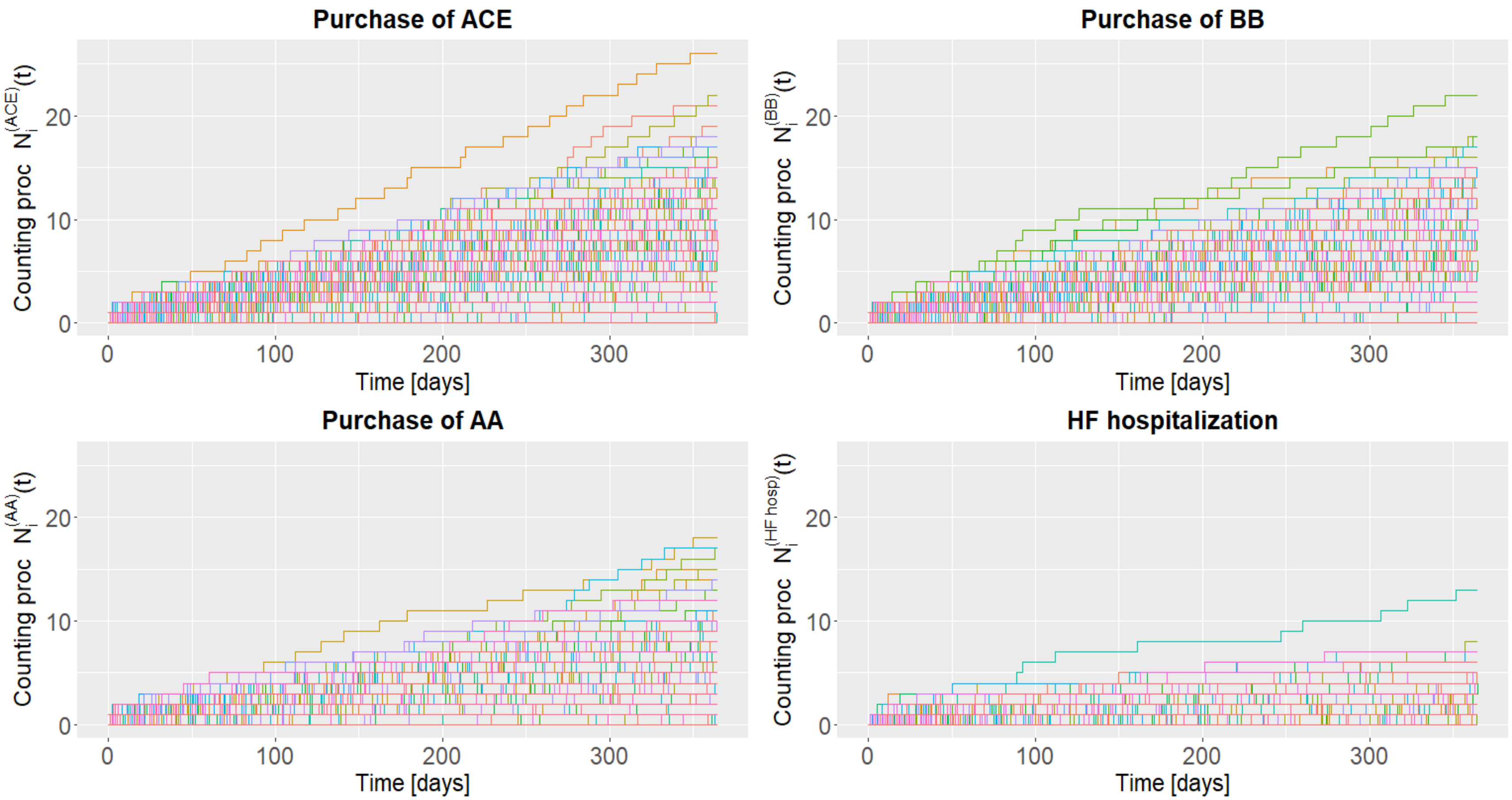
$$\Big\downarrow$$

$$\hat{\Lambda}_i^{(h)}(t) = \sum_{j=1}^{N_i^{(h)}(t)} \exp\left\{\hat{\boldsymbol{\beta}}^{(h)^T}\mathbf{x}_i^{(h)}(t_{i,j-1})+\hat{\boldsymbol{\gamma}}^{(h)^T}\mathbf{z}_i^{(h)}(t_{i,j-1})\right\}\left[\tilde{\Lambda}_0^{(h)}\left(min\left(t_{i,j}^{(h)},t\right)\right)-\tilde{\Lambda}_0^{(h)}\left(t_{i,j-1}^{(h)}\right)\right]$$

# MPP approach for drug purchases and HF re-hospitalizations

**Drug purchases** (ACE or BB or AA) and **HF re-hospitalizations** events can be seen as a **marked point processes** (**MPPs**) with:

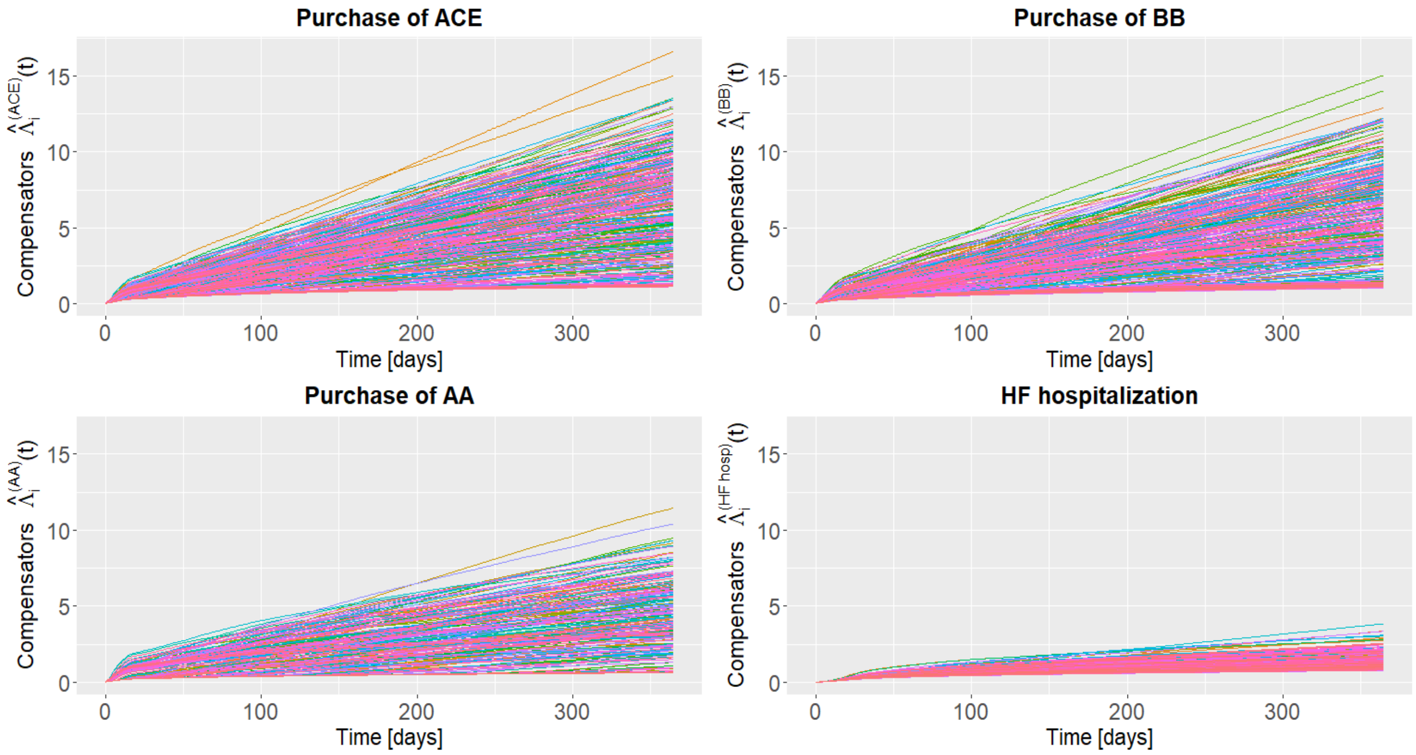
- *jump times* equal to event times
- *jump marks* equal to the *duration of the prescription* or *length of stay in hospital*



Four types of recurrent event processes: $h \in \mathcal{H} = \{ACE, BB, AA, HF\}$

# Functional compensators of drug purchases and HF re-hospitalizations MPPs

- Four time-varying processes (MPPs): **drug purchases** (ACE or BB or AA) and **HF re-hospitalizations**
- Functional compensators of the MPPs: $\left\{\widehat{\Lambda}_i^{(h)}\right\}_{h\in\mathcal{H}} = \left\{\widehat{\Lambda}_i^{(ACE)}, \widehat{\Lambda}_i^{(BB)}, \widehat{\Lambda}_i^{(AA)}, \widehat{\Lambda}_i^{(HF)}\right\}$



*Compensators are our functional data used to enrich the information available for modelling survival*

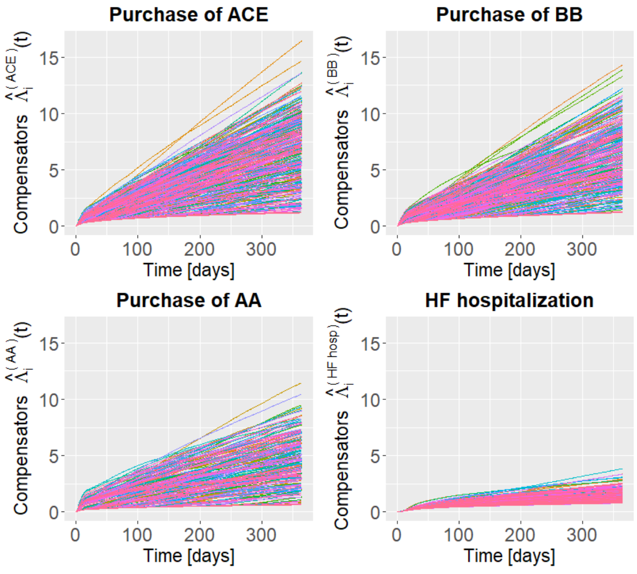
↓

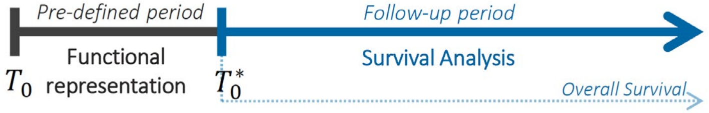
*Highlight **trends and variations in the shape** of the processes over time*

R

- ✓ **Expected number of events by time $t$,** given the covariates → *Dynamic evolution of the events risk*
- ✓ Higher the curve → higher the cumulative risk of a new event
- ✓ The variability of the compensators across different patients reflects the **variability of the realizations** of their recurrent events.

# Multivariate Functional Linear Cox Regression Model for long-term survival





## Multivariate Functional Linear Cox Regression Model

includes the functional compensators $\left\{\hat{\Lambda}_i^{(h)}\right\}_{h\in\mathcal{H}}$ with $\mathcal{H} = \{ACE, BB, AA, HF\}$ in the classical Cox model using the following form:

$$\eta_i\left(t \middle| \boldsymbol{\omega}_i, \left\{\hat{\Lambda}_i^{(h)}\right\}_{h\in\mathcal{H}}\right) = \eta_0(t) \exp\left\{\boldsymbol{\theta}^T \boldsymbol{\omega}_i + \sum_{h\in\mathcal{H}} \int_{T_0}^{T_0^*} \hat{\Lambda}_i^{(h)}(s)\alpha^{(h)}(s)ds\right\}$$

- Patient's index: $i \in \{1, \dots, N\}$
- Event index: $h \in \mathcal{H} = \{ACE, BB, AA, HF\}$
- $\eta_0(t) =$ baseline hazard function
- $\boldsymbol{\omega}_i =$ vector of baseline covariates with regression parameters $\boldsymbol{\theta}$
- $\left\{\hat{\Lambda}_i^{(h)}\right\}_{h\in\mathcal{H}}$ realizations of the functional compensators for the $i$-th individual, with functional regression parameters $\alpha^{(h)}(s)$

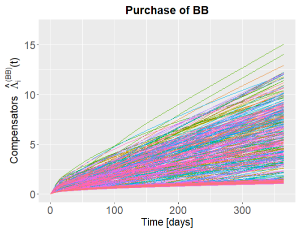
→ *Functional Principal Component Analysis (FPCA)* applied to compensators ends up with a Cox type regression model where the *FPC scores* $f_{ik}^{(h)}$ are treated as standard covariates.

$$\eta_i\left(t \middle| \boldsymbol{\omega}_i, \left\{\hat{\Lambda}_i^{(h)}\right\}_{h\in\mathcal{H}}\right) = \eta_0^*(t) \exp\left\{\boldsymbol{\theta}^T \boldsymbol{\omega}_i + \sum_{h\in\mathcal{H}} \sum_{k=1}^{K_h} f_{ik}^{(h)} \alpha_k^{(h)}\right\}$$

M

$^*\boldsymbol{\omega}_i$ and $K_h$ are chosen by cross-validation

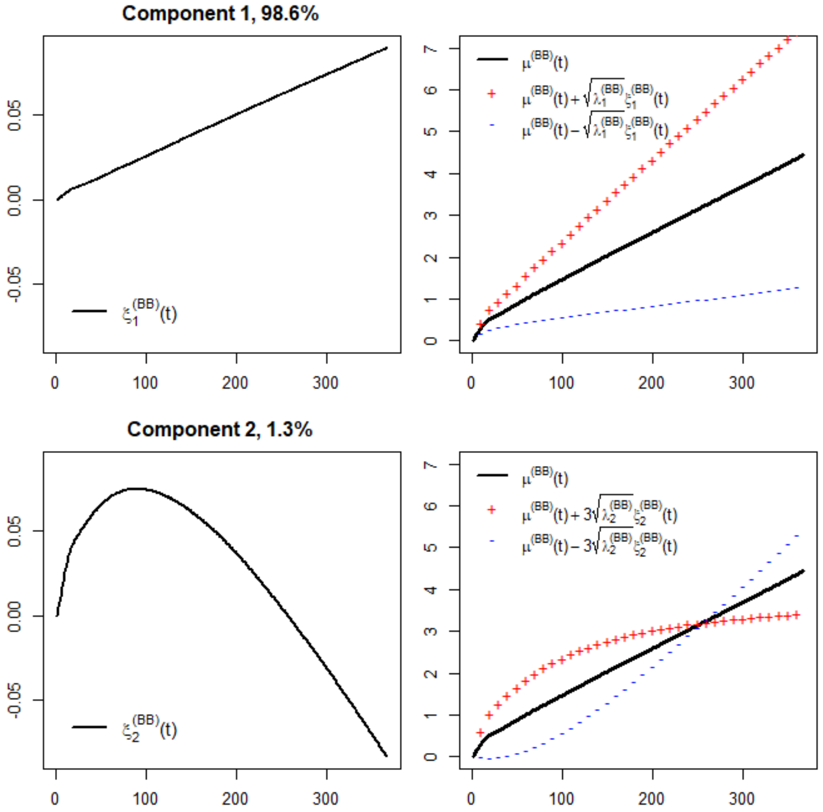
# FPCA on functional compensators



*Dynamic evolution of the events risk*

## FPCA

Components across the four time-varying processes have similar shapes



*Dimensionality reduction to summarise information emerging from the functional compensators to a finite set of covariates, while losing a minimum part of the information*

FPC I: **Different events risk**

↓

A patient with a high score is likely to experience more events than a patient with a low score.

FPC II: **Different events timing**

↓

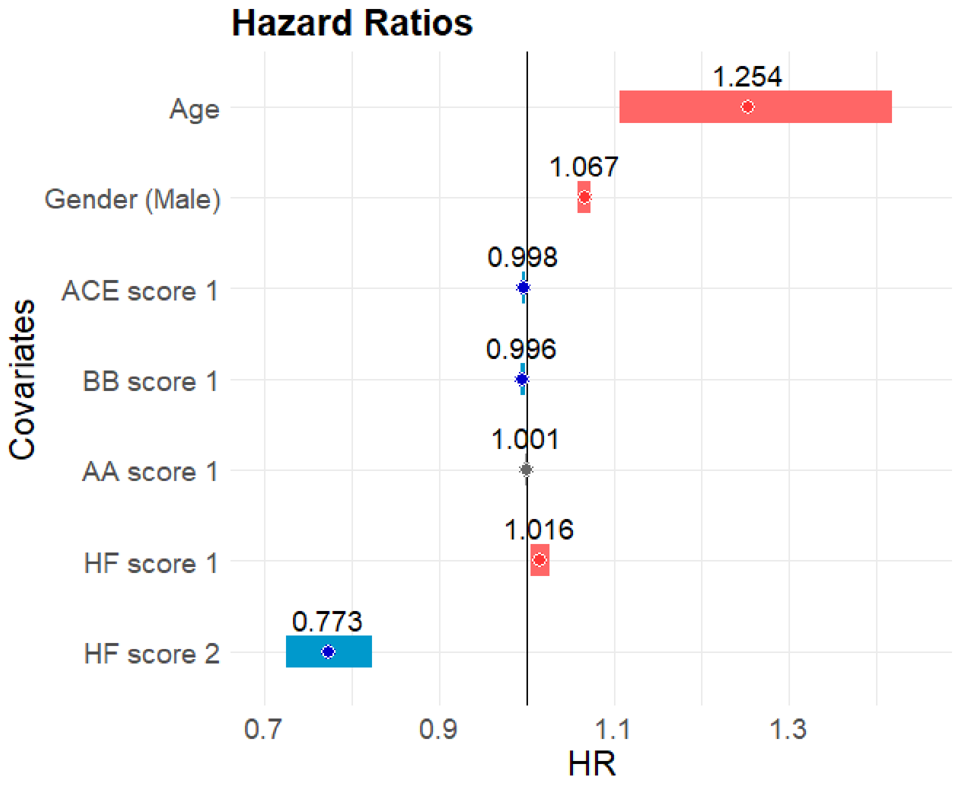
A patient with a high score is likely to experience more events in the first part of the year and less events in the last months of the year than a patient with a lower score.

# Multivariate Functional Linear Cox Regression Model for long-term survival

According to the highest median Concordance Index, the selected MFLCRM was the following:

$$\eta_i\left(t|\boldsymbol{\omega}_i,\left\{\Lambda_i^{(h)}\right\}_{h\in H}\right) = \eta_0^*(t)\exp\left\{\theta_1 age_i + \theta_2 gender_i + \alpha_1^{(ACE)} f_{i1}^{(ACE)} + \alpha_1^{(BB)} f_{i1}^{(BB)} + \alpha_1^{(AA)} f_{i1}^{(AA)} + \alpha_1^{(HF\,hosp)} f_{i1}^{(HF\,hosp)} + \alpha_2^{(HF\,hosp)} f_{i2}^{(HF\,hosp)}\right\}$$

M



<span style="color:red">Higher risk</span> of death for:

- elder patients (6% each year)
- male patients
- patients having experienced many hospitalizations

Lower risk of death for:

- patients assuming more ACE inhibitors
- patients assuming more BB agents
- patients who had many hospitalizations at the beginning of the year and few in the end correspond to the ones who have already experienced a critical phase of the disease and survived from it *(effect of the hospitalizations trend over time)*

# Block I – take home messages

- **Starting from the need for novel and tailored methodologies** capable of **extracting additional information** from Real-World Data (e.g., Administrative Data), our method is able to characterize the association between **time-varying** covariates and **time-to-event** data.

- **New** methodology based on **stochastic processes theory** and **Functional Data Analysis** able to effectively extract and resume information from functional data, intended as **trajectories of compensators representing recurrent events**.
  → *Marked Point Process formulation for Recurrent Events*

- Functional compensators contains information about different events **risk** and different events **timing**.
  → Highlight **trends and variations in the shape** of the processes over time

- One of the first attempts in literature of **merging potential of Functional Data Analysis** and **Survival Analysis**.

- Flexible methodology to quantify the **effect of personal behaviours and therapeutic patterns on survival.**
  → *New insights for personalized treatment*

PB: Observation period and immortal time bias

# Block II

Data sources: genomic/epigenomic data, SNPs, expression data

Methods: Deep Sparse Autoencoders – Network Theory – Itemset Rule Mining

# Research Question

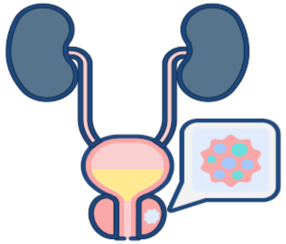
- Precision medicine framework often has the need to **model the relationship between some phenotypic trait or health outcome and one or more omics-based information sources**.
- However, irrespectively of the clinical inquiry, **raw genotype data (and -omics data, in general) naturally carry characteristics that hinder the applicability of most traditional statistical and biostatistical methods**.
- Indeed, traditional approaches often rely on strict assumptions (s.a. independence between predictors, linear and additive effect on the outcome, normally distributed predictors, etc) that are unrealistic to model the complexity of the genotype, and oftentimes suffer some practical facets of these information sources and of their real-world application settings.

- Need for development of methodologies that construct **effective biological system complexity-aware representations** to enhance and complement interpretable and robust statistical approaches to classification, regression or survival modeling

  => <u>map the input into informative and manageable spaces where complexities are resolved</u>

  => **tackle the complexity of genomic data** (unbalanceness, interactions, high dimensionality, computational scalability,…), extracting meaningful information (feature selection,)

# RadPrecise: personalize radiotherapy



- **Prostate cancer** is the most diffused cancer affecting the male population in Europe
- Complications (**toxicity side effects**) resulting **from radiotherapy** *in the long run* may arise, but **are very rare**

→ Traditional methods (**Normal Tissue Complication Probability Models, NTCP**) based on patients' phenotypic characteristics and treatment details **fail in stratifying** the treated population.

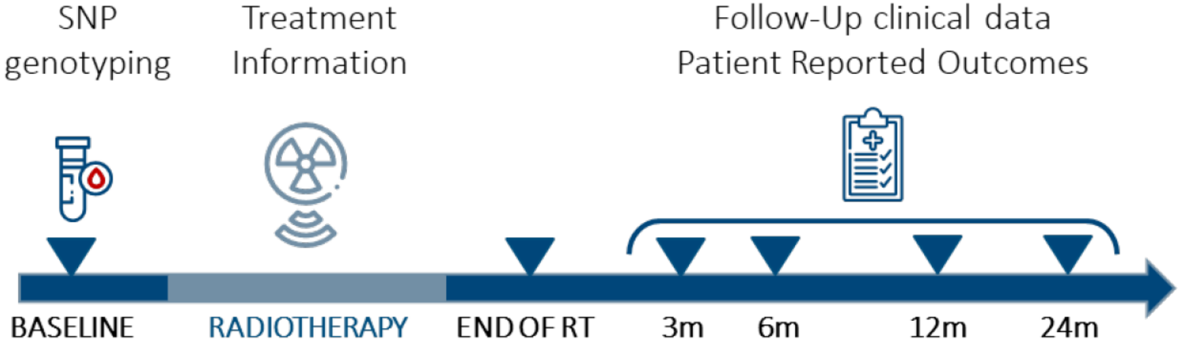
## DATA, COHORT & OUTCOMES

**1405 patients** were included

**43 SNPs** from literature

**5 endpoints:**

- rectal bleeding 11.7%,
- urinary frequency 4%,
- haematuria 5.5%,
- nocturia 7.8%,
- decreased urinary stream 17.1%.



SNP genotyping

Treatment Information

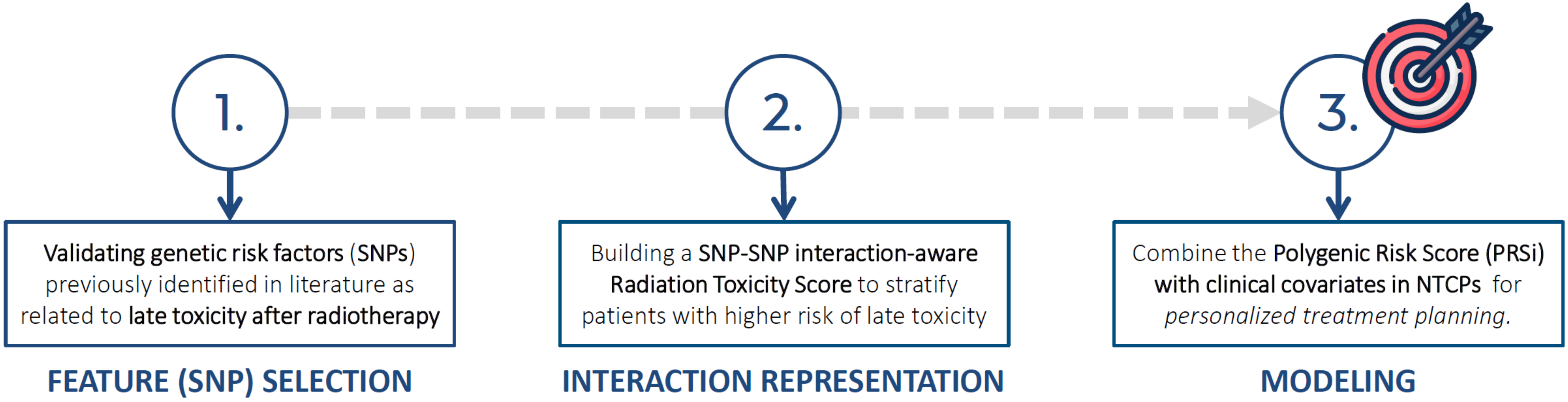
Follow-Up clinical data
Patient Reported Outcomes

REQUITE

# RadPrecise: personalize radiotherapy

→ Including genotype information may aid treatment outcome modeling and allow personalized treatment planning

Combined model to stratify patients and drive treatment decision-making



**Validating genetic risk factors** (**SNPs**) previously identified in literature as related to **late toxicity after radiotherapy**

**FEATURE (SNP) SELECTION**

Building a **SNP-SNP interaction-aware Radiation Toxicity Score** to stratify patients with higher risk of late toxicity

**INTERACTION REPRESENTATION**

Combine the **Polygenic Risk Score (PRSi) with clinical covariates in NTCPs** for *personalized treatment planning.*

**MODELING**

# RadPrecise: personalize radiotherapy

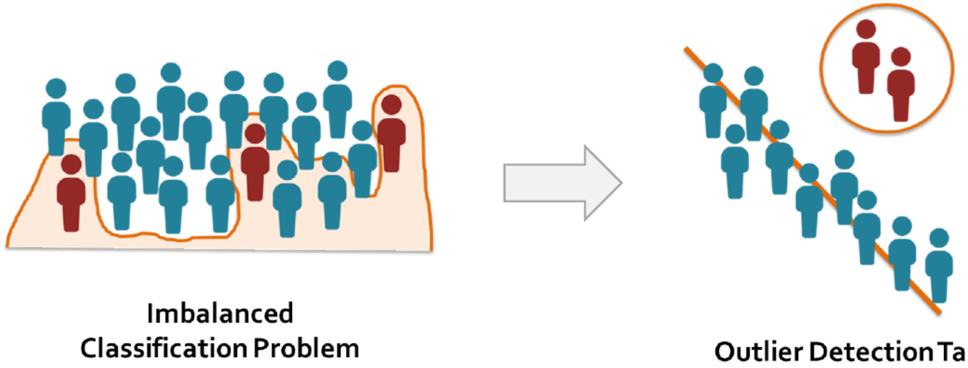
**1.**
Validating
genetic risk factors (SNPs)

## METHODOLOGICAL PROBLEM SETTING

- We seek to find differences in features (SNPs) between two **strongly imbalanced groups**, with a
- very **small minority class sample size**.
- The method need to be scalable to **very high dimensionalities**
- We want to consider **complex non-linear interactions** between SNPs
- Data can be **noisy** (imputed SNPs)

## OUR SOLUTION

*What characteristics (features) make the underrepresented population appear as an outlier of the overall population?*



Massi M.C., Gasperoni F., Ieva F. *et al.* (2020). A Deep Learning Approach Validates Genetic Risk Factors for Late Toxicity After Prostate Cancer Radiotherapy in a REQUITE Multi-National Cohort, *Frontiers in Oncology, Vol. 10 : 2033*

# RadPrecise: personalize radiotherapy

## AutoEncoders to ***characterize outliers***

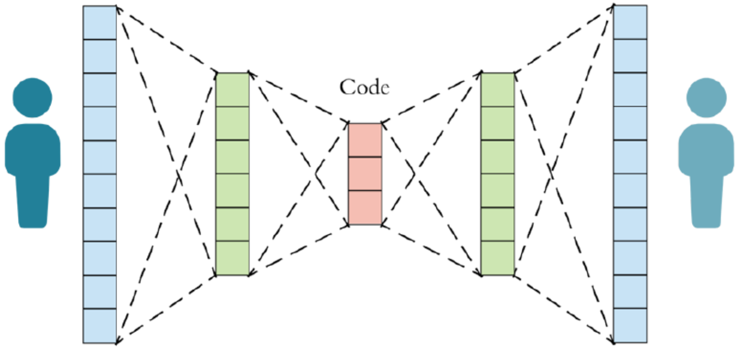
1. Train a **Deep Sparse Autoencoder** (DSAE) to learn how to reconstruct *majority class* observations.

   → *The learnt data distribution does not include the characterization aspects of minority class instances*

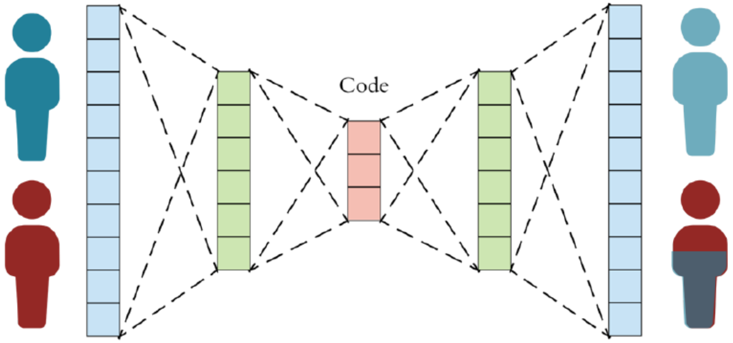
2. **Test** the model on *majority* and *minority* classes

3. The model is expected to make higher **Reconstruction Errors** (RE) on *anomalous observations* (minority class)

TRAINING PHASE



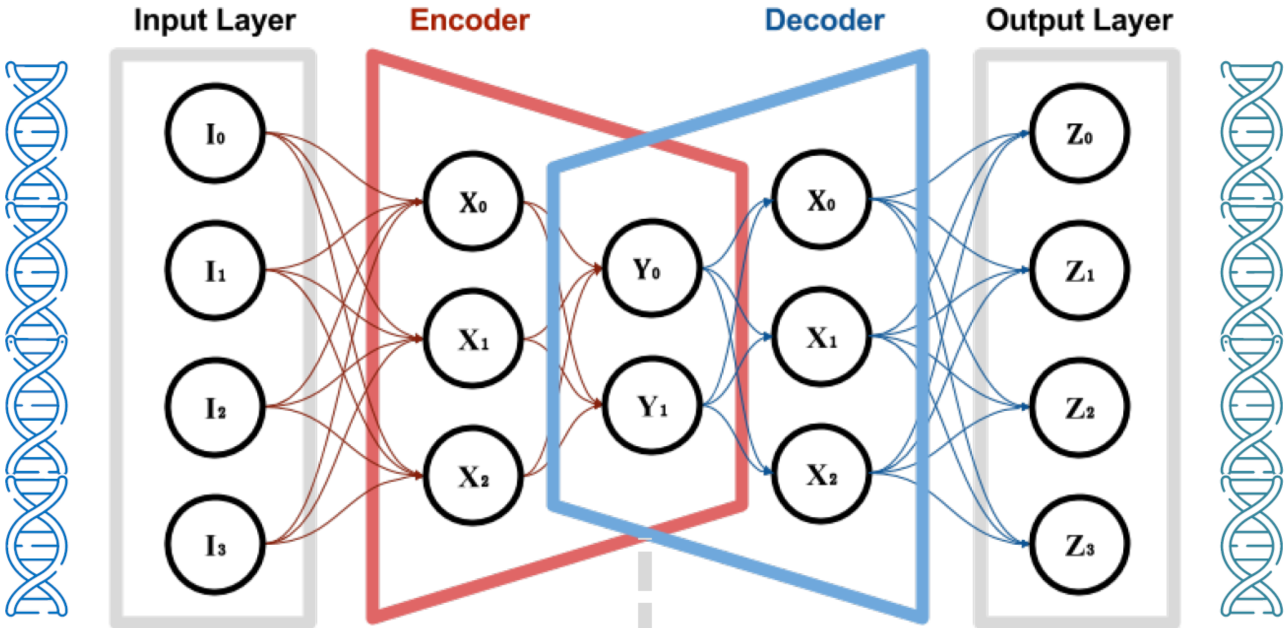
TEST PHASE



# RadPrecise: personalize radiotherapy

AutoEncoders (AE) are Neural Networks trained to reconstruct their input.

*They are powerful* ***non-linear dimensionality reduction models***



The **bottleneck layer** forces the model to learn a representation of the input that is reduced in dimensionality and informative enough to reconstruct the input precisely

Complex and non-linear mapping that models interrelationships between features

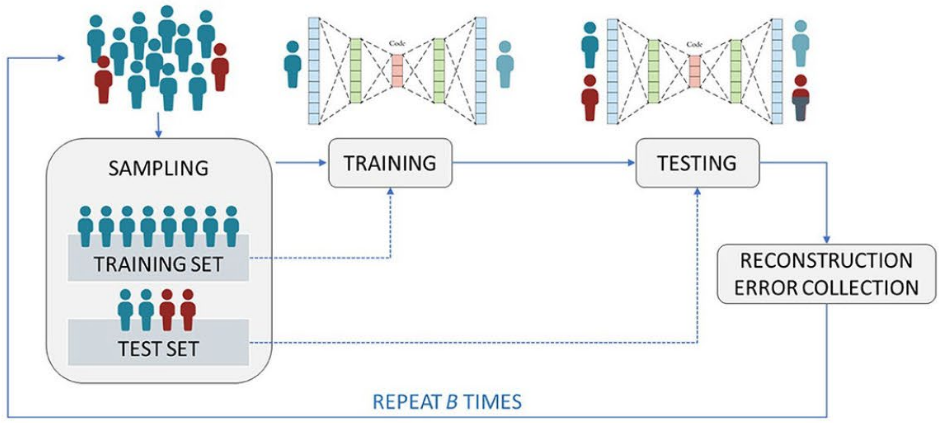
Learns the most relevant aspects of the input

Can be used for outlier detection…

…how?

# RadPrecise: personalize radiotherapy

## ENSEMBLE LEARNING



SAMPLING

TRAINING SET

TEST SET

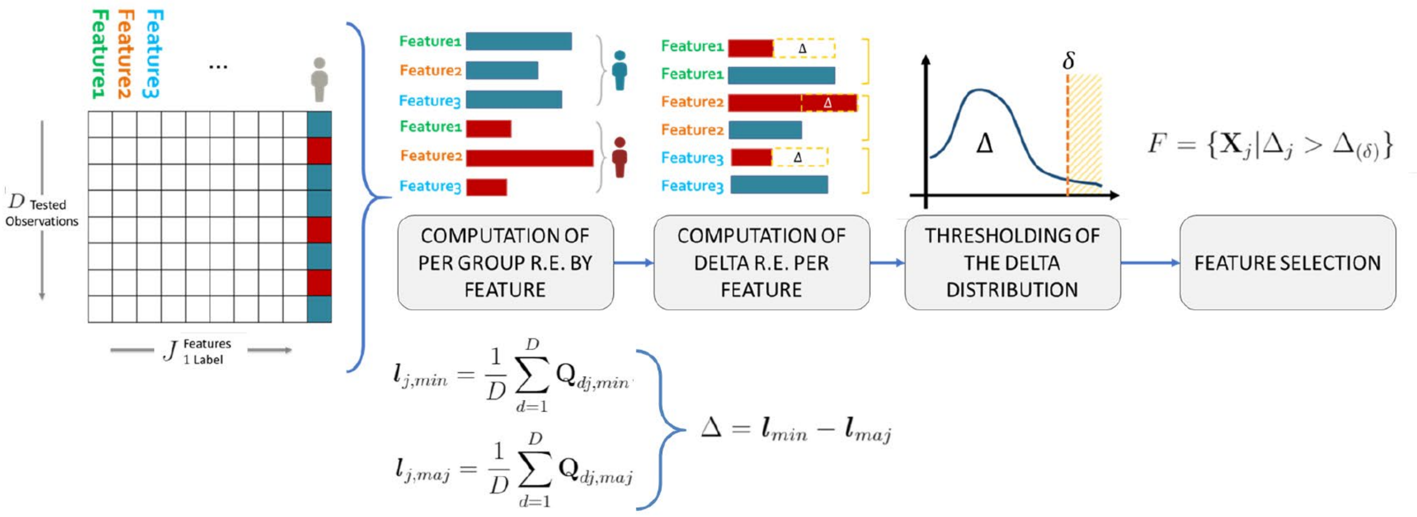
TRAINING

TESTING

RECONSTRUCTION ERROR COLLECTION

REPEAT *B* TIMES

## AGGREGATION AND FEATURE SELECTION



COMPUTATION OF PER GROUP R.E. BY FEATURE

COMPUTATION OF DELTA R.E. PER FEATURE

THRESHOLDING OF THE DELTA DISTRIBUTION

FEATURE SELECTION

$$F = \{\mathbf{X}_j | \Delta_j > \Delta_{(\delta)}\}$$

$$l_{j,min} = \frac{1}{D}\sum_{d=1}^{D} \mathbf{Q}_{dj,min}$$

$$l_{j,maj} = \frac{1}{D}\sum_{d=1}^{D} \mathbf{Q}_{dj,maj}$$

$$\Delta = l_{min} - l_{maj}$$

# RadPrecise: personalize radiotherapy

- Cohort of **1,296** patients
- 55 (**4.2%**) of which experiencing Late Toxicity
- **43 SNPs**
- **9 SNPs identified in literature**[a] for this endpoint

Table. Association between SNPs and toxicity endpoint when using logistic regressionon REQUITE cohort

| **ODDS RATIO**<br>[Kerns et al.] | **SNPs to validate**<br>DSAEE 85th quantile | **p-value**<br>**[REQUITE]** |
|---|---|---|
| **3,2** | rs7366282 | **0.05** |
| **3,12** | rs17599026 | 0.61 |
| **2,66** | rs10209697 | 0.86 |
| **2,41** | rs8098701 | 0.48 |
| 1,8 | rs10101158 | 0.70 |
| 1,74 | rs7356945 | 0.47 |
| 0,51 | rs342442 | 0.79 |
| 0,51 | rs6003982 | 0.63 |
| 0,49 | rs4997823 | 0.44 |
| TOTAL SELECTED | 7 | |
| TOTAL VALIDATED | 4 | |
| PERCENTAGE VAL/SEL | **57.14%** | |
| PERCENTAGE SEL/TOT | 16.28% | |

Table 2. in green SNPs selected by DSAEE

# RadPrecise: personalize radiotherapy

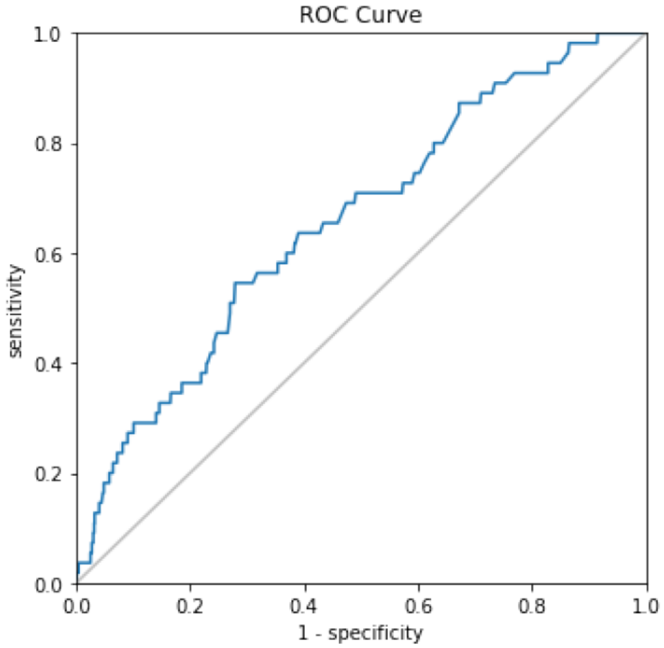
## 2.
Polygenic Risk Scoring

### PROBLEM

**The DSAE accounted for SNPs interactions to perform feature selection**, but we have **no direct access** to such information for later use.

Most relevant SNPs filtered by DSAEE

| $SNP_2$ | $SNP_5$ | $SNP_{10}$ | … | $SNP_{23}$ | $SNP_n$ | |
|---|---|---|---|---|---|---|
| 2 | 0 | 0 | 0 | 2 | 2 | 🟥 |
| 0 | 0 | 1 | 1 | 1 | 1 | 🟦 |
| 2 | 1 | 0 | 1 | 2 | 2 | 🟥 |
| 1 | 2 | 2 | 1 | 2 | 1 | 🟦 |
| 1 | 2 | 1 | 1 | 2 | 1 | 🟦 |
| 1 | 2 | 1 | 1 | 2 | 2 | 🟥 |



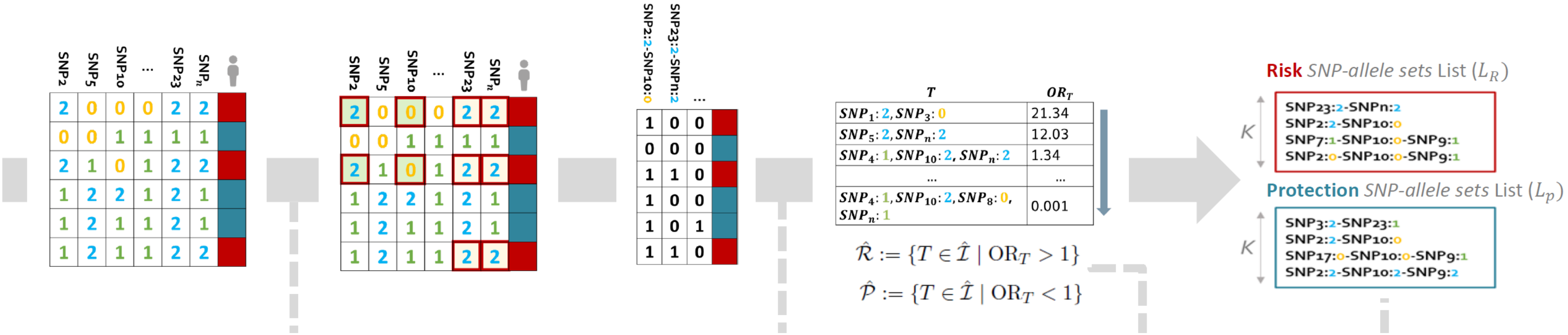
Area under the curve: 0.66

Statistics at optimality:

- Odds-ratio: 2.74
- Sensitivity: 63.64%
- Specificity: 61.00%
- Neg. predictivity: 97.43%
- Pos. predictivity: 6.74%

**Ignoring** interaction terms **results in** classifiers with **bad performances**

# RadPrecise: personalize radiotherapy



| SNP2 | SNP5 | SNP10 | … | SNP23 | SNP$_n$ | |
|---|---|---|---|---|---|---|
| 2 | 0 | 0 | 0 | 2 | 2 | 🟥 |
| 0 | 0 | 1 | 1 | 1 | 1 | 🟦 |
| 2 | 1 | 0 | 1 | 2 | 2 | 🟥 |
| 1 | 2 | 2 | 1 | 2 | 1 | 🟦 |
| 1 | 2 | 1 | 1 | 2 | 1 | 🟦 |
| 1 | 2 | 1 | 1 | 2 | 2 | 🟥 |

| SNP2 | SNP5 | SNP10 | … | SNP23 | SNP$_n$ | |
|---|---|---|---|---|---|---|
| 2 | 0 | 0 | 0 | 2 | 2 | 🟥 |
| 0 | 0 | 1 | 1 | 1 | 1 | 🟦 |
| 2 | 1 | 0 | 1 | 2 | 2 | 🟥 |
| 1 | 2 | 2 | 1 | 2 | 1 | 🟦 |
| 1 | 2 | 1 | 1 | 2 | 1 | 🟦 |
| 1 | 2 | 1 | 1 | 2 | 2 | 🟥 |

| SNP2:2-SNP10:0 | SNP23:2-SNPn:2 | … | |
|---|---|---|---|
| 1 | 0 | 0 | 🟥 |
| 0 | 0 | 0 | 🟦 |
| 1 | 1 | 0 | 🟥 |
| 1 | 0 | 0 | 🟦 |
| 1 | 0 | 1 | 🟦 |
| 1 | 1 | 0 | 🟥 |

| $T$ | $OR_T$ |
|---|---|
| $SNP_1: 2, SNP_3: 0$ | 21.34 |
| $SNP_5: 2, SNP_n: 2$ | 12.03 |
| $SNP_4: 1, SNP_{10}: 2, SNP_n: 2$ | 1.34 |
| … | … |
| $SNP_4: 1, SNP_{10}: 2, SNP_8: 0, SNP_n: 1$ | 0.001 |

$$\hat{\mathcal{R}} := \{T \in \hat{\mathcal{I}} \mid \mathrm{OR}_T > 1\}$$

$$\hat{\mathcal{P}} := \{T \in \hat{\mathcal{I}} \mid \mathrm{OR}_T < 1\}$$

**Risk** *SNP-allele sets* List ($L_R$)

SNP23:2-SNPn:2
SNP2:2-SNP10:0
SNP7:1-SNP10:0-SNP9:1
SNP2:0-SNP10:0-SNP9:1

**Protection** *SNP-allele sets* List ($L_p$)

SNP3:2-SNP23:1
SNP2:2-SNP10:0
SNP17:0-SNP10:0-SNP9:1
SNP2:2-SNP10:2-SNP9:2

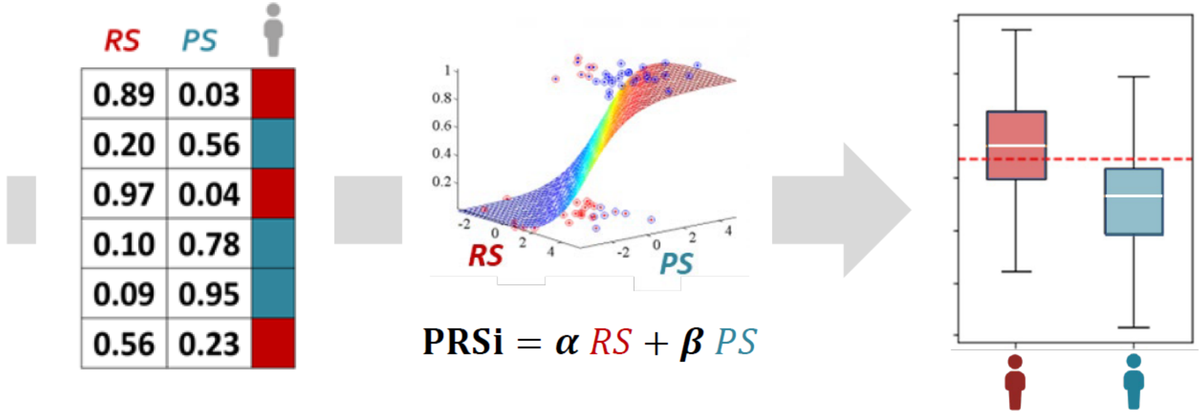
For each patient $\boldsymbol{i}$ we define the two scores $\boldsymbol{RS_i}$ and $\boldsymbol{PS_i}$, as the percentage of risk or protection SNP-sets in $\boldsymbol{x}_i$.

Fit a Logistic Model of the form:

$$log(\,\mathbb{P}(\,y = 1)\,)\; = \boldsymbol{\alpha}\mathbf{RS}\; + \boldsymbol{\beta}\mathbf{PS}\; + \gamma$$

Once obtained $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the combined Interaction-aware PRS is

$$\boldsymbol{PRSi} = \boldsymbol{\alpha}\mathbf{RS}\; + \boldsymbol{\beta}\mathbf{PS}$$

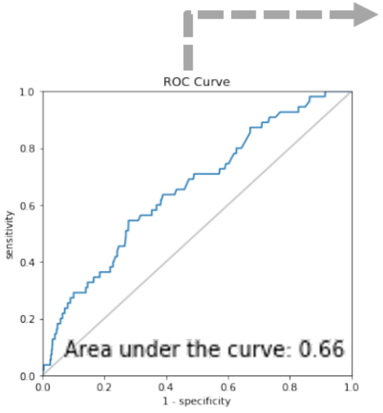


| RS | PS | |
|---|---|---|
| 0.89 | 0.03 | 🟥 |
| 0.20 | 0.56 | 🟦 |
| 0.97 | 0.04 | 🟥 |
| 0.10 | 0.78 | 🟦 |
| 0.09 | 0.95 | 🟦 |
| 0.56 | 0.23 | 🟥 |

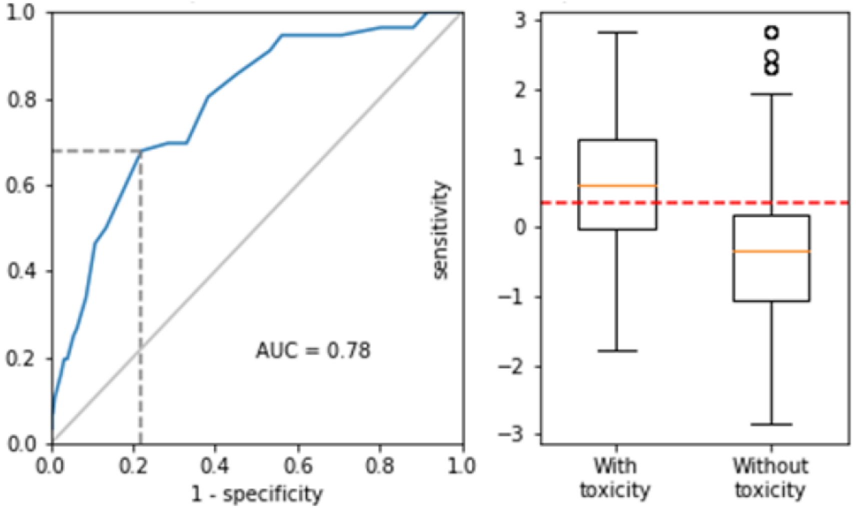
$$\mathbf{PRSi} = \boldsymbol{\alpha}\; RS + \boldsymbol{\beta}\; PS$$

Franco N.R., Massi M.C., Ieva F. et al. (2021) **Development of a method for generating SNP interaction-aware polygenic risk scores for radiotherapy toxicity**, Radiotherapy and Oncology

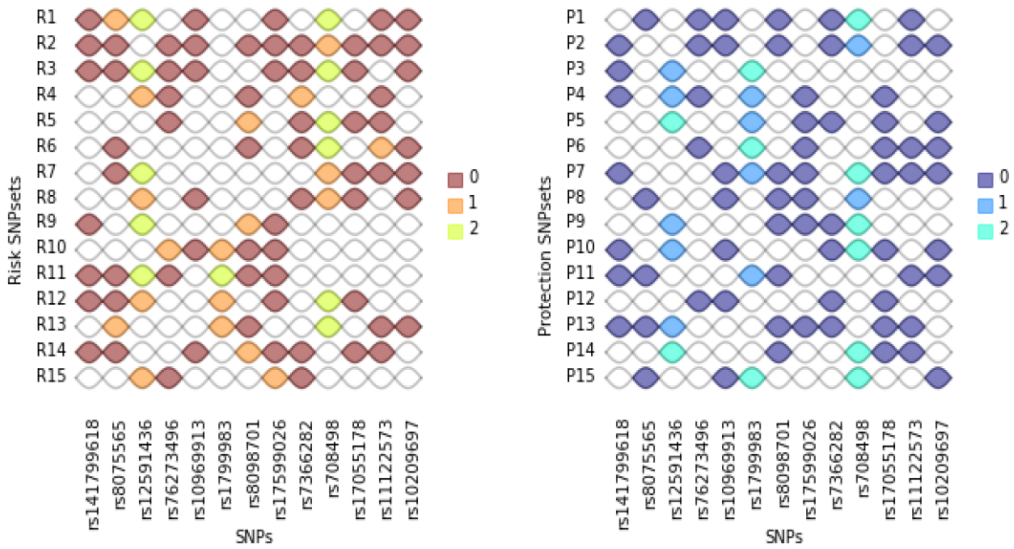
# RadPrecise: personalize radiotherapy



## PRSi classification performance



## Risk and Protection SNP-sets



- Distributions of PRSi differed significantly in patients with/without toxicity with AUCs ranging from 0.61 to 0.78.
- PRSi performed **better than the classical Polygenic Risk Score** based on SNPs additive effect
- **Readable and interpretable list of predictive interactions**

# RadPrecise: personalize radiotherapy

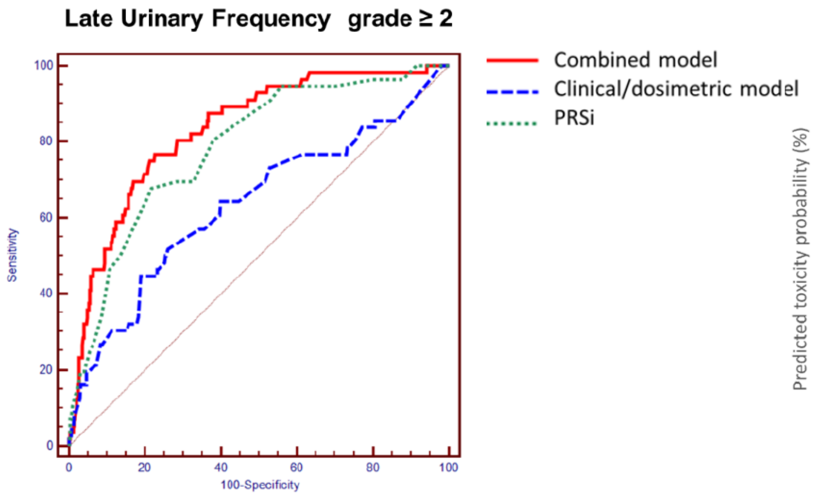
## 3. Combined Modeling

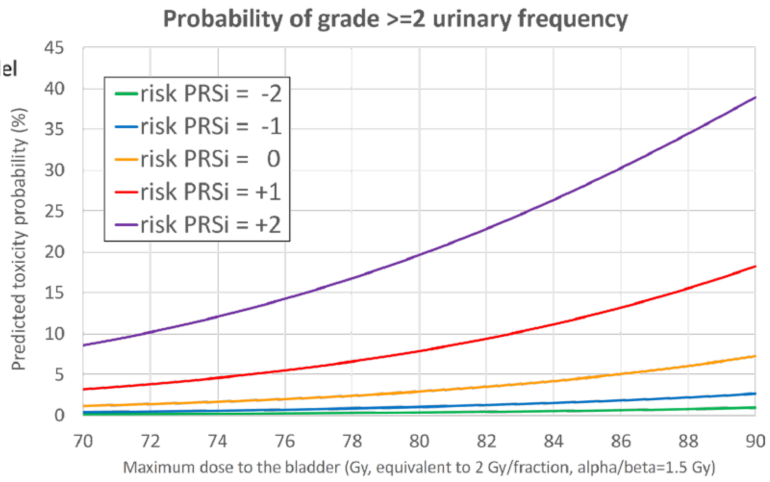
Combined NTCP model with **PRSi** and **clinical/dosimetric data**

→ Evaluation of added value of the **genetic information**

## RESULTS

| Late urinary frequency grade ≥ 2 | OR clinical/dosimetric | OR PRSi | OR combined model |
|---|---|---|---|
| Bladder maximum dose alpha/beta=1.5 Gy (1 Gy increase) | 1.09 | | 1.1 |
| baseline urinary frequency symptoms (no symptoms vs mild) | 2.5 | | 2.7 |
| diabetes | 1.65 | | 1.8 |
| prostatectomy | 2.1 | | 2.3 |
| Polygenic risk score with SNP-allele interaction (PRSi) | | 2.7 | 2.9 |
| *AUC* | *0.64* | *0.78* | *0.83* |





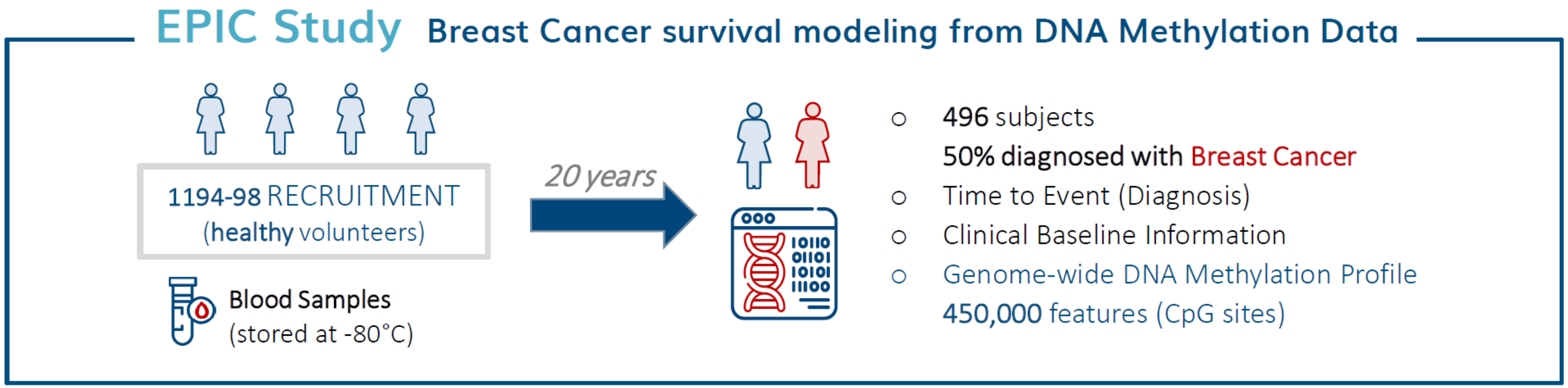
## CONCLUSIONS

Toxicity probability depends on

→ **PRSi**, i.e. genetic background of the patient: can't be changed, should be acknowledged
→ **Maximum dose to the bladder**: *this could be optimized for personalized treatment*

T. Rancati, M. Massi, et al (2021). "**PH-0656 Prediction of toxicity after prostate cancer RT: the value of a SNP-interaction polygenic risk score**", Radiotherapy and Oncology

# EPIC: non-invasive prediction of cancer development



## EPIC Study Breast Cancer survival modeling from DNA Methylation Data

1194-98 RECRUITMENT (healthy volunteers)

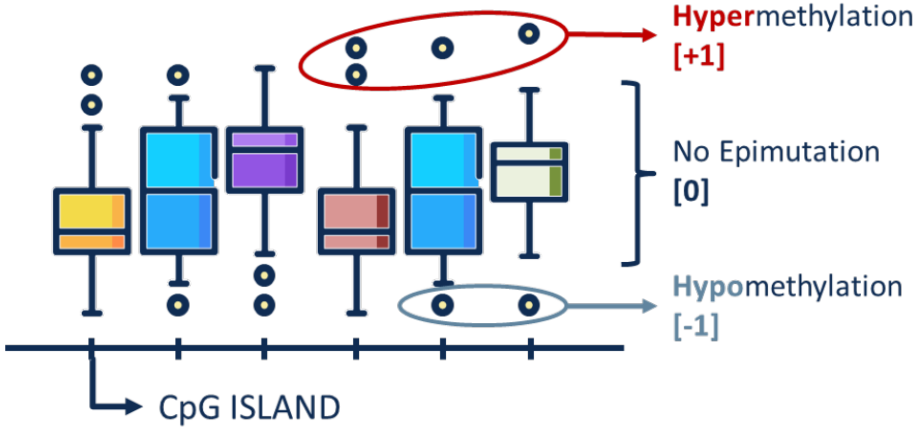
Blood Samples (stored at -80°C)

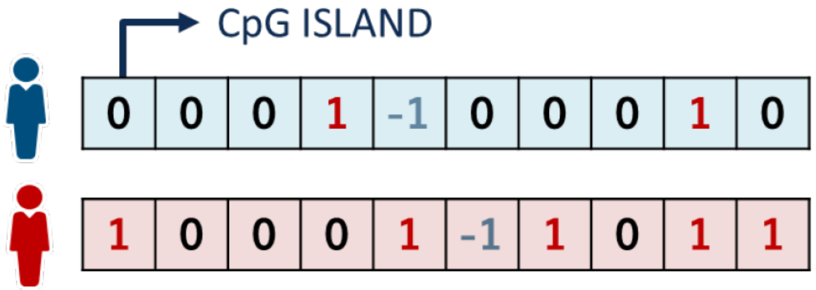
- 496 subjects
  50% diagnosed with Breast Cancer
- Time to Event (Diagnosis)
- Clinical Baseline Information
- Genome-wide DNA Methylation Profile
  450,000 features (CpG sites)

DNA methylation data (DNAm) can be codified as:

A. Continuous β values [0,1]

B. Categorical epimutation data





| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | -1 | 0 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 | 1 | -1 | 1 | 0 | 1 | 1 |

# EPIC: non-invasive prediction of cancer development

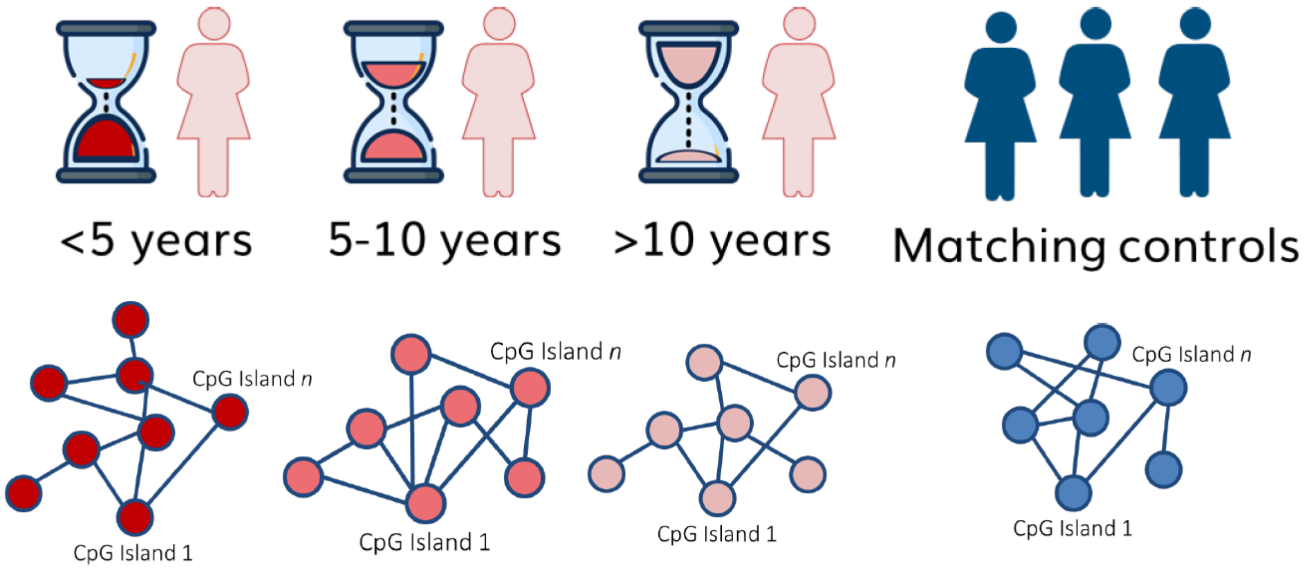
*What differs* in terms of **DNA methylation patterns** between patients with an early or a late cancer diagnosis w.r.t. the healthy control population?

Huge **dimensionality**, noisy **categorical** data, **sparse** information

**Failure of traditional Survival Models** with or without regularization

Breast Cancer Cohort

Co-Occurrence Networks



<5 years

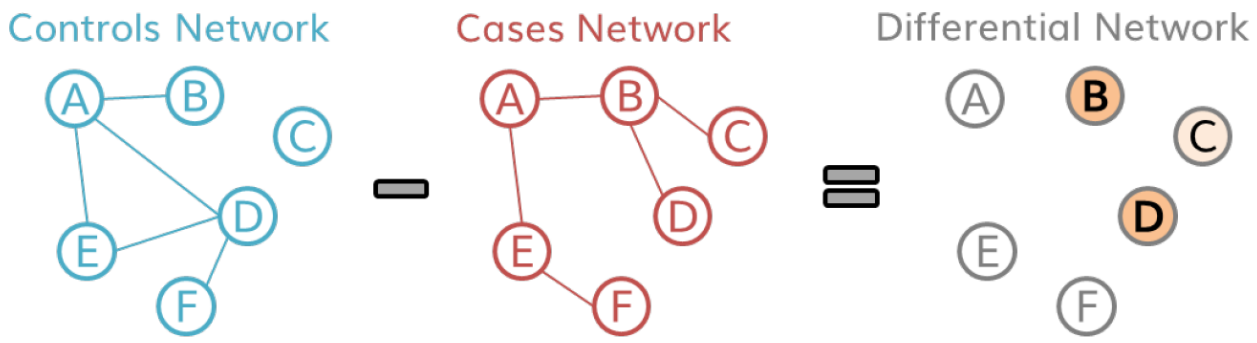
5-10 years

>10 years

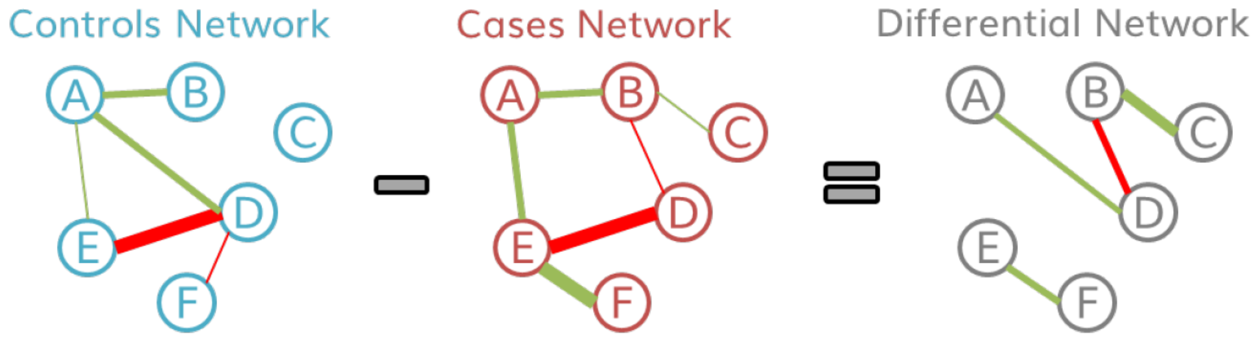
Matching controls

***Differential Network Analysis***

# EPIC: non-invasive prediction of cancer development



Comparing **network topologies**
(i.e. degrees of vertices, modularities, network flow, etc.)

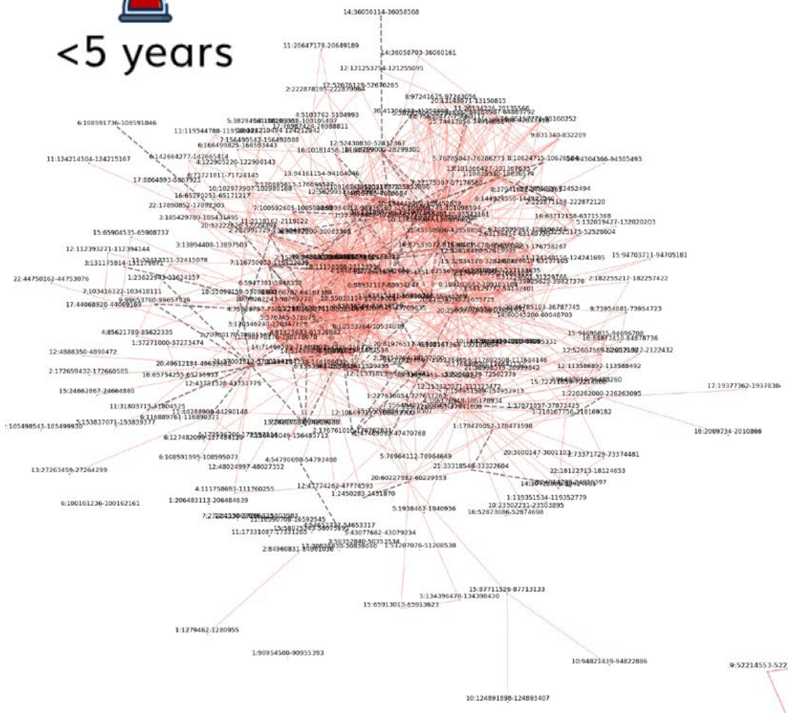


Comparing **weighted group-specific networks** for **edge-specific weight differences**

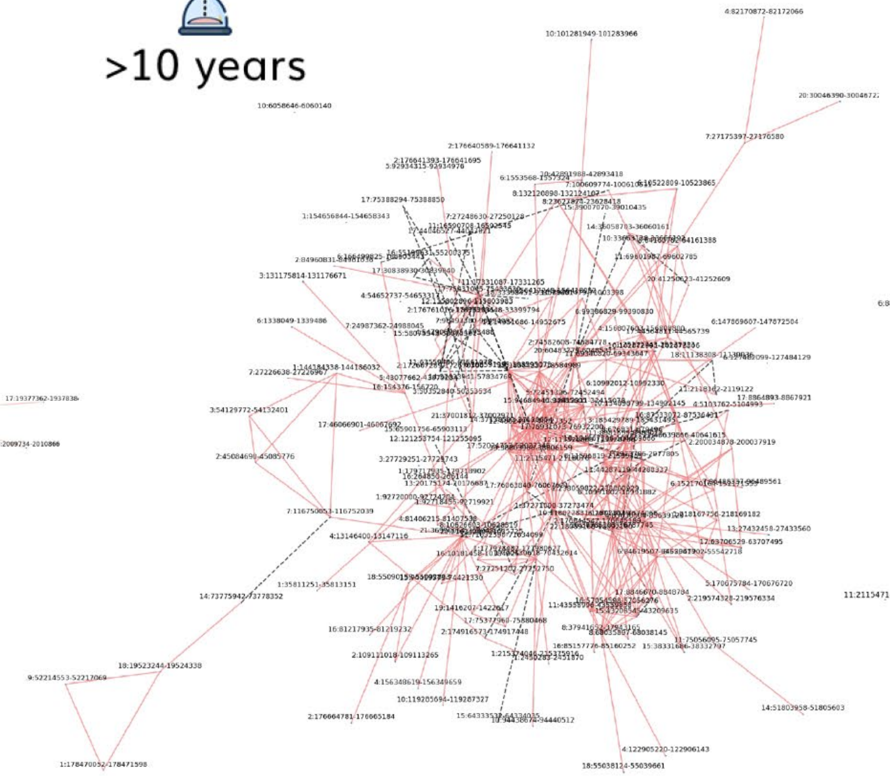
# EPIC: non-invasive prediction of cancer development

*What differs* in terms of **DNA methylation patterns** between patients with an **early** or a late cancer diagnosis **w.r.t. the healthy control** population?

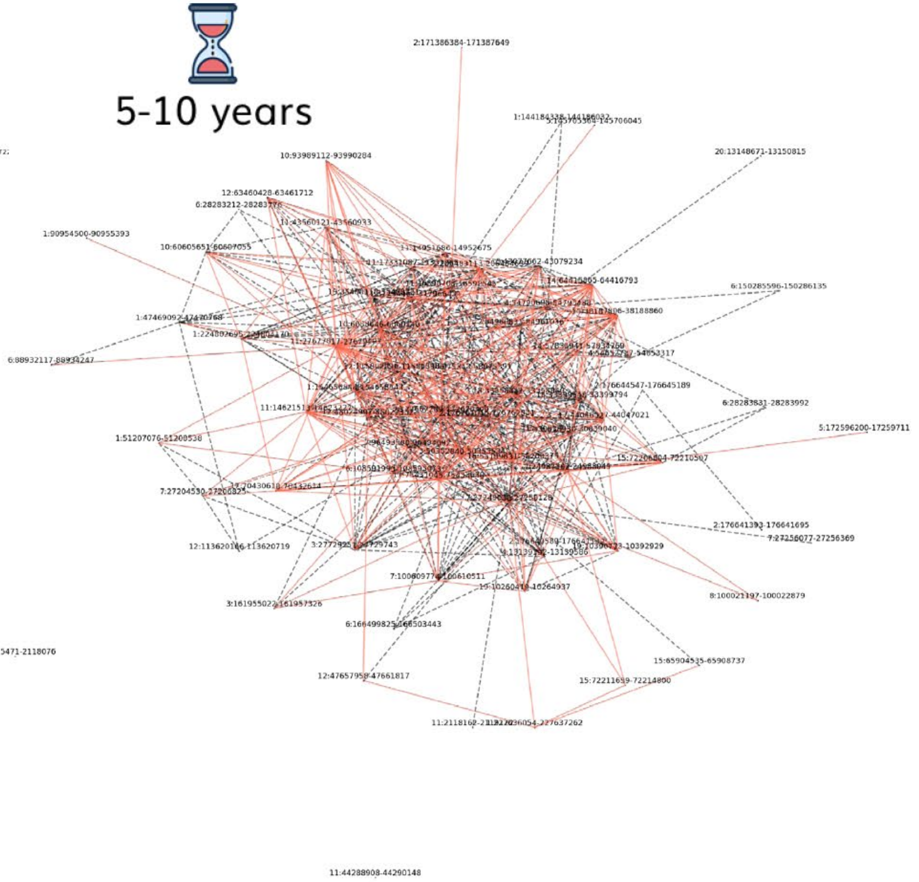
<5 years



>10 years



5-10 years



# EPIC: non-invasive prediction of cancer development

Survival data is comprised of three elements: a patient's baseline data *x*, a failure event time *T*, and an event indicator *E*.
The **hazard function** is the probability an individual will not survive beyond *t*, given they have already survived up to time *t*.

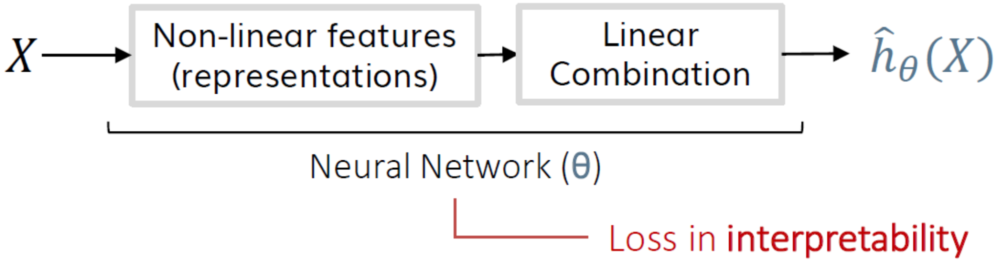
## Cox Proportional Hazards model (CoxPH)

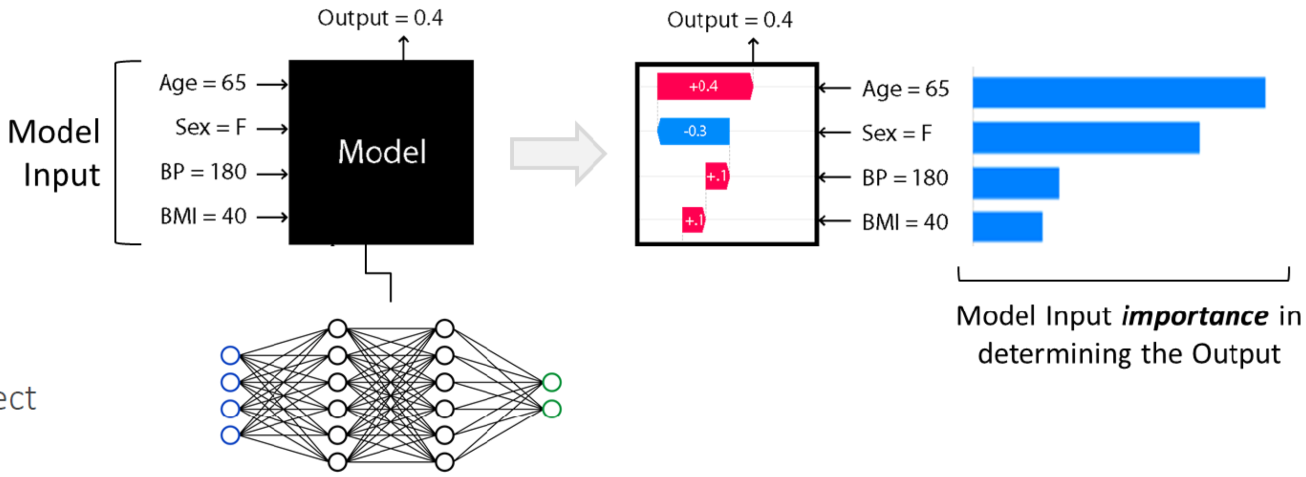
$$\lambda(t|X) = \underbrace{\lambda_0(t)}_{\textit{Baseline hazard}} \cdot e^{\overbrace{h(X)}^{\textit{Log-risk}}} = \lambda_0(t) \cdot e^{(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_M x_M)}$$

Easily **interpretable** model

We cannot assume the data as-is satisfies the linear proportional hazards condition
→ We should include **high-order interaction** terms

## Non-linear Survival Analysis


$X$
Non-linear features (representations)
Linear Combination
$\hat{h}_\theta(X)$
Neural Network (θ)
Loss in **interpretability**

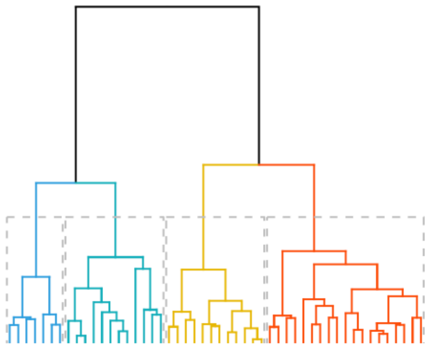


Model Input ***importance*** in determining the Output

We can exploit powerful non-linear models and then trace back the effect of each input thanks to ***explanation methods***.

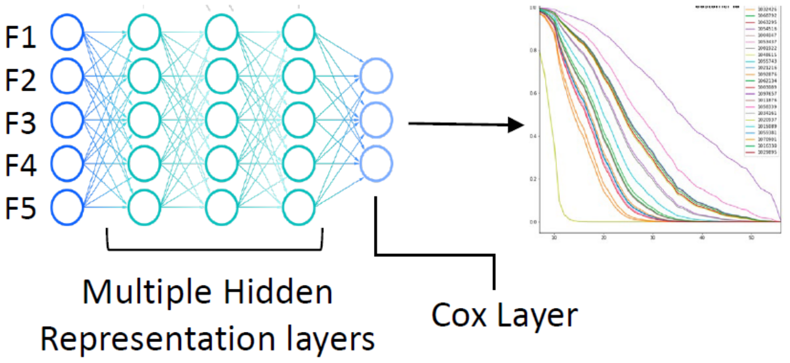
# EPIC: non-invasive prediction of cancer development

### Feature Aggregation



CpG Islands

### Deep (non-linear) Survival Modeling

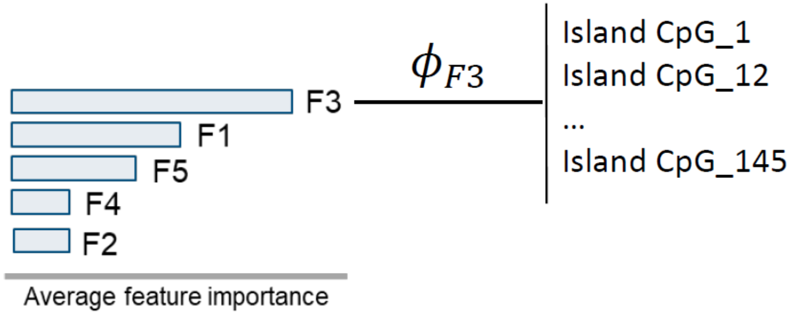


F1
F2
F3
F4
F5

Multiple Hidden Representation layers

Cox Layer

### CpG Islands Importance Ranking



F3
F1
F5
F4
F2

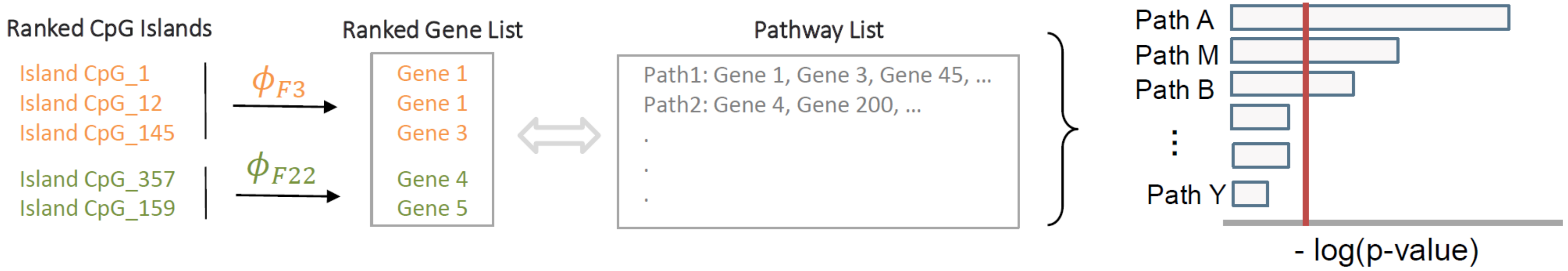
Average feature importance

$\phi_{F3}$

Island CpG_1
Island CpG_12
…
Island CpG_145

Shapley Values $\phi_j$

### Validation: Weighted Pathway Enrichment Analysis



Ranked CpG Islands

Island CpG_1
Island CpG_12
Island CpG_145

Island CpG_357
Island CpG_159

$\phi_{F3}$

$\phi_{F22}$

Ranked Gene List

Gene 1
Gene 1
Gene 3

Gene 4
Gene 5

Pathway List

Path1: Gene 1, Gene 3, Gene 45, …
Path2: Gene 4, Gene 200, …
.
.
.

Path A
Path M
Path B
⋮
Path Y

- log(p-value)

# Block II – take home messages

- Through ML and proper representations of the input data we can account for, and alleviate, data and context-specific complexities, overcoming the limitations of the traditional approaches to several precision medicine-oriented analyses of biological and medical data.

- Exploiting a Deep Representation Learning (RL) model as abuilding block of our ensemble algorithm allows to **model the complex non-linear interactions between all genetic features together and their relationship with the phenotype while performing feature selection**, accounting for high-order interaction between SNPs.

- **Co-Occurrence Network-based algorithm for categorical and extremely sparse genotype data**, tailored to deal with imbalanced settings such as studies seeking rare variants' association with Extreme Phenotypes.

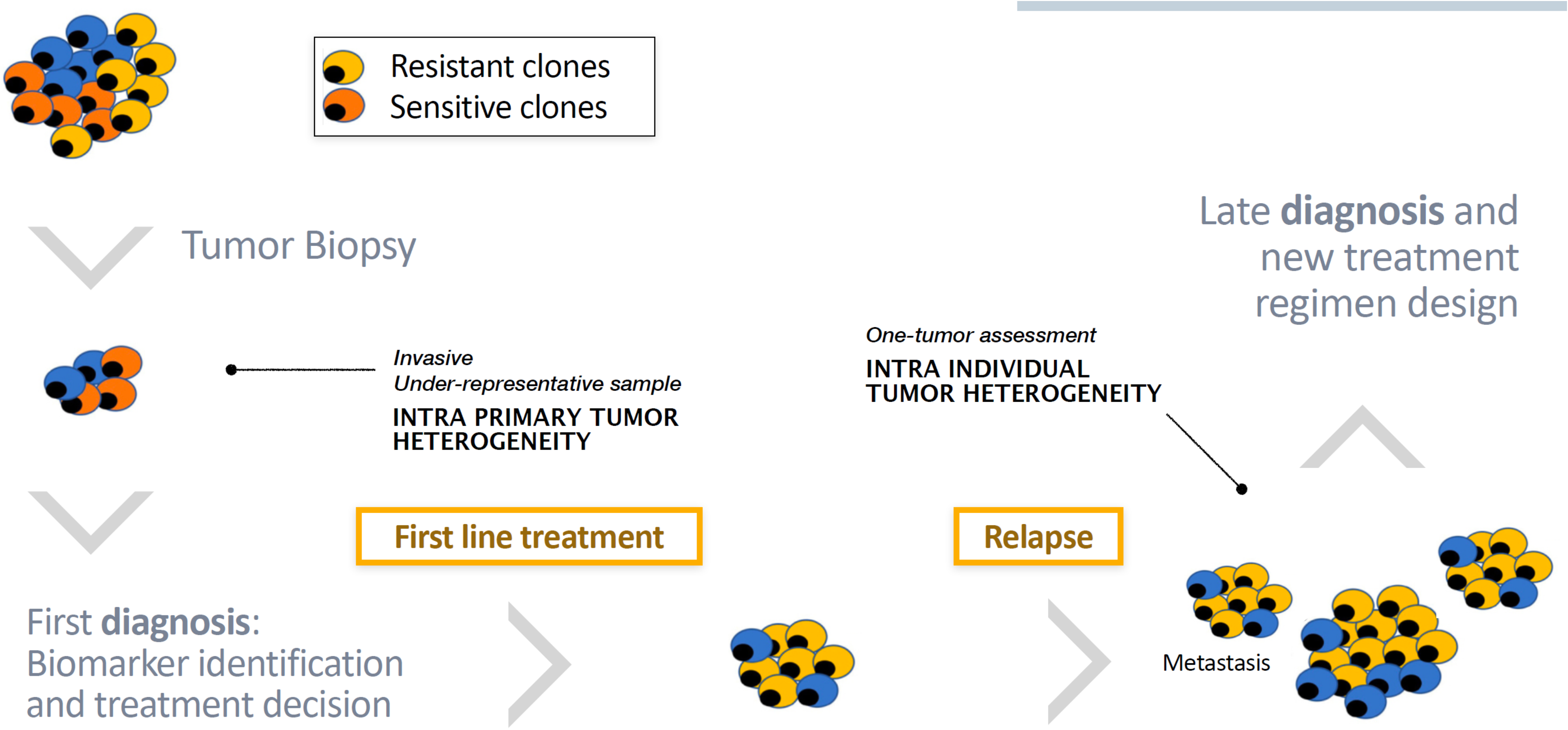
- Several of the methods we presented have the ability to manage data sources that are different in nature, i.e. omics but also unstructured medical data in general. Indeed, by picking the right tool to represent each data type-specific view, and by finding the best way to combine them, we will aim at building truly 360 degree Patient Representations, that have the potential to being informative and effective in dealing with all the facets of the complex system of biological and clinical information each of those patients embodies.

# Block III

Data sources: Medical Imaging

Methods: Trees – Concolutional Neural Netwotks – Depth Measures – Penalized Regressions – Survival Clustering

# The standard scenario



Resistant clones

Sensitive clones

Tumor Biopsy

*Invasive*
*Under-representative sample*

**INTRA PRIMARY TUMOR HETEROGENEITY**

First **diagnosis**:
Biomarker identification and treatment decision

First line treatment

Relapse

*One-tumor assessment*

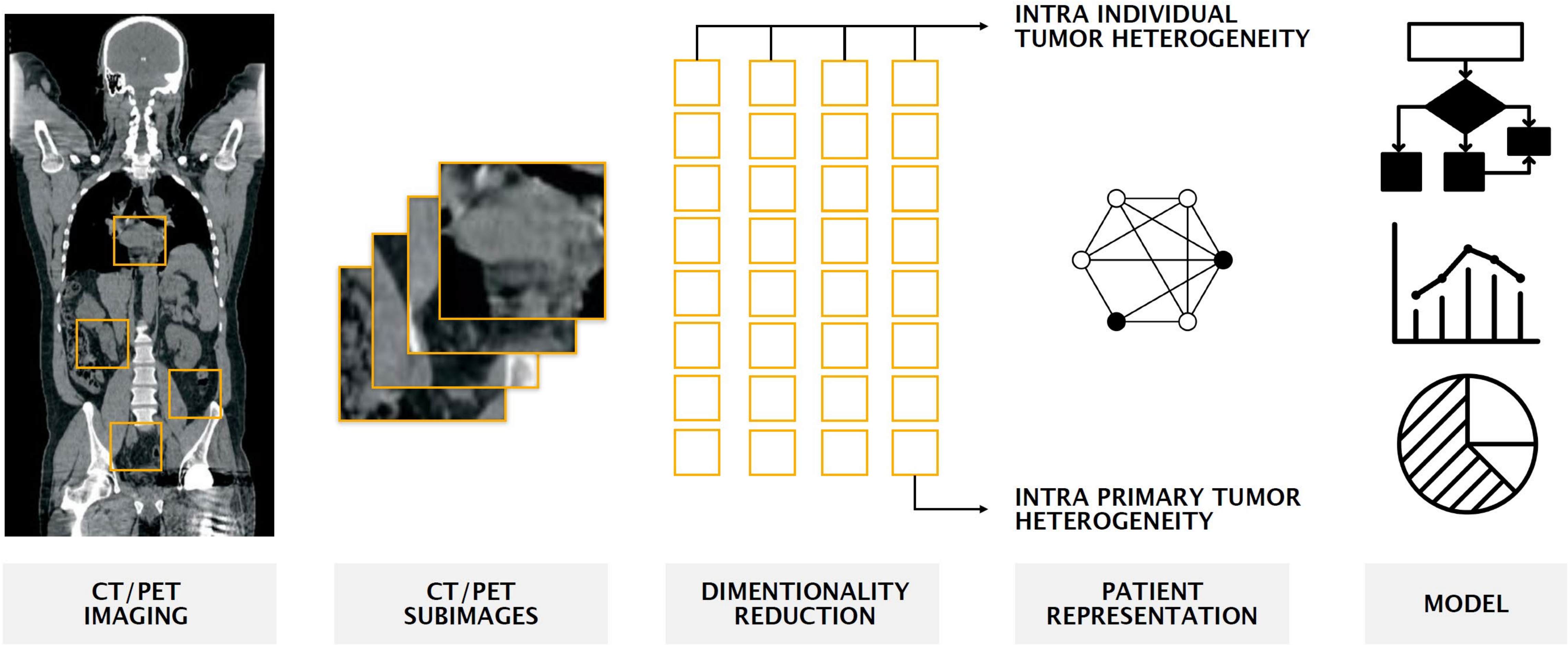
**INTRA INDIVIDUAL TUMOR HETEROGENEITY**

Metastasis

Late **diagnosis** and new treatment regimen design

# The precision medicine scenario: Virtual Biopsy

## Tumor **Virtual** Biopsy



INTRA INDIVIDUAL TUMOR HETEROGENEITY

INTRA PRIMARY TUMOR HETEROGENEITY

CT/PET IMAGING

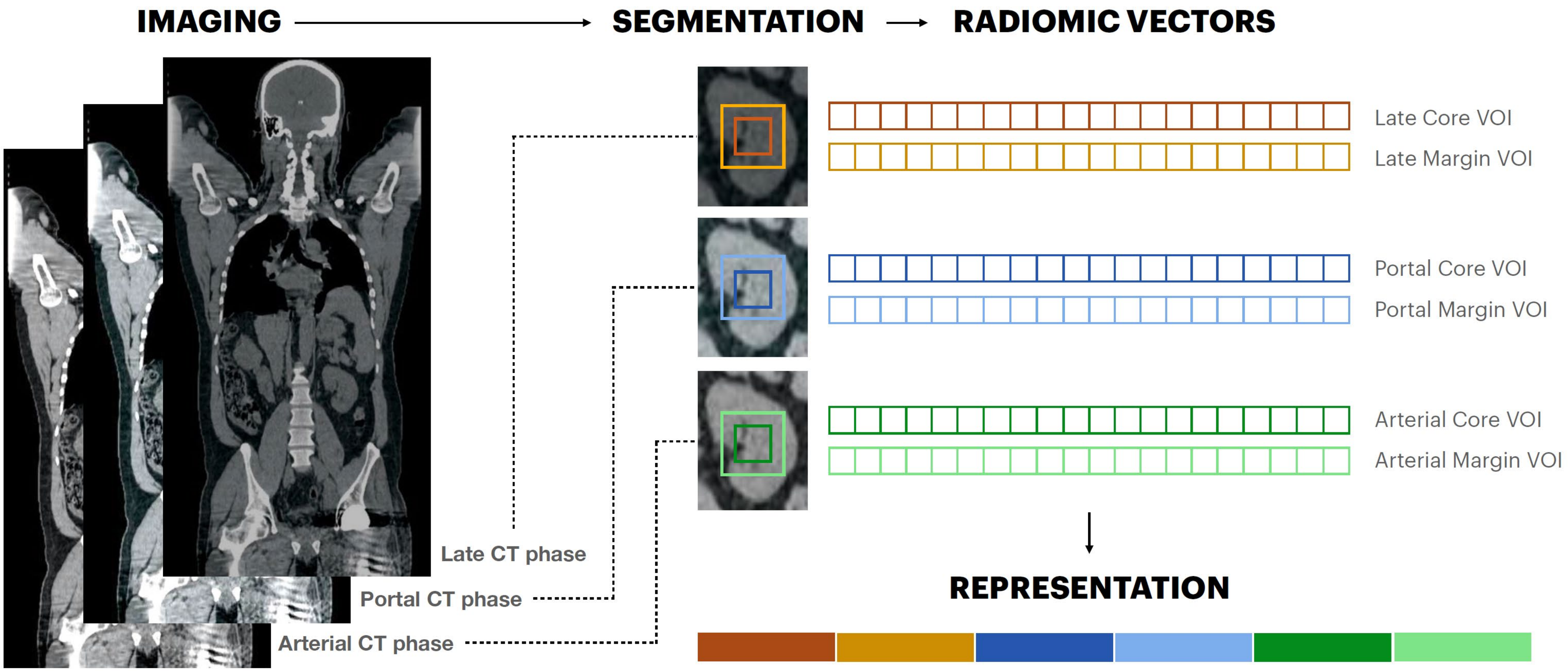
CT/PET SUBIMAGES

DIMENTIONALITY REDUCTION

PATIENT REPRESENTATION

MODEL

# The precision medicine scenario: Virtual Biopsy



# Research Questions

- How can we summarize the complex multi-view information about the patient?

=> Representation issue

- Can radiomic be of added value in predicting pathology evolution and survival response in IHC patients?
- Which radiomics information are the most informative?

=> Dimensionality reduction issue

- Reliable identification of prognostic factors and cohort stratification criteria
- Cancer subtyping

=> Explainability issue

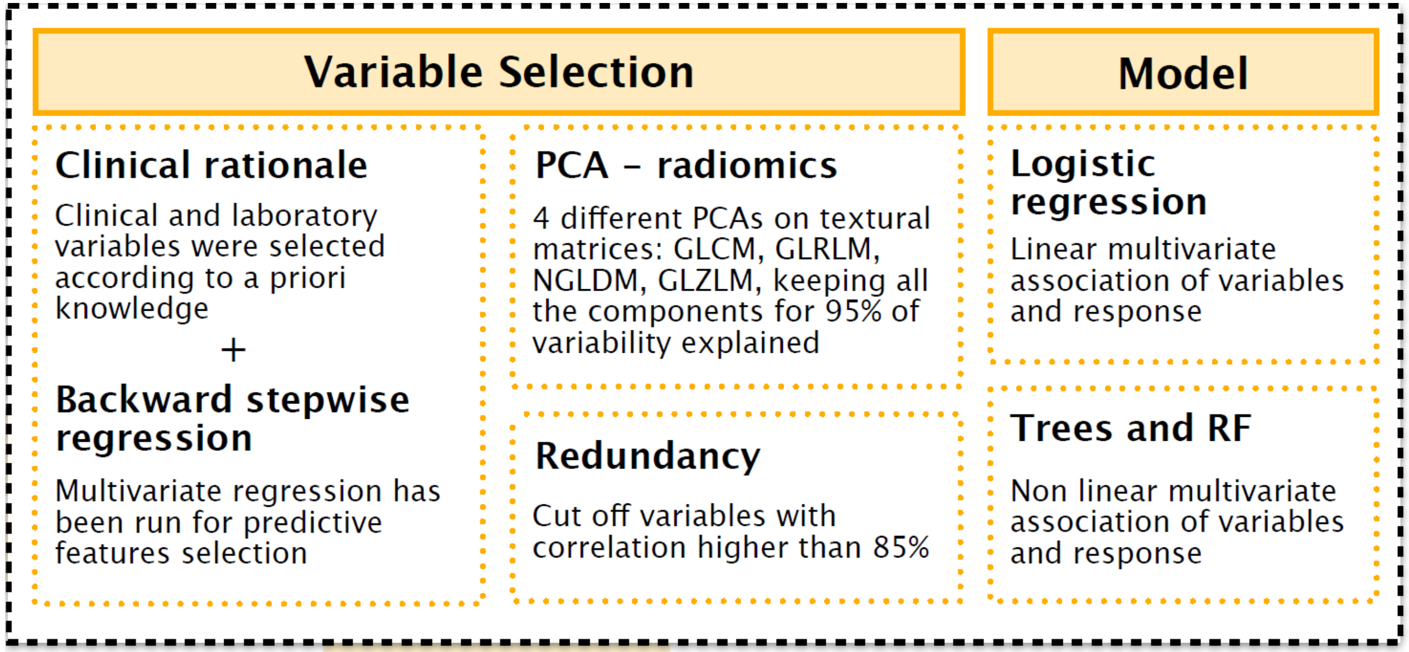
- Assessment of the role of core vs margin information
- Assessment of the information content of the different phases of the CT scan
- Are there any differences between centers?

=> Transfearability issue

# Hodgkin Lymphoma

Sollini, M., Bartoli, F., Cavinato, L., Ieva, F., Ragni, A., Marciano, A., Zanca, R., Galli, L., Paiar, F., Pasqualetti, F., Erba, P.A. (2021)
[18F]FMCH PET/CT biomarkers and similarity analysis to refine the definition of oligometastatic prostate cancer.
EJNMMI Research, Nov 27; 11(1): 119 PMID: 34837532

Sollini, M., Kirienko, M., Cavinato, L., Ricci, F., Biroli, M., Ieva, F., Calderoni, L., Tabacchi, E., Nanni, C., Zinzani, P.L., Fanti, S., Guidetti, A., Alessi, A., Corradini, P., Seregni, E., Carlo-Stella, C., Chiti, A. (2020)
Methodological framework for radiomics applications in Hodgkin's Lymphoma
European Journal of Hybrid Imaging. 4: 1-17 .



**Variable Selection**

**Clinical rationale**

Clinical and laboratory variables were selected according to a priori knowledge

+

**Backward stepwise regression**

Multivariate regression has been run for predictive features selection

**PCA – radiomics**

4 different PCAs on textural matrices: GLCM, GLRLM, NGLDM, GLZLM, keeping all the components for 95% of variability explained

**Redundancy**

Cut off variables with correlation higher than 85%

**Model**

**Logistic regression**

Linear multivariate association of variables and response

**Trees and RF**

Non linear multivariate association of variables and response

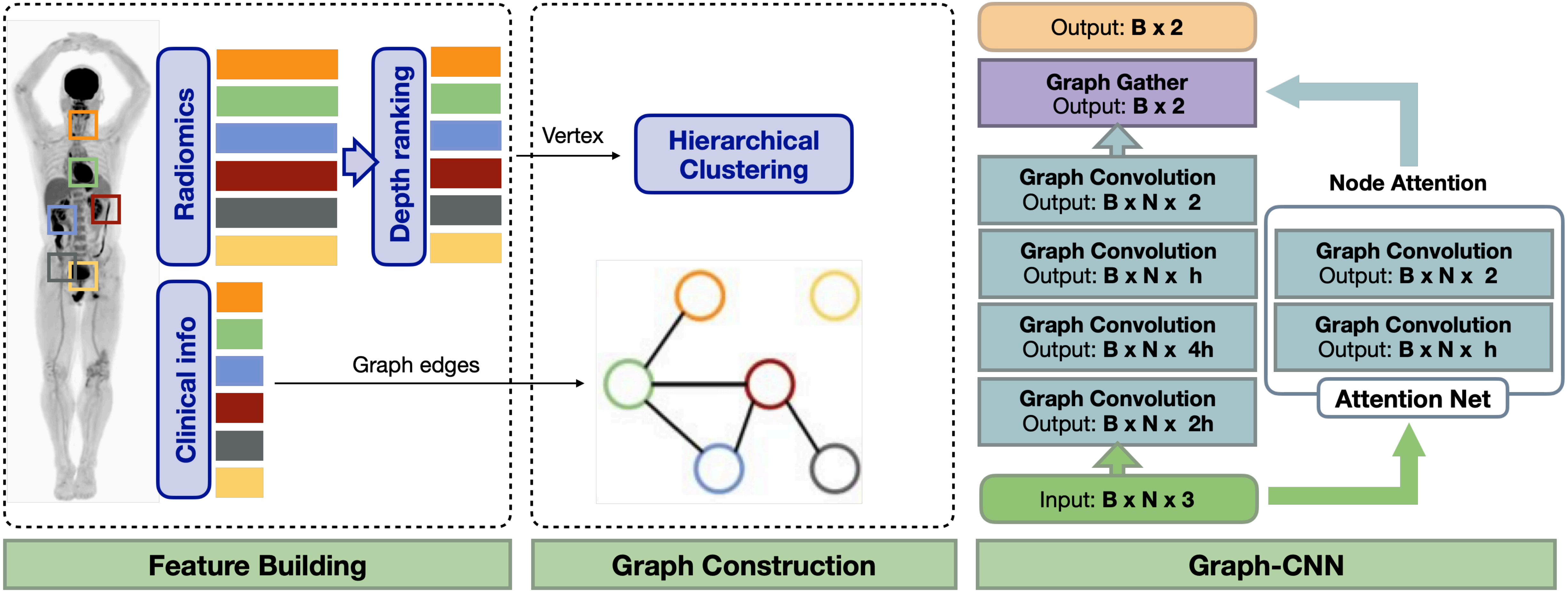
## Methodological contribution

Rigorous feature selection framework for healthy tissue

Relevance of imaging information as prognostic factor (wrt only clinical prognosticators)

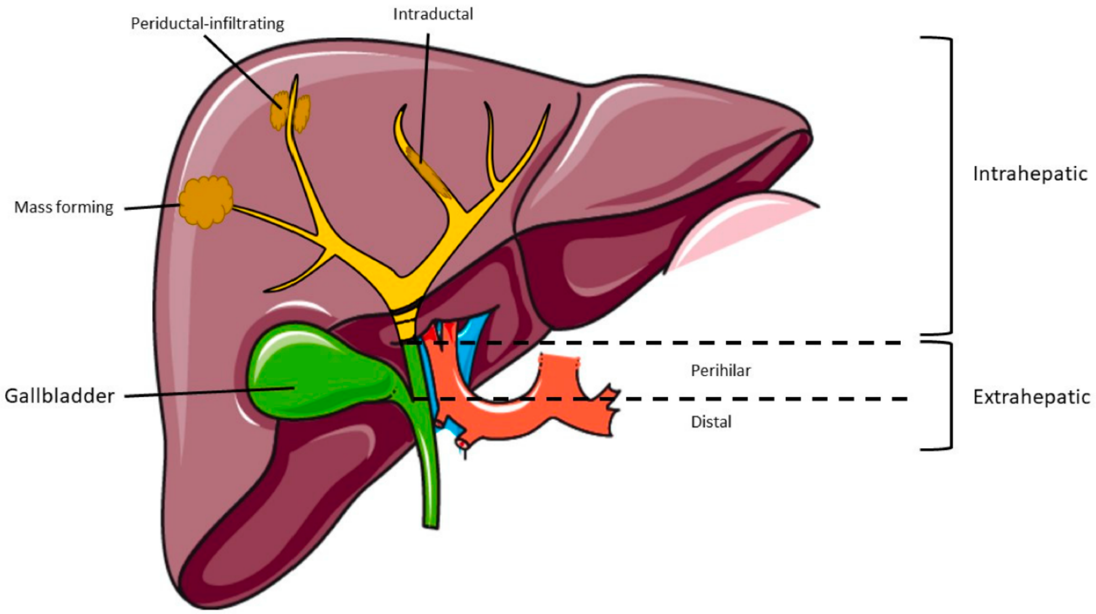
High performance application of virtual biopsy engine workflow

# Hodgkin Lymphoma



Cavinato, L. et al (2021). Recurrence-specific supervised graph clustering for subtyping Hodgkin Lymphoma radiomic phenotypes. 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

# Intrahepatic Cholangiocarcinoma



Intrahepatic cholangiocarcinoma (IHC) is an aggressive disease that affects the liver.

It is the second most common primary hepatic tumor and its incidence is increasing over last decades.

Diagnosis is difficult at early stages, due to IHC complicated biology.

The main treatment is surgery, chemotherapy has a limited effectiveness.

Five-years survival rate ranges from 25% to 40%.

## *Prognostic factors*

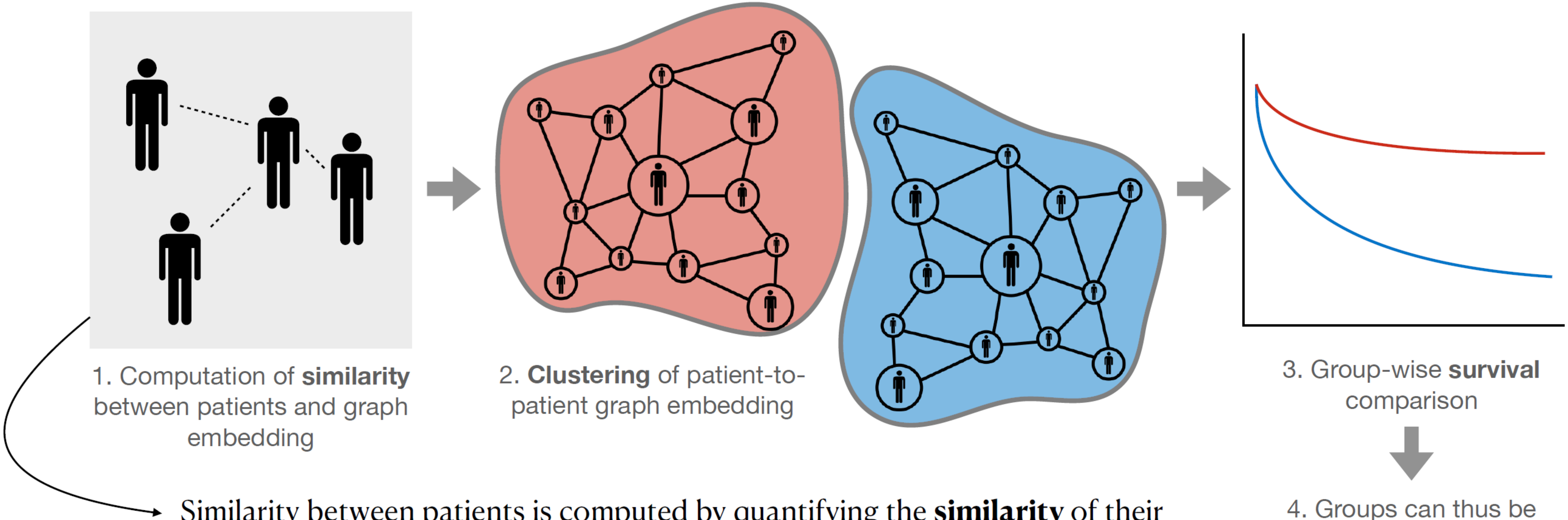
- Tumor size, number and distribution
- Tumor differentiation
- Vascular invasion
- Lymph nodes metastases
- Metabolic tumor volume
- R Status

## BUT !!

They are still debated, robust biomarker are lacking and precision medicine approach with an adequate **non-invasive preoperative assessment** of tumor biology and prognosis is still not available.

Cavinato, L., et al. (2021). Virtual Biopsy for Diagnosis of Chemotherapy-Associated Liver Injuries and Steatohepatitis: A Combined Radiomic and Clinical Model in Patients with Colorectal Liver Metastases. Cancers, 13(12), 3077.
Viganò, L. et al. (2021) Chemotherapy-associated liver injuries. Unmet needs and new insights for surgical oncologists. Annals of Surgical Oncology, 28(8): 4074–4079 doi: 10.1245/s10434-021-10069-z

# Intrahepatic Cholangiocarcinoma



1. Computation of **similarity** between patients and graph embedding

2. **Clustering** of patient-to-patient graph embedding

3. Group-wise **survival** comparison

4. Groups can thus be characterized in terms of significant covariates, both **endogenous** and **exogenous**

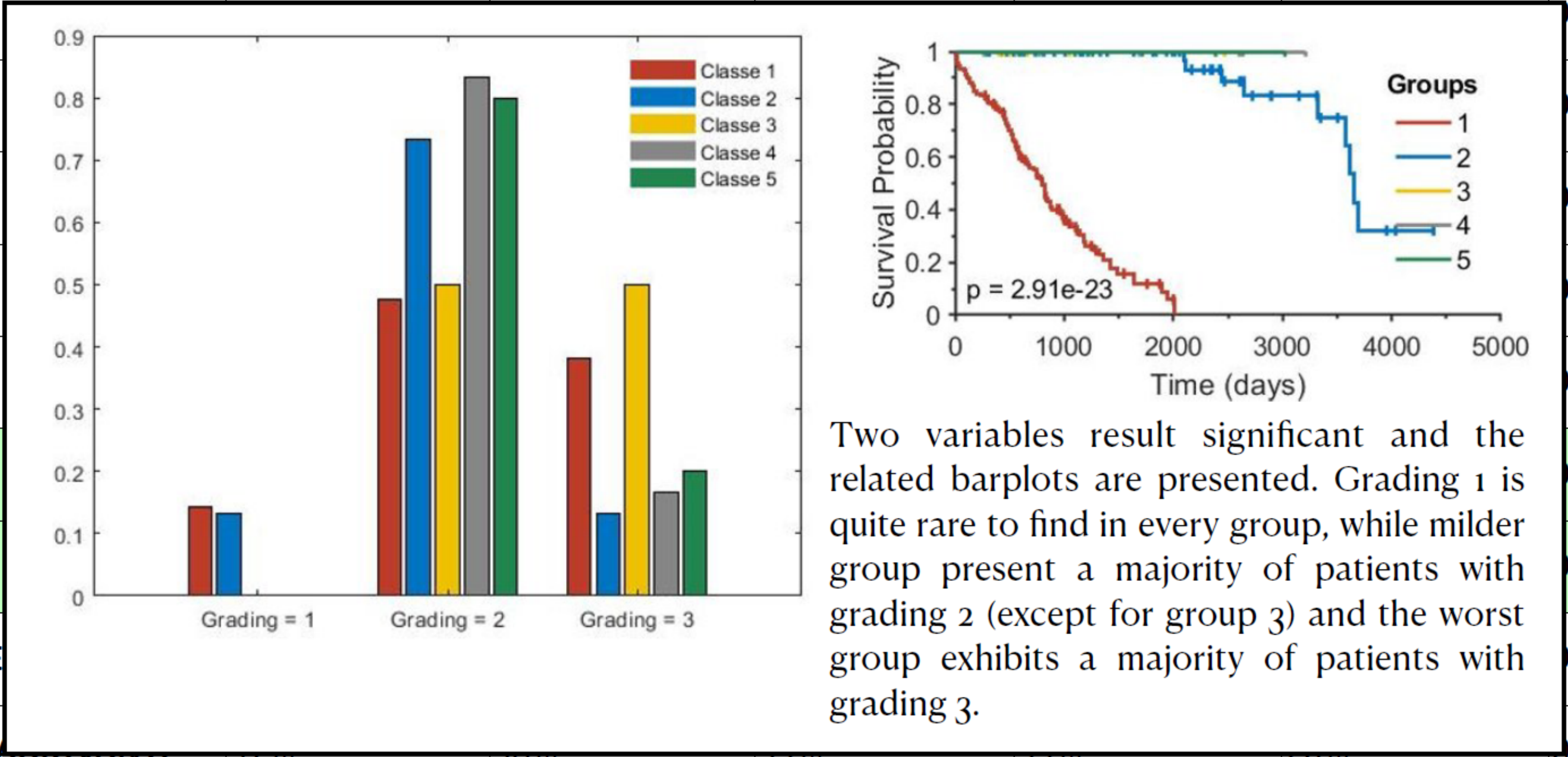
Similarity between patients is computed by quantifying the **similarity** of their **imaging** characteristics and similarity of their **time to event** (i.e., death or recurrence). According to similarities, patients are arranged in a **graph** where distance between nodes (patients) represents pair-wise similarity.

Chapfuwa, P. et al. (2020) Survival Cluster Analysis. ArXiV https://doi.org/10.48550/arXiv.2003.00355

# Intrahepatic Cholangiocarcinoma

## Cluster interpretation according to exogenous clinical variables

| Variabili (% nel gruppo) | GROUP 1 | GROUP 2 | GROUP 3 | GROUP 4 | GROUP 5 | P-value |
|---|---|---|---|---|---|---|
| **PATTERN = 1** | | | | | | .1730 |
| **PATTERN = 2** | | | | | | .4856 |
| **PATTERN = 3** | | | | | | .3603 |
| **SINGLE NODULE** | | | | | | .4682 |
| **GRADING = 1** | | | | | | .5990 |
| **GRADING = 2** | | | | | | .0074 |
| **GRADING = 3** | | | | | | .0081 |
| **INFILTRAZIONE PE** | | | | | | .0653 |
| **CHEMIOTERAPIA ADIUVANTE** | 17% | 38% | 33% | 33% | 20% | .5717 |



Two variables result significant and the related barplots are presented. Grading 1 is quite rare to find in every group, while milder group present a majority of patients with grading 2 (except for group 3) and the worst group exhibits a majority of patients with grading 3.

# Case Study III – take home messages

- **Early detection** of responders/not responders or long/short-term survivors may allow for personalized and more effective treatments

  => Medical imaging is the most promising driver of the non-invasive predictive medicine.

- Unfortunately, the lack of standardization in image processes, the need of human intervention for segmentation and reconstruction still pose issues in transfearability of results and general assessment of efficacy in personalized prediction

  => Suitable representation methods are of crucial importance, as well as methods which are able to account for hierarchical structure of the data in multicenter trials

- Despite its limitations, radiomics is one of the most common way to process medical images in order to plug their information into a predictive machinery

  => Balance between interpretability and predictive power to enforce clinical actionability

# Take Home Messages

# Take Home Messages

- The increasing complexity of healthcare research and data require nowadays a **major effort in developing novel statistical models and algorithms for personalized prediction**.

- Such effort should be devoted to the development of robust evidence to support the development of precision policies, in a context of **Evidence Based Decision Making.**

- This is definitively not an easy task, since many issues still remain (lack of standardization, regulation of data access, privacy, among others).

- **Data are not enough.**

  => More sophisticated and tailored analytics methods (new systems of health analytics, i.e., **integrated pipelines** going from data collection, to pre-processing tools and statistical models)

  - Shared (transdisciplinary) attention to a critical interpretation of the evidences generat as well as to their transfer to the decision level.

- **Complexity ask for new methods, not for more data**

- **Data cannot replace decisions** => Keep humans into the loop

# Acknowledgments

POLITECNICO
MILANO 1863

MOX
stat

PhD DADS

**Marta Spreafico**
PhD student in Mathematical Models and Methods for Engineering – 34th cycle

**Michela Massi**
PhD student in Data Analytics and Decision Sciences – 34th cycle

**Lara Cavinato**
PhD student in Data Analytics and Decision Sciences – 35th cycle

# Acknowledgments

Piercesare Secchi

Anna Paganoni

Laura Sangalli

Simone Vantini

Alessandra Menafoglio

Andrea Cappozzo

Chiara Masci

Michela Massi

Marta Spreafico

Lara Cavinato

Caterina Gregorio

Laura Savarè

Nicola Rares Franco

Vittorio Torri

Federica Corso

Letizia Clementi

Eleonora Arnone

Matteo Pegoraro

Riccarrdo Scimone

Riccardo Peli

Marta Cannistrà

Alessandra Ragni

Arianna Burzacchi

Teresa Bortolotti

Lidia Rossi

Thanks for your attention!