



gmds | Deutsche Gesellschaft für
Medizinische Informatik,
Biometrie und
Epidemiologie e.V.

Summer Academy

23.-26.9.2024, Schloss
Fürstenried, München

Workshop 3:

Use of AI and ML in
Statistics, with a focus on
application, interpretation
and small data challenges

Day 2 - Handling of
missing data in clinical
trials using machine
learning

GMDS Summer Academy: Handling of missing data in clinical trials using machine learning

Dr. Halimu Haliduola¹ (presenter)

Prof. Ulrich Mansmann²

Prof. Frank Bretz^{3,4}

23-26Sep2024

¹BioNTech SE, Germany

²IBE, LMU Munich, Germany

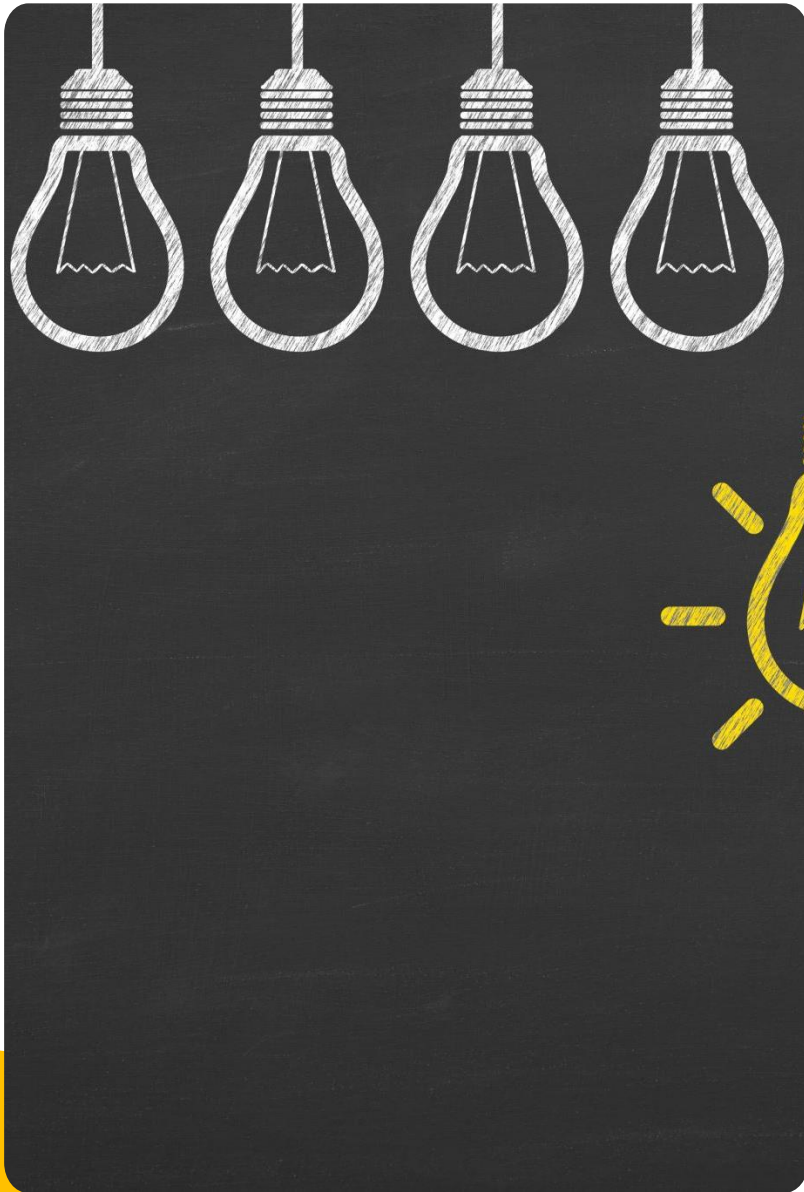
³Novartis Pharma AG, Basel, Switzerland

⁴Section for Medical Statistics, Medical University of Vienna, Austria



Disclaimers

- **This presentation reflects the views of the author and should not be construed to represent BioNTech's views or policies.**



Agenda

- **Missing Data (MD)**
 - Brief introduction on MD
 - Practicals - MI
- **Brief introduction on model evaluation**
- **Tree-based models**
 - Regression Tree, Random Forest (RF), QRF
 - Practicals - RF
- **Proposed method – UBR**
 - Motivation
 - Main idea
 - Workflow and methods
 - Results
- **UBR Practicals**
 - SMOTER
 - UBR
- **Discussion and Q&A**

Three types of missing mechanism (Rubin, 1976)

MD in outcome variable

MCAR

- Missing Completely at Random: if the probability of missingness **does not depend on observed or unobserved measurements**

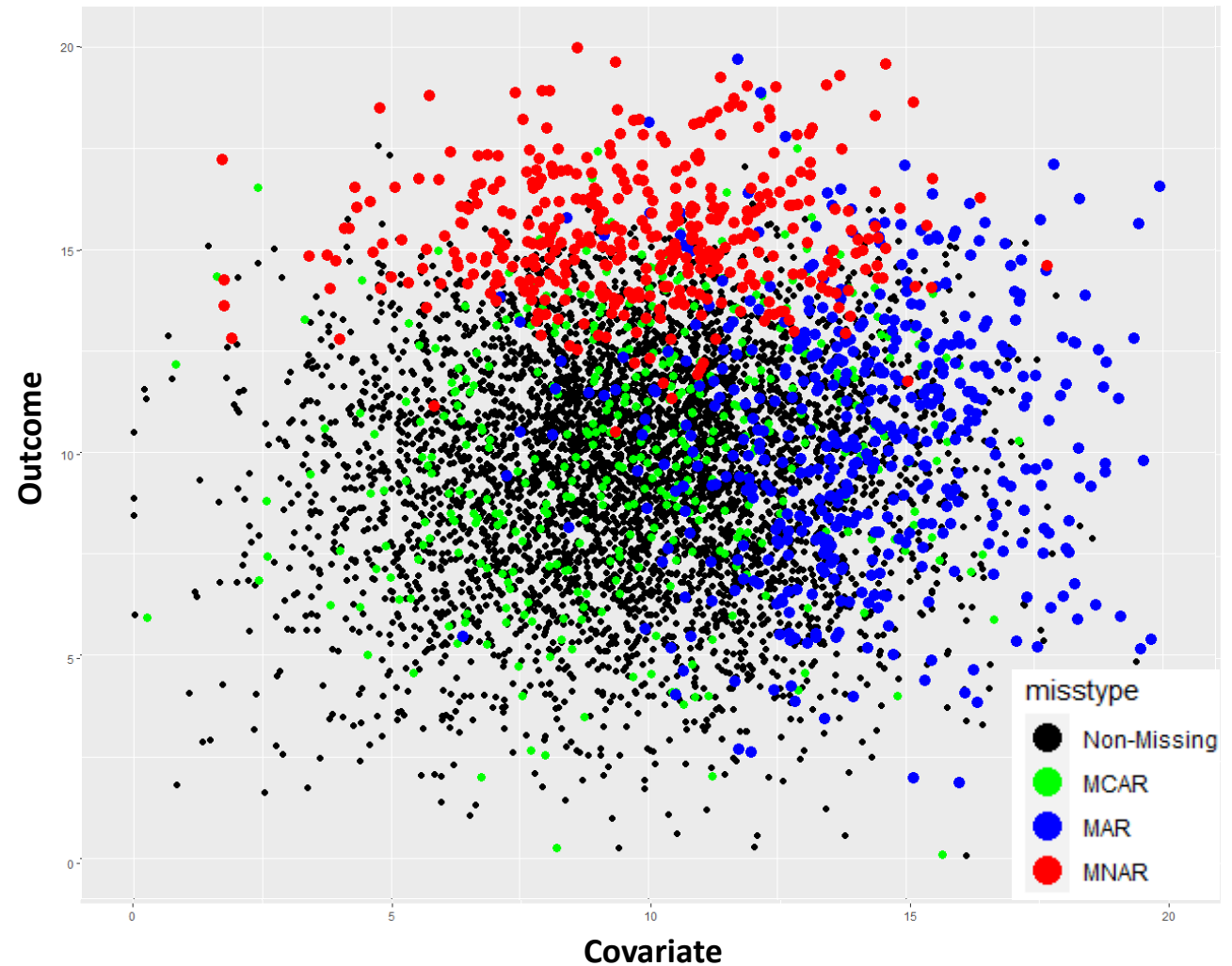
MAR

- Missing at Random: if the probability of missingness **depends only on observed measurements conditional on the covariate** in the model

MNAR

- Missing Not at Random: if the probability of missingness **depends on unobserved measurements**.

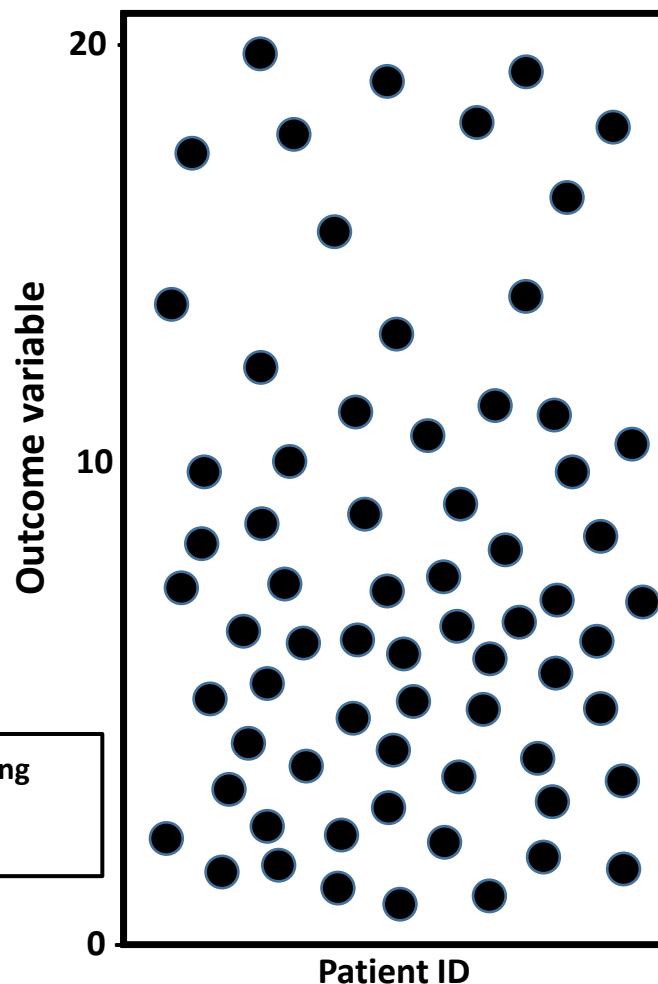
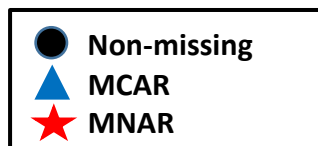
Simulation data with three types of MD



Why should we care about MD?

MD may cause:

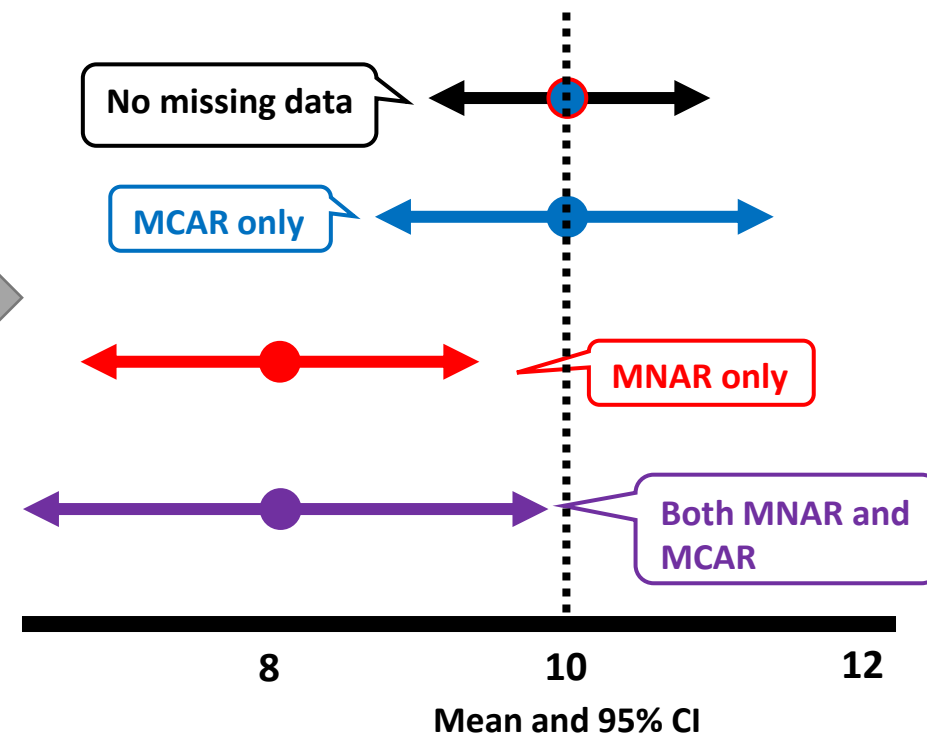
- lower power
- Bias
- Underestimation of variability
- Loss of external validity



Simulation data with MCAR and MNAR

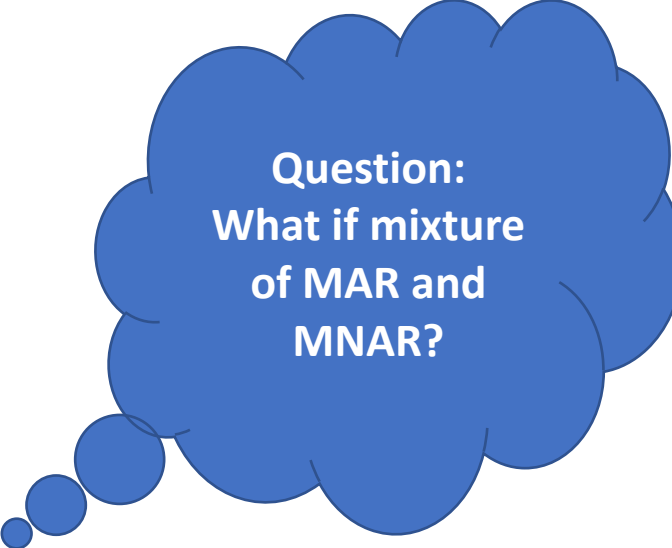
Analysis

Aggregated analysis results for different hypothetical scenarios



Traditional Methods to handle MD

- **Traditional data imputation**
 - Single imputation (LOCF, BOCF, Unconditional/conditional mean, worst/best case...)
 - **Multiple Imputation (MI)**
 - Under **MAR** or **MNAR** assumption respectively
- **Handling the MD by analysis method – no imputation.** E.g., for longitudinal continuous data
 - Under **MAR**: Mixed Model for Repeated Measurement (**MMRM**)
 - Under **MNAR**: Selection Model (**SM**), Shared Parameter Model (**SPM**), Pattern Mixture Model (**PMM**)
- **Sensitivity analysis**
 - **FDA**: sensitivity analyses should be planned that correspond to **alternative assumptions** about missing data.

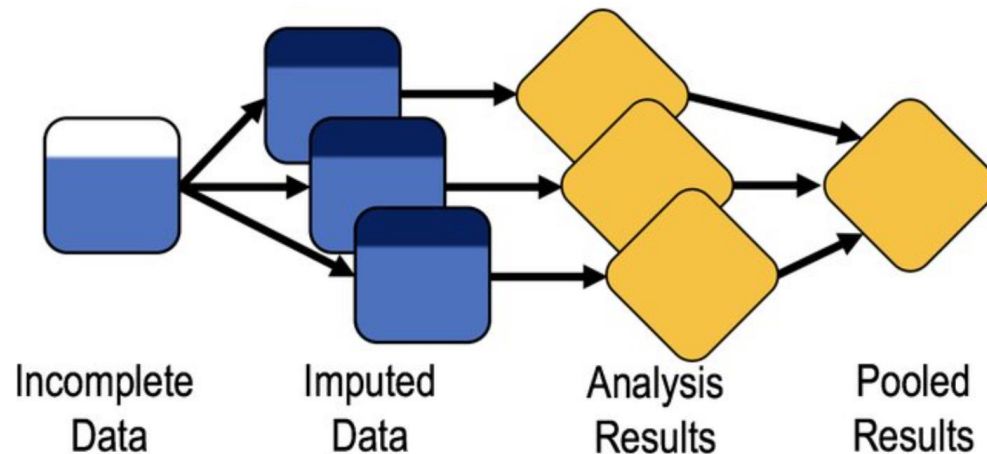


Question:
What if mixture
of MAR and
MNAR?

Traditional Methods to handle MD (cont'd)

Rubin's MI procedure (1987)

- **Step 1.** The missing data are filled in m (e.g., $m=50$) times to generate m complete datasets.
- **Step 2.** The m complete datasets are analyzed by using standard procedures.
- **Step 3.** The results from the m complete data sets are combined for the inference (using Rubin's rule).



of β . Let Q_j and W_j denote the point estimate and variance respectively from the j th ($j = 1, \dots, m$) complete dataset. The multiple-imputation point estimate Q^* of Q is the arithmetic mean of the m complete-data estimates. The estimated variance T of Q is obtained by a components-of-variance argument, leading to the following formulas

$$T = W + \left(1 + \frac{1}{m}\right) B$$

where

$$W = \frac{1}{m} \sum_{j=1}^m W_j$$
$$B = \frac{1}{m-1} \sum_{j=1}^m (Q_j - Q^*)^2$$

are the within- and between-imputation components of variance, respectively. For confidence intervals, [Rubin \(1987\)](#) gives the approximation

$$Q^* \pm t_\nu \sqrt{T}$$

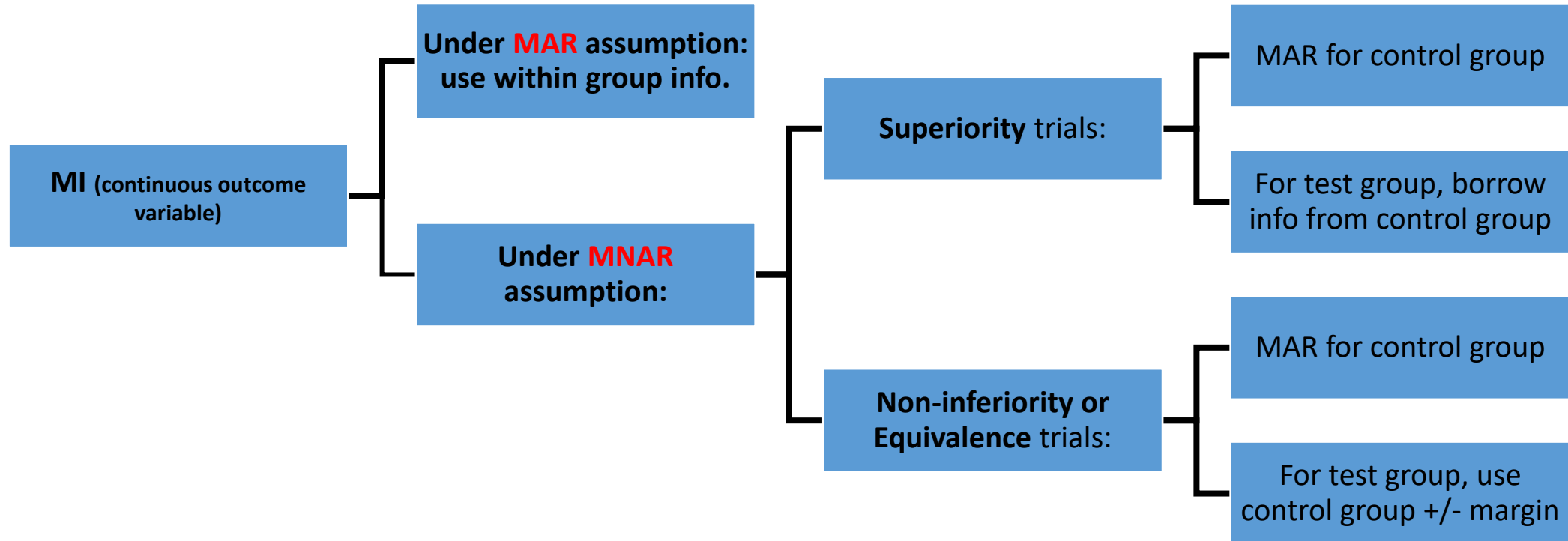
where the degrees of freedom ν are estimated by

$$\nu = (m-1) \left\{ 1 + \frac{W}{\left(1 + \frac{1}{m}\right) B} \right\}^2$$

and where t_ν is the appropriate fractile of the central t distribution on ν degree of freedom. Note that both ν and T are estimated from the data and that both depend on the quantity B . Note also that ν depends on $(W/B)^2$, a quantity that may have a large variance and a highly skew distribution. B itself is an estimated variance with $m-1$ degrees of freedom.

Multiple imputation (MI)

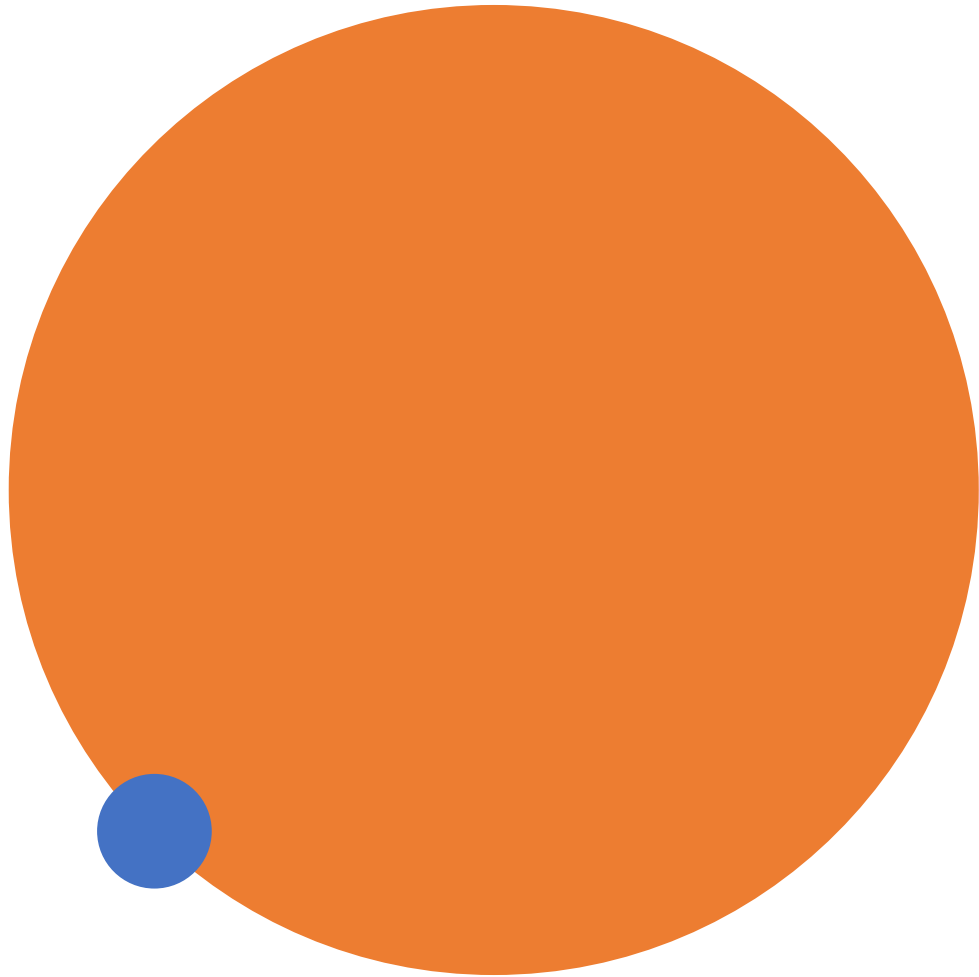
--- example scenarios



Practicals

“What is learned without pleasure is forgotten without remorse.” --- Epictetus

- To implement MI on a simulation data
- R package: “mice”
- **Method:** Predictive-Mean Matching (PMM).
- For a given subject with MD on the target variable, **PMM identifies subjects with no MD whose linear predictors** (created using the regression coefficients from the fitted imputation model) **are close to the linear predictors of the given subject**. Of those subjects who are close, one subject is selected at random and the observed value (**empirical data**) of the given variable for that randomly selected subject is used as the imputed value of the variable for the subject with missing data.



Brief introduction on ML in clinical research and model evaluation



ML: Challenges in Clinical Research

Research Quality & Mindset

- Breaking traditional mind set to research questions
- Relaxing the statistical assumptions
- Pooling from different domains and studies

Technical & Scientific

- Do we know how to make the best use of ML
 - Problems that seem to fit ML successes
 - Developing new methods for clinical research
- Causal inference
- High degree of uncertainty surrounding patient outcomes and ranking of patients

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

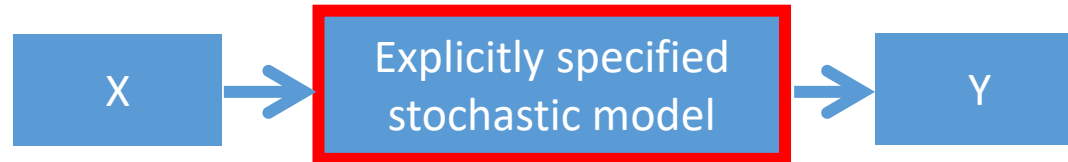
The Two Cultures

Nature



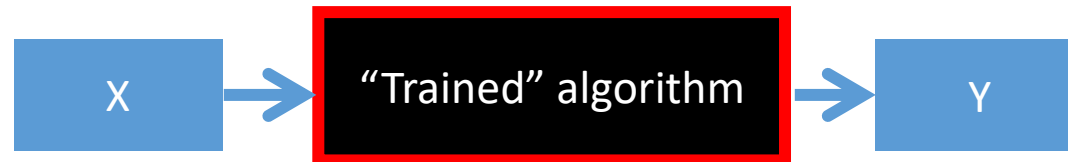
- Nature is a black box

«Data modeling culture»



- Simple models with interpretable parameters
- Emphasis on interpretability and inference

«Algorithmic modeling culture»



- Complex models that are trained rather than explicitly specified
- Emphasis on prediction rather than interpretability

Leo Breiman's opinion

- Model validation based on goodness of fit and residual examination – should be based on predictive accuracy.
- Led to irrelevant theory and questionable scientific conclusions.
- Kept statisticians from using more suitable algorithmic models and from working on interesting problems.
- Estimated 98% of statisticians follow this approach.
- The goal should be accurate information, not interpretability.

Comments on the ML culture



In 2001 Breiman claimed about 2% of statisticians would follow the machine learning or algorithmic approach



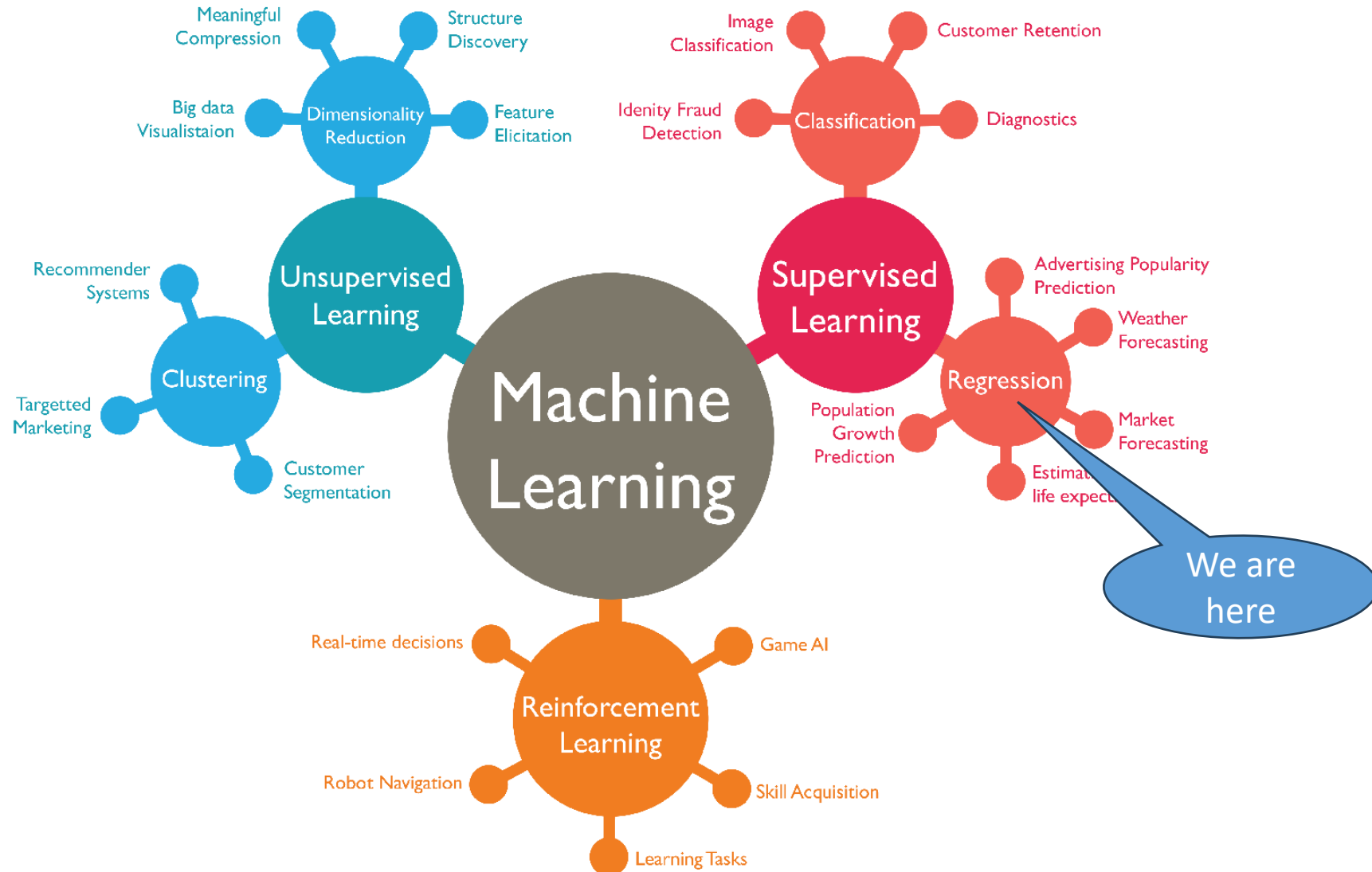
However, since then a large literature has developed in statistical machine learning.



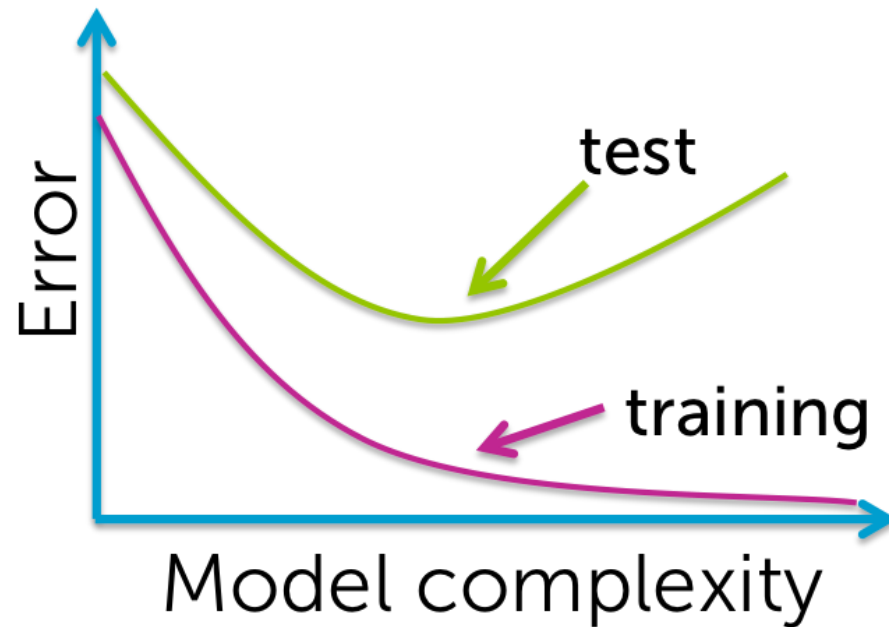
More recent development

Generative Model;
Individual prediction, e.g., treatment outcome, MD imputation;
Causal ML;
Targeted learning;
Interpretable ML; and others

Quick Summary

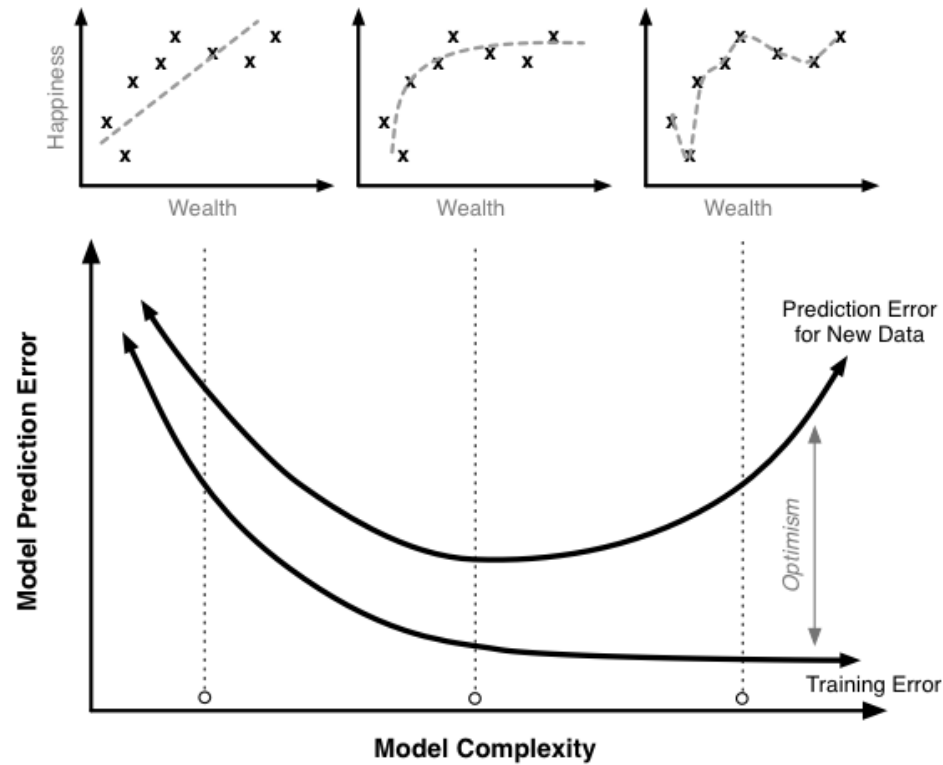


Model Evaluation



- The **training error** can easily be calculated by comparing predicted value and the actual value (e.g., MSE) in the original (training) dataset.
- The **test error** on the other hand is the average error that results from using the model (e.g., Random Forest) to predict the value on a **new observation** (one that was not used in training the model).
- **training error \neq test error**
the former can often dramatically underestimate the latter

Discrepancy between training and test performance hints at overfitting



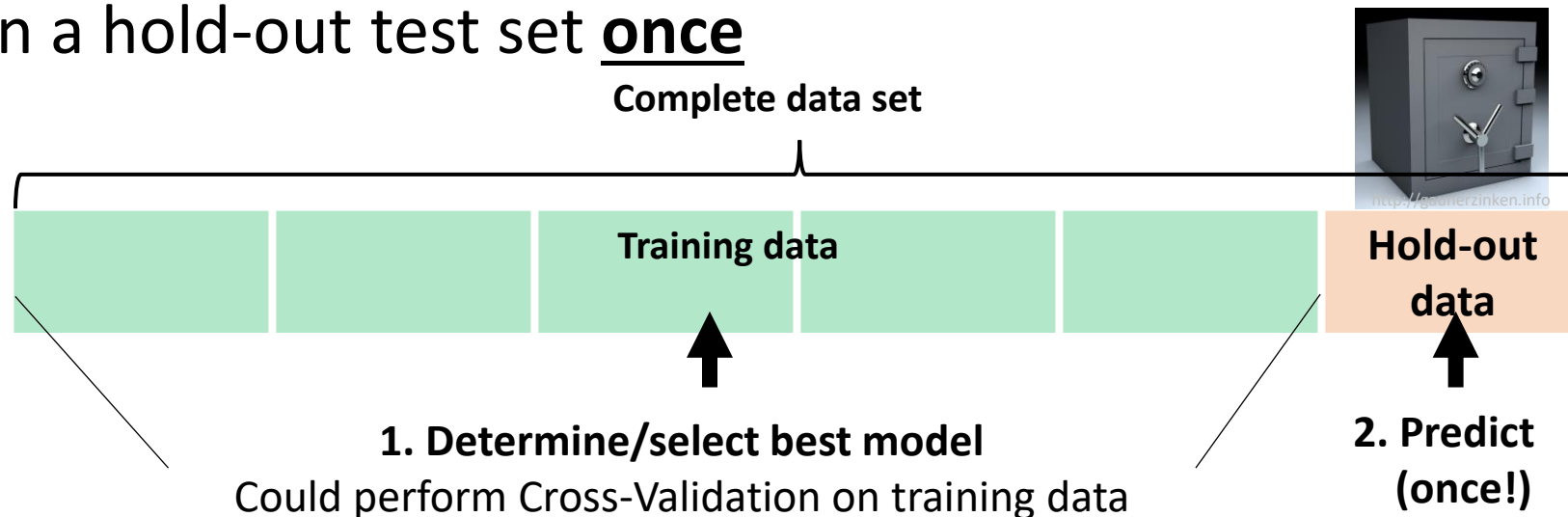
The **model might work very well on training set**, but that **doesn't mean that it will work well on new patients (test set)**

Hold-out test sets are the gold standard for model evaluation

- Training and testing on the same data set will overestimate performance → **Don't do this!**

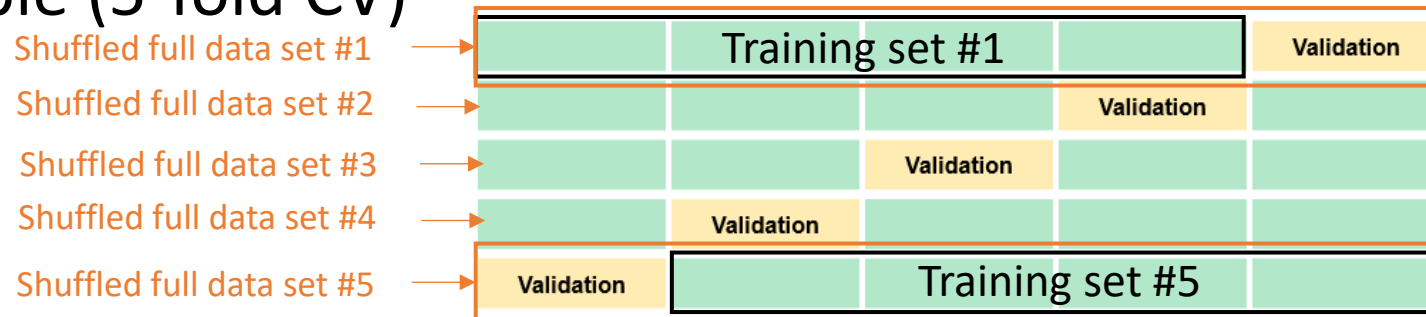


- The **gold standard** is to evaluate the trained, optimized and selected model on a hold-out test set **once**



K-fold Cross-Validation (CV)

- Widely used method for
 - estimating test error (model validation)
 - selecting best model or tuning parameter (model selection)
- Example (5-fold CV)



- For each shuffle, fit model on training set and predict on validation set!
- The average error of these predictions is called Cross-Validation error

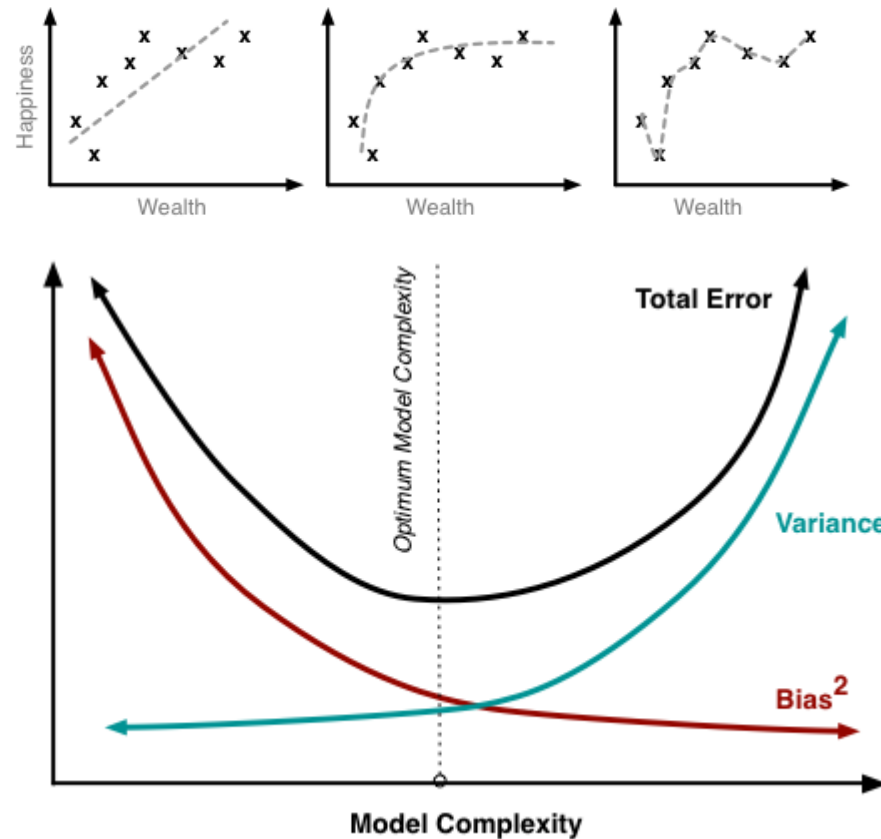
K-fold Cross Validation

for model validation and model selection

	Training set #1			Validation
			Validation	
		Validation		
	Validation			
Validation		Training set #5		

- ... for model validation
 - **CV error provides a reasonable estimate of the test error** of a given model (as long as you fit only that one model across the 5 shuffles)
- ... for model selection
 - You can fit multiple models (e.g. with different tuning parameters) to each of the 5 shuffles and choose the model with smallest CV error.
 - In this case however the **CV error of best model may not reflect test error**

Over-fitting can be understood as bias-variance trade-off



Aim to predict Y using covariate X , we may assume a relationship $Y = f(x) + \varepsilon$ (where $\varepsilon \sim N(0, \sigma_\varepsilon)$).

We estimate a model $\hat{f}(X)$ of $f(X)$

Then the MSE is: $Err(x) = E[(Y - \hat{f}(x))^2]$

This error can be decomposed into

$$Err(x) = \underbrace{\left(E[\hat{f}(x)] - f(x)\right)^2}_{\text{Bias}^2} + \underbrace{E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right]}_{\text{Variance}} + \underbrace{\sigma_e^2}_{\text{Irreducible error}}$$

Model prediction under infinite data True unknown relation Model prediction under actual data Model prediction under infinite data

$$Err(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$



Tree-based ML models

- **Simple Regression Tree**
- **Random Forest**
- **Quantile Regression Forest**
- **Others:** E.g., Adaptive Boosting and Gradient Boosting Machine, etc.

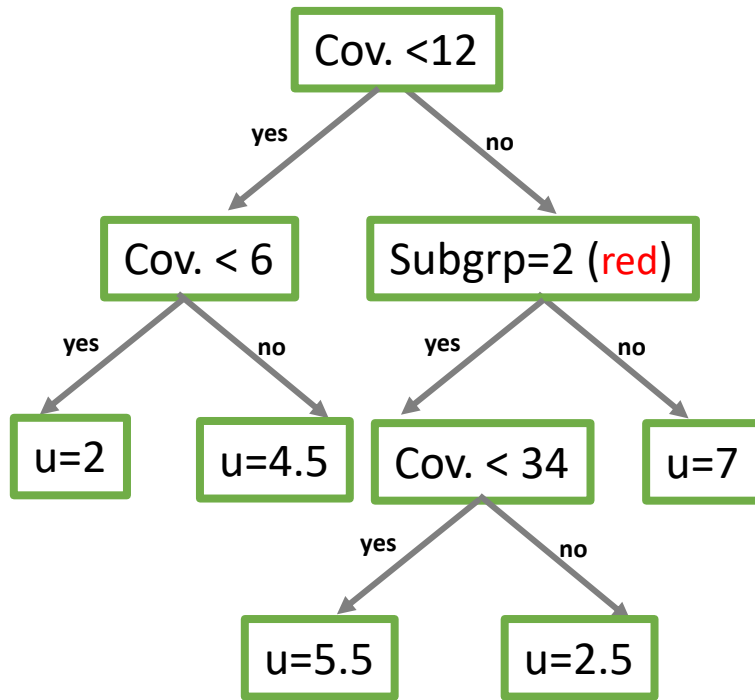
Regression Tree

- A branch of the decision trees (with continuous target variable) that divides the dataset into smaller regions and then fits a simple model (**average**) for each region.
- The model begins with the entire dataset, and searches **every distinct value** of **every input variable** to find the **predictor** and **split value** that partitions the data into two regions such that the **mean squared error (MSE) are minimized**:

$$E(Y|X = x) = \arg \min_z E\{(Y - z)^2|X = x\}$$

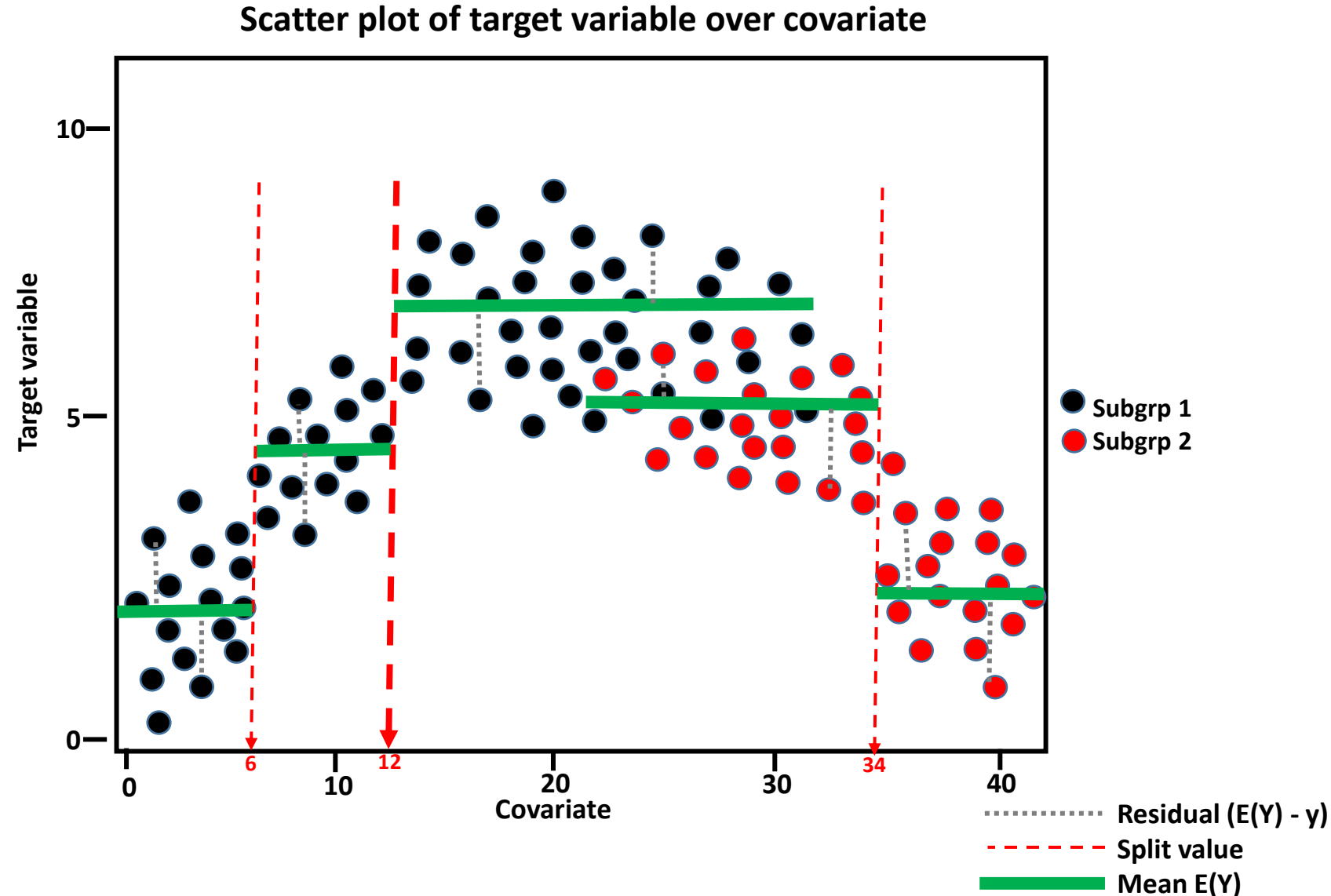
- Having found the best split, the data is split into the two resulting regions and **repeat the splitting process** on each of the two regions. This process is continued until some stopping criterion is reached (to avoid model overfitting).

Regression Tree (cont'd) – Example



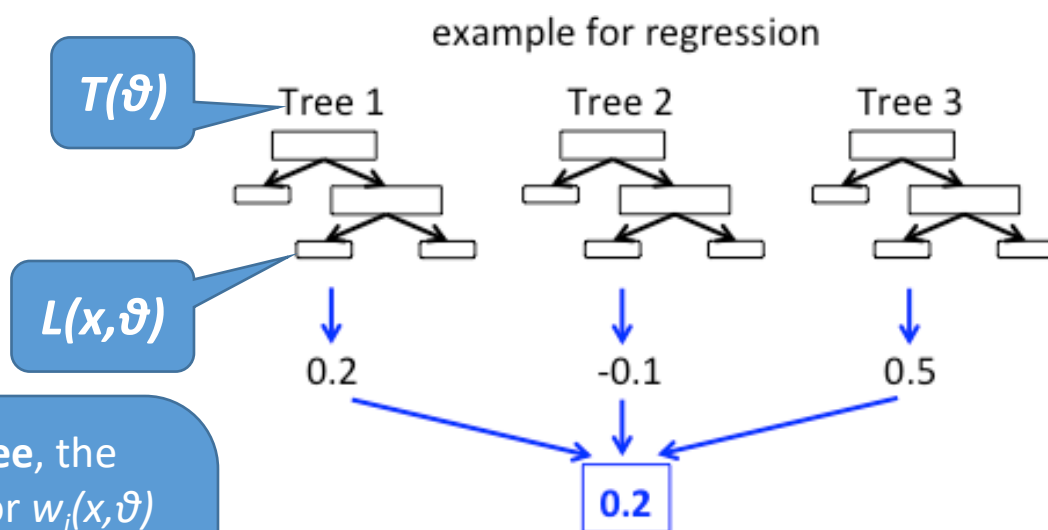
New data:

- P1: Cov=10, subgrp=1 → predicted value = 4.5
- P2: Cov=20, subgrp=1 → predicted value = 7.0
- P3: Cov=30, subgrp=2 → predicted value = 5.5
- P4: Cov=40, subgrp=2 → predicted value = 2.5



Random Forest (Breiman, 2001)

- Build many trees in parallel (e.g., $N_{\text{tree}}=500$), each tree is based on:
 - **bootstrapped data:** Random sampling with replacement each time, e.g., 2/3 as the original data size
 - **random subset of variables:** Random selecting predictive variables, e.g., 1/3 of the all variables
- The variety is what makes random forest more effective than individual decision tree.



For **single tree**, the weight vector $w_i(x, \vartheta)$ is a positive constant if observation X_i is part of leaf $L(x, \vartheta)$ and 0 if it is not. $w_i(x, \vartheta)$ sum to 1.

Tool: R package “randomForest”

Random forests prediction is weighted **conditional mean**.

For a single tree: $\hat{\mu}(x) = \sum_{i=1}^n w_i(x, \theta) Y_i$.

Weights of random forests with k trees:

$$w_i(x) = k^{-1} \sum_{t=1}^k w_i(x, \theta_t).$$

Random forests prediction: $\hat{\mu}(x) = \sum_{i=1}^n w_i(x) Y_i$.

The conditional mean **minimizes the mean squared error (MSE)**:

$$E(Y|X = x) = \arg \min_z E\{(Y - z)^2 | X = x\}$$

Quantile Regression

(Koenker, 2005)

- Beyond the Conditional Mean
 - The conditional distribution function $F(y|X = x)$ is given by the probability that, for $X = x$, Y is smaller than y , $F(y|X = x) = P(Y \leq y|X = x)$
 - For a continuous distribution function, the α -quantile $Q_\alpha(x)$ is then defined such that the probability of Y being smaller than $Q_\alpha(x)$ is, for a given $X = x$, exactly equal to α . The quantiles give more complete information about the distribution of Y as a function of the predictor variable X than the conditional mean alone.
- Quantile regression aims to estimate the conditional quantiles from data. Let the loss function L_α be defined for $0 < \alpha < 1$ by the weighted absolute deviations:
- Conditional quantiles minimize the expected loss $E(L_\alpha)$,

$$L_\alpha(y, q) = \begin{cases} \alpha |y - q| & y > q \\ (1 - \alpha) |y - q| & y \leq q \end{cases}$$

$$Q_\alpha(x) = \arg \min_q E\{L_\alpha(Y, q)|X = x\}$$

Quantile Regression Forest

(Meinshausen, 2006)

- For quantile regression forests, **trees are grown as in the standard random forest algorithm**. The conditional distribution is then estimated by the weighted distribution of observed response variables, where the weights attached to observations are identical to the original random forests algorithm.
- The **key difference** from RF, for each node in each tree,
 - **RF keeps only the mean** of the observations that fall into this node and neglects all other information.
 - **QRF keeps the value of all observations** in this node (not just their mean), and **assesses the conditional distribution** based on this information.

- The conditional distribution function of Y , for $X = x$, is given by

$$F(y|X = x) = P(Y \leq y|X = x) = E(1_{\{Y \leq y\}}|X = x)$$

- define an approximation to $E(1_{\{Y \leq y\}}|X = x)$ by the weighted mean over the observations of $1_{\{Y \leq y\}}$.

$$\hat{F}(y|X = x) = \sum_{i=1}^n w_i(x) 1_{\{Y_i \leq y\}}$$

- where $w_i(x)$ is the same weights as in RF algorithm.
- **Tool:** R package “quantregForest”

- **Practicals - Random Forest**

- R package “randomForest”

Three types of missing mechanism (Rubin, 1976)

MCAR

- Missing Completely at Random: if the probability of missingness **does not depend on observed or unobserved measurements**

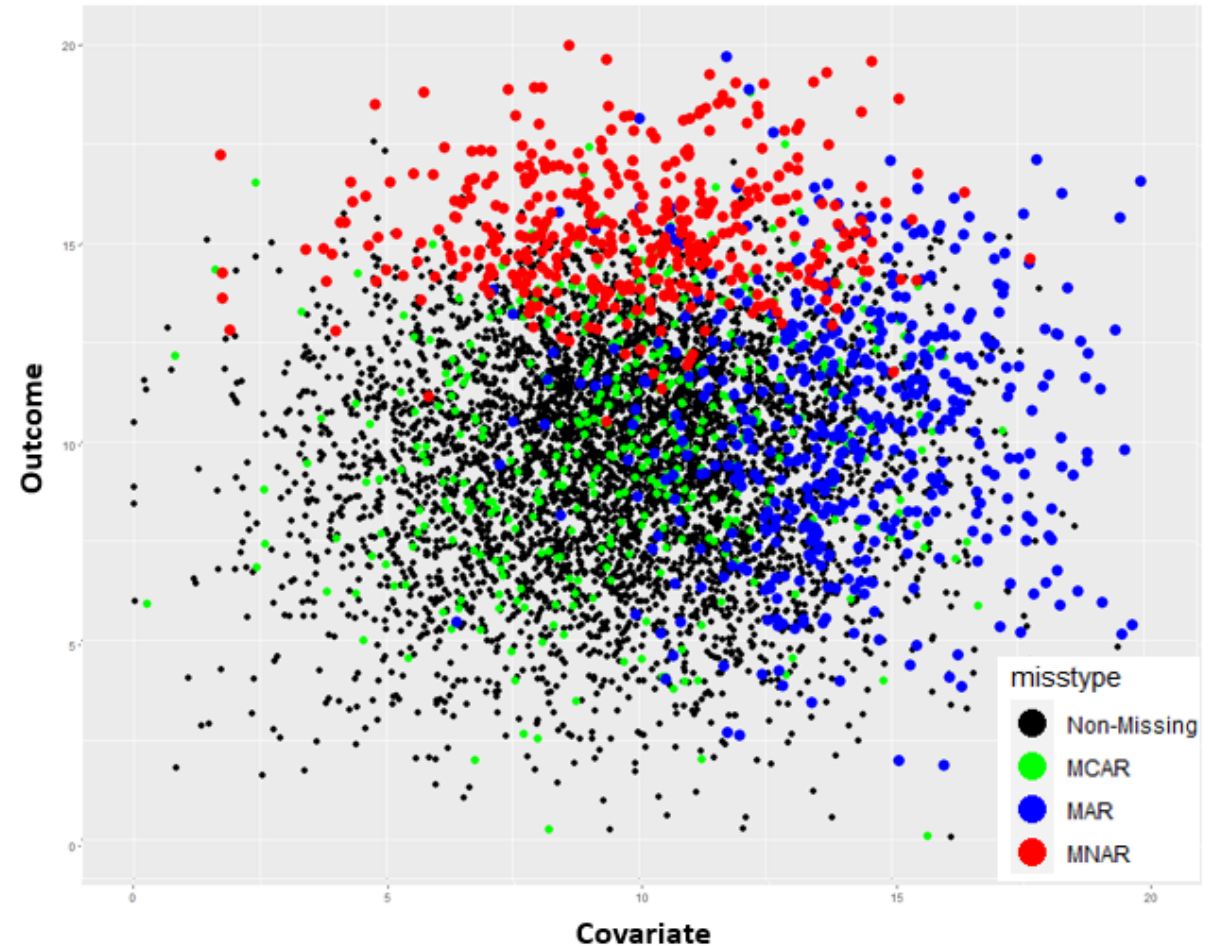
MAR

- Missing at Random: if the probability of missingness **depends only on observed measurements conditional on the covariate** in the model

MNAR

- Missing Not at Random: if the probability of missingness **depends on unobserved measurements**.

Simulation data with three types of MD



Difficult to make assumption about the missing mechanism at the population level

- It is difficult to ascertain whether there is a relationship between missing values and the unobserved outcome variable for the entire study population.
- It is not possible to ascertain whether the MAR/MCAR assumptions are appropriate in any practical situation.
- A proposition that none of the data missing in a confirmatory clinical trial are MNAR seems implausible.
- **EMA (2010):** *“A combined strategy incorporating several methods for handling missingness (e.g. assume dropouts due to lack of efficacy are MNAR and lost to follow-up are MAR) may also be considered.”*

Our proposal

- Handling of realistic missing data scenarios in clinical trials using machine learning (ML) techniques

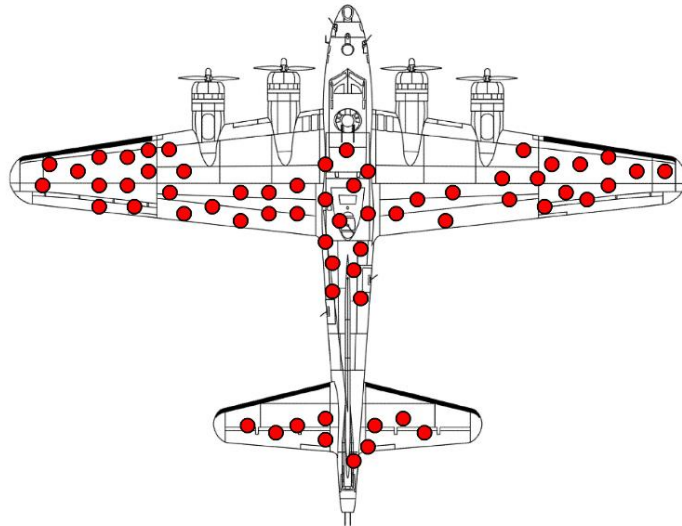
- We aim to handle the realistic missing data scenarios (e.g., mixture of MAR and MNAR).
- We use ML techniques to **provide accurate individual prediction** for missing data.
 - *Breiman (2001), in statistical learning “the goal is not interpretability, but accurate information.” -- Statistical modeling: the two cultures.*
- We treat the MNAR as an **imbalanced learning** task, i.e., we give more focus to the MNAR data by assigning more sample weights to the data points in that area.
 - *Abraham Wald and the missing bullet holes – a war time story*

Abraham Wald and the missing bullet holes



The armor, said Wald, doesn't go where the bullet holes are. It goes where the bullet holes aren't: on the engines.

The Survivor

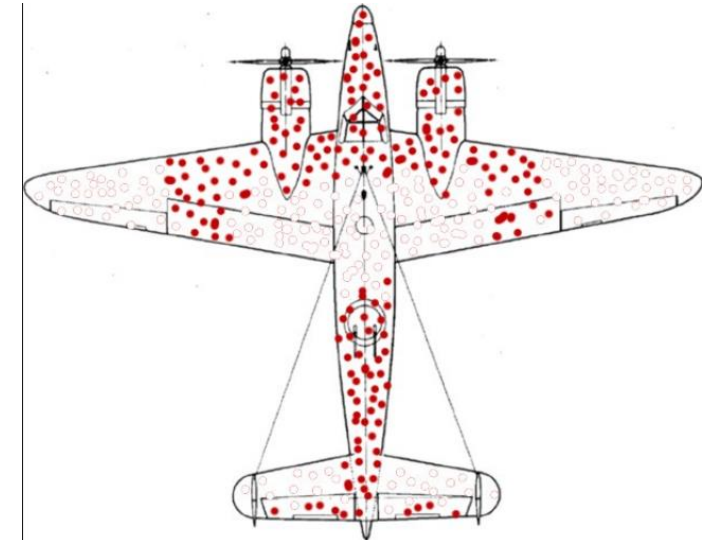


Data from the Survivors

Section of plane	Bullet holes per square foot
Engine	1.11
Fuselage	1.73
Fuel system	1.55
Rest of plane	1.8

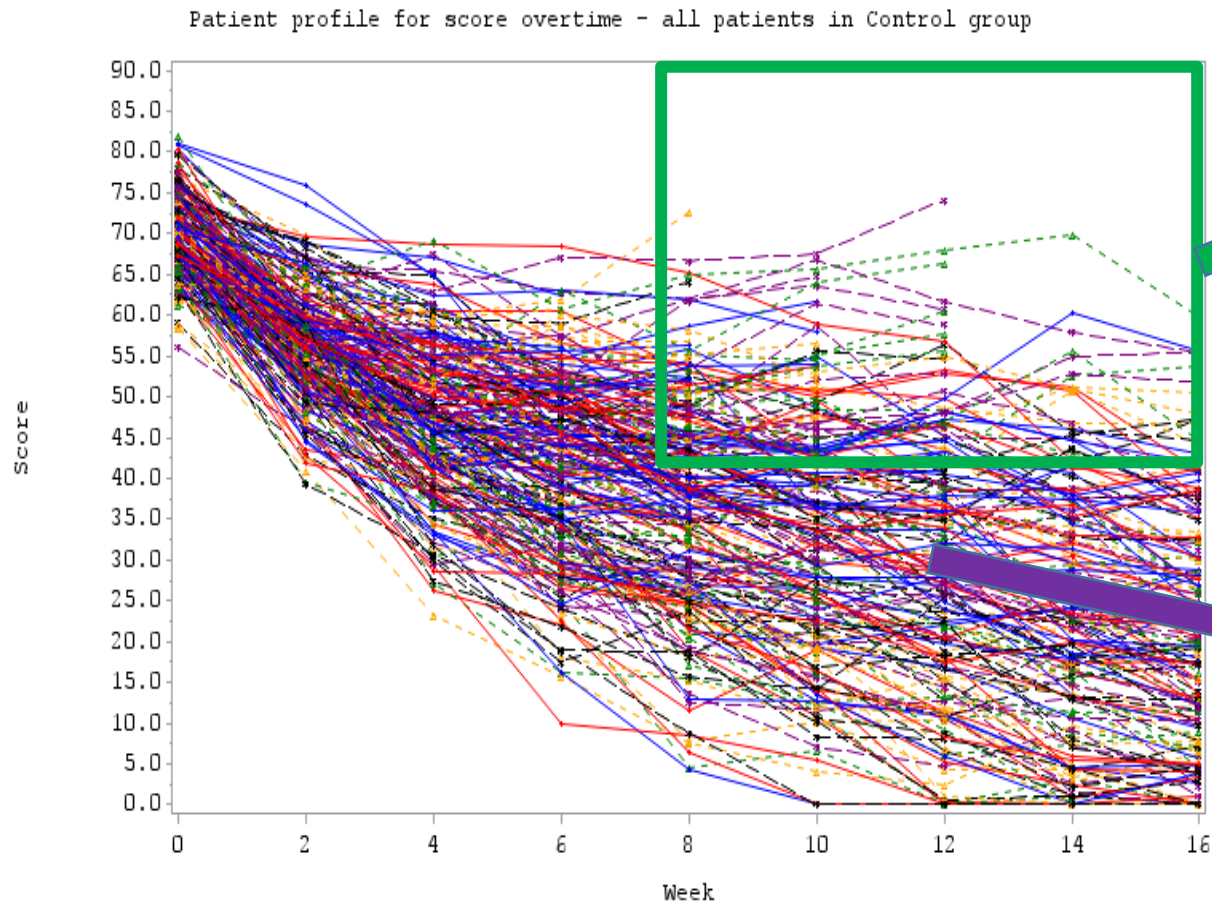
The airplane that never came back

???

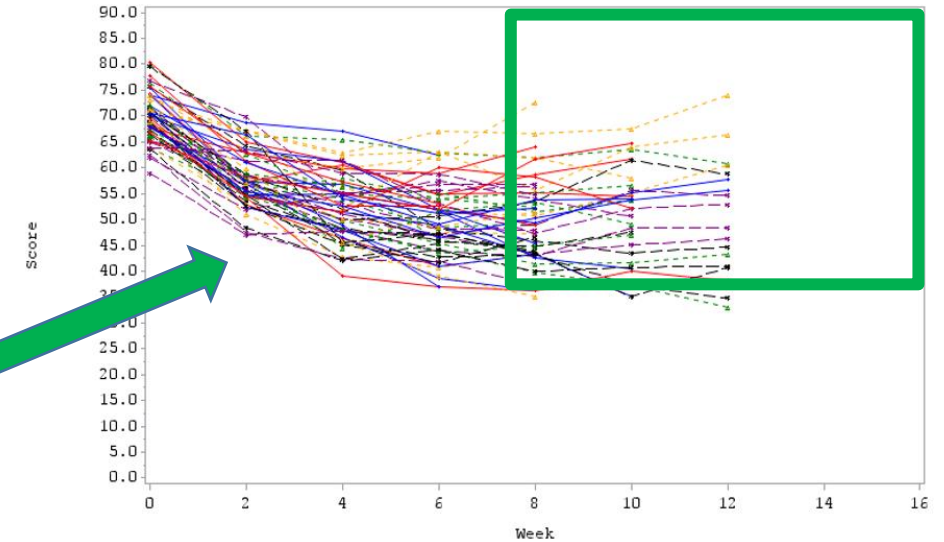


Clinical Trial: longitudinal data example

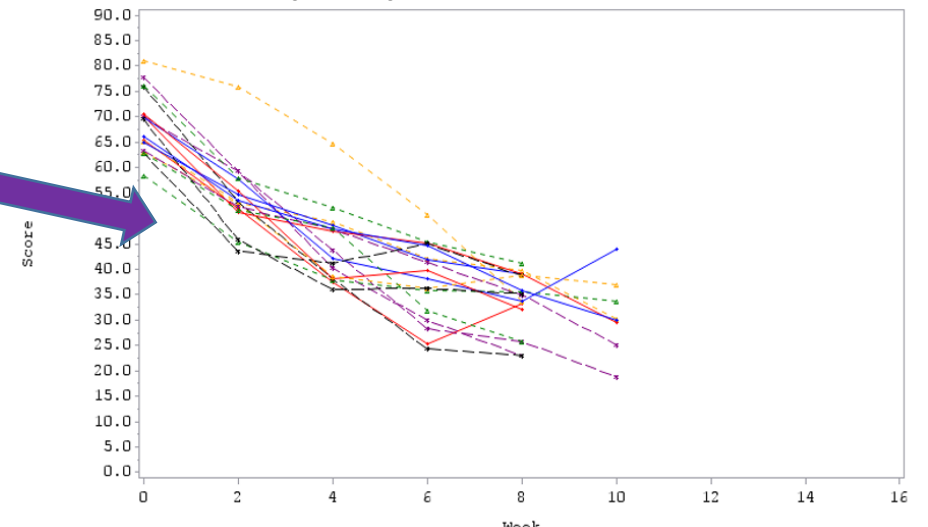
All patients



Patients who discontinued due to lack of efficacy (MNAR)

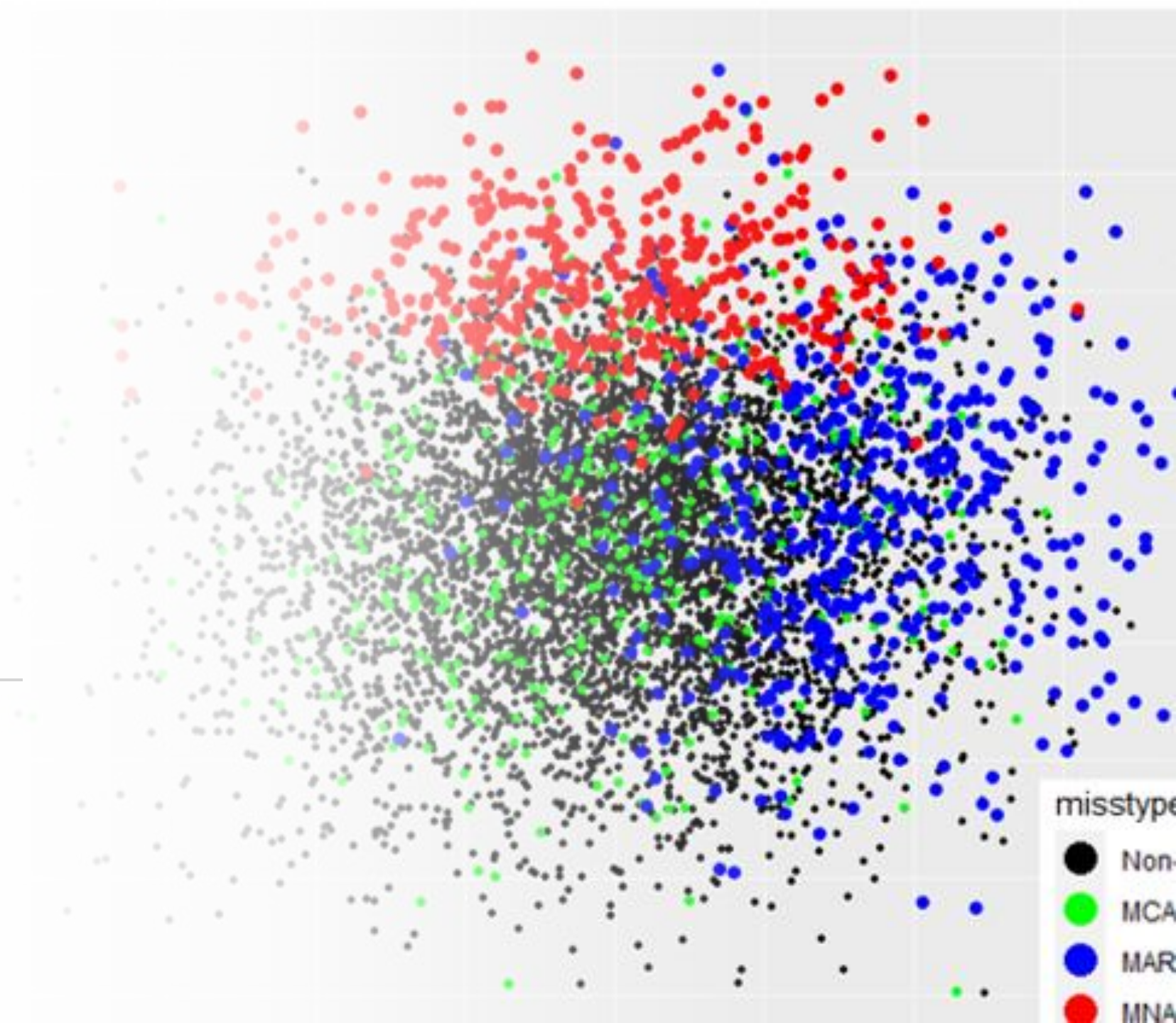


Patients who discontinued due to other reason (MAR)



Main problem

MNAR results in an **imbalanced distribution** of available data for the target variable

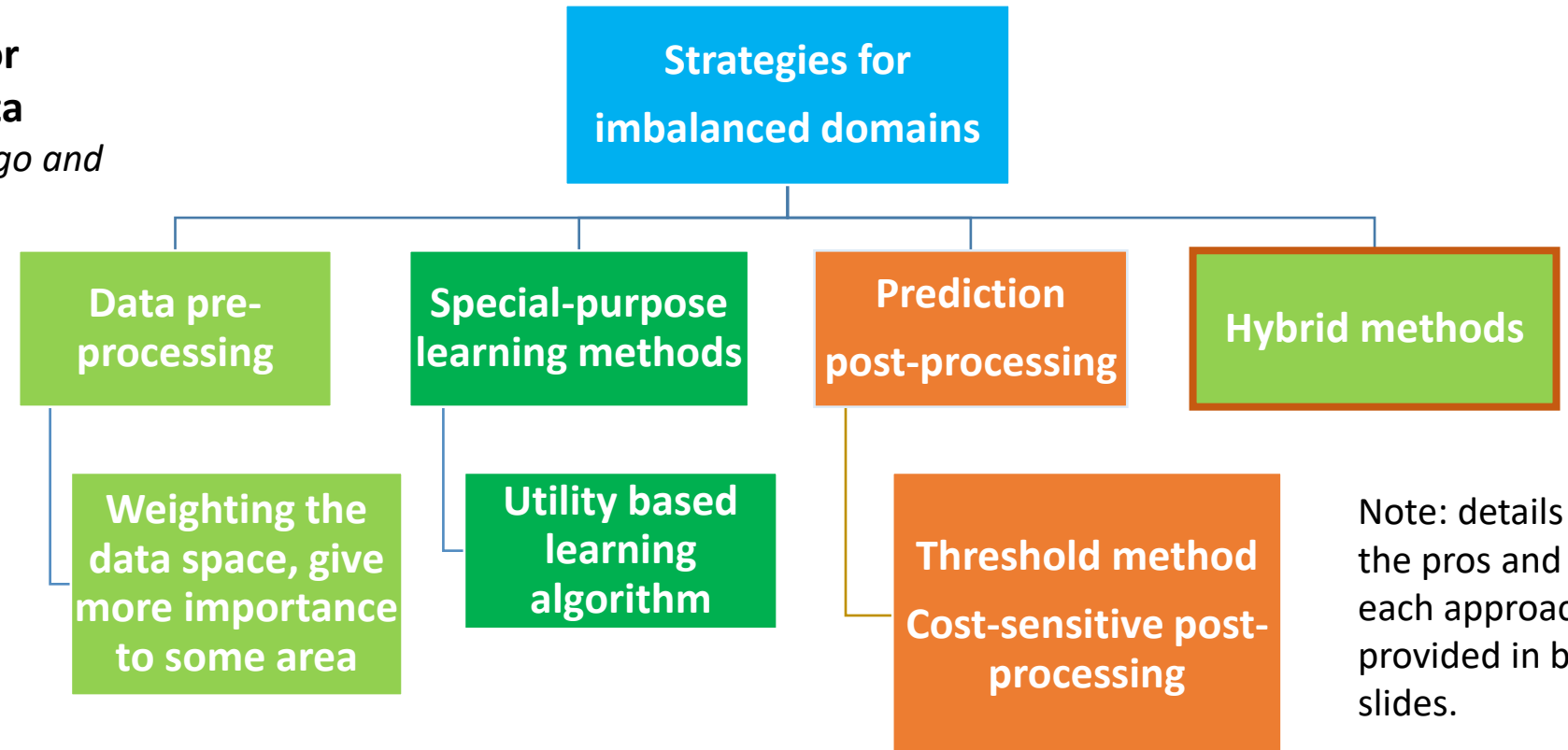


Predictive modeling with imbalanced distribution

- Many predictive tasks involve handling a target variable that has an imbalanced distribution in the available training data.
- Problem thoroughly explored in classification tasks.
- Similar problems occur in some regression tasks.

A taxonomy of strategies for handling of imbalanced data

Modified based on [Branco, Torgo and Ribeiro, 2016]



Note: details about the pros and cons of each approach are provided in backup slides.

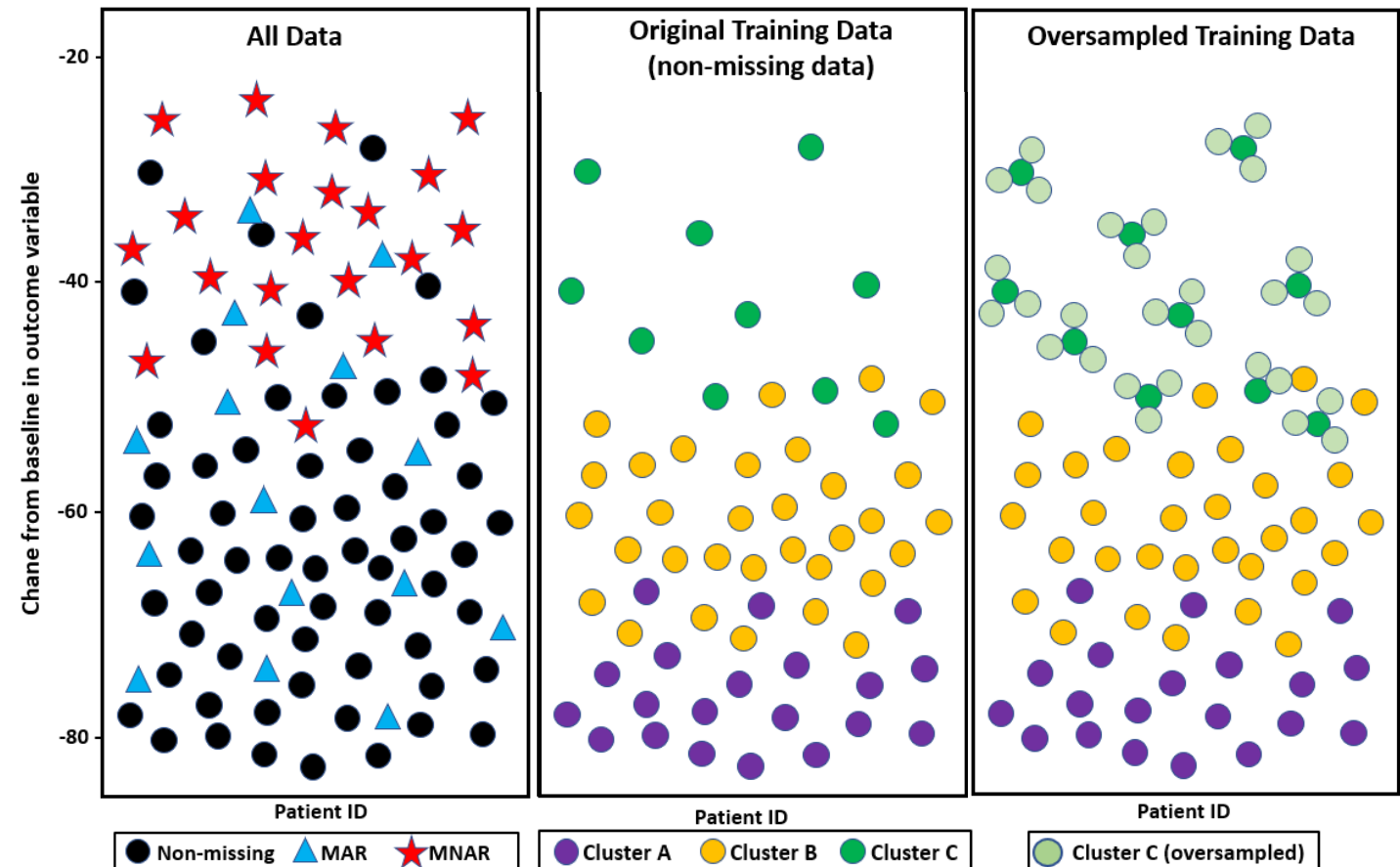
Missing data imputation in clinical trials using recurrent neural network facilitated by clustering and oversampling

Halimu N. Haliduola^{1,2} | Frank Bretz^{3,4} | Ulrich Mansmann¹

Main idea in paper I longitudinal data

- The problem of MNAR is seen as an **imbalanced machine learning** task, i.e., to **oversample the minority cases** (Weiss, 2013) to **compensate for the MNAR** data in certain area. It can also handle the MAR data by looking for accurate individual information.
- To be able to use the oversampling in longitudinal continuous variable, **clustering through k-mean algorithm** (Gower, 1971) needs to be performed the first.
- Recurrent neural networks (RNN) model to fit the longitudinal trajectories. RNN can learn from the past to predict the future outcomes.
- Error metrics: MSE**

Illustration of clustering and oversampling for target variable



Main idea in paper II

- **Data pre-processing:**

- the simple random oversampling may lead to **too much duplicated cases** in the data, and hence cause **overfitting** of the model.
- new method **SMOTER** – generates synthetic data points, which is proven to be more generalizable.

- **Special-purpose Learning - UBL**

- the inadequacy of standard error measures: e.g., MSE and MAD, they are not suitable in the context of a regression problem with non-uniform costs. **Their weakness lies in taking all the prediction errors equally across the range of the target variable.** [Torgo and Ribeiro, 2009]
- UBL considers both the prediction error and the relevance (importance) of target variable values

Considering the MNAR (red dots), distribution of available non-missing data (black dots) are **imbalanced** in the range of target variable.

Computer Methods and Programs in Biomedicine 226 (2022) 107172

Contents lists available at ScienceDirect

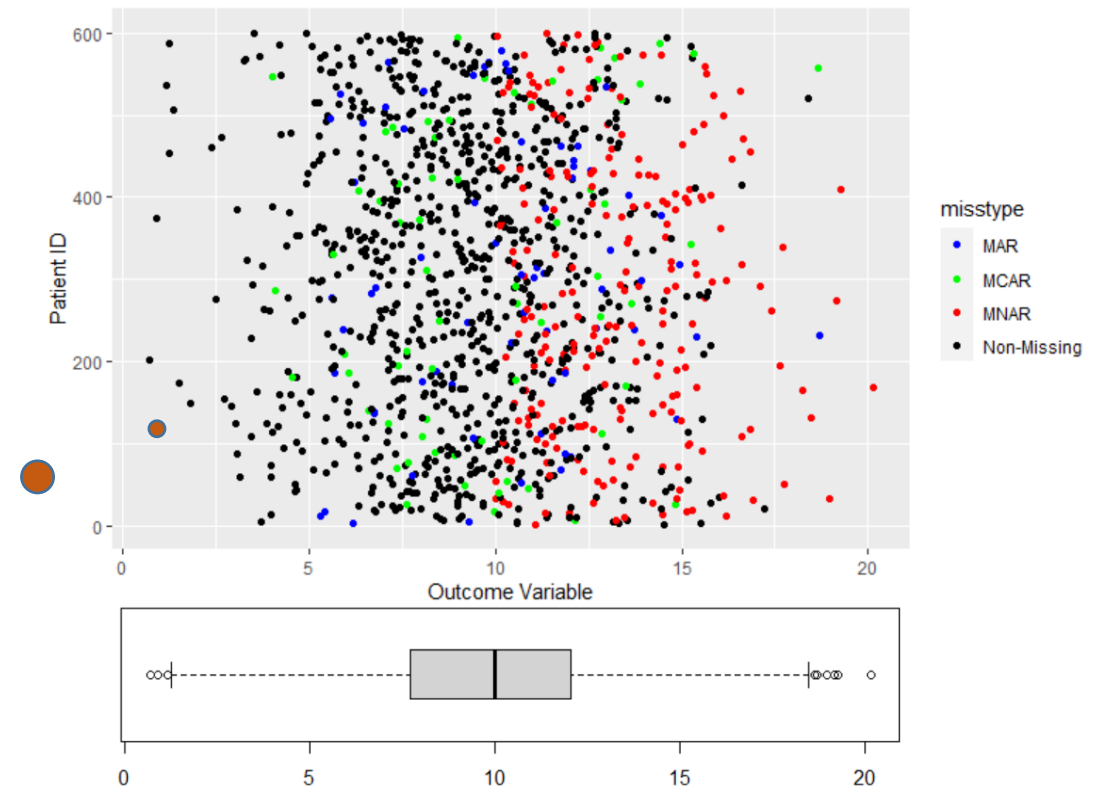
Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Missing data imputation using utility-based regression and sampling approaches

Halimu N. Haliduola^a, Frank Bretz^{b,c}, Ulrich Mansmann^{a,*}

Check for updates

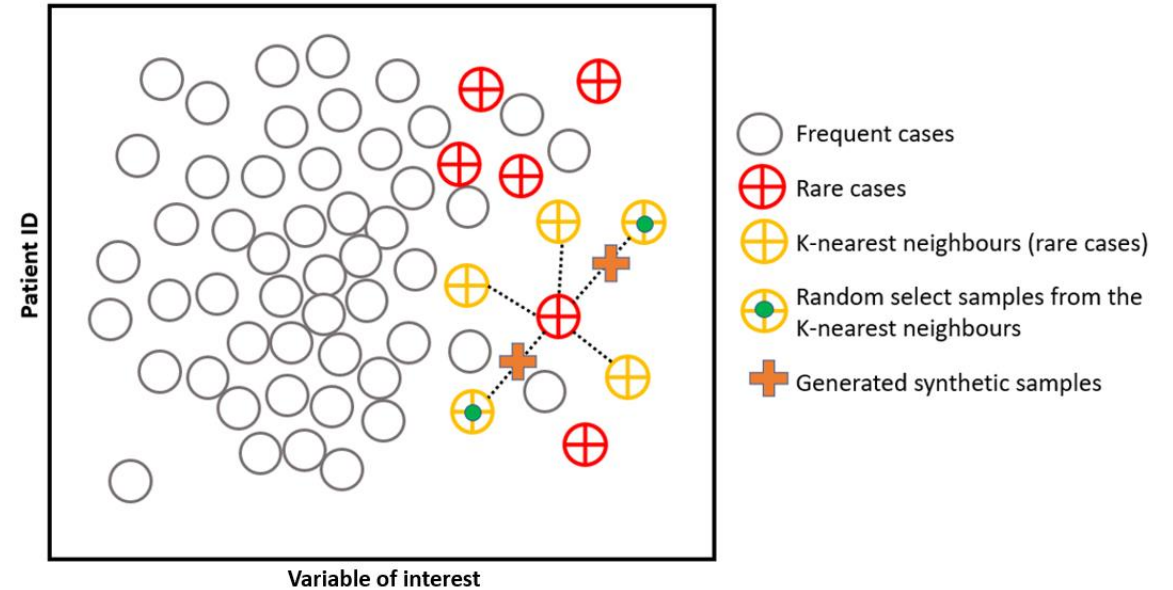


Simulation data: scatter and boxplot for the target variable.
back = non-missing data, blue = MAR, green = MCAR, red = MNAR.

Data pre-processing: SMOTER

SMOTE: Synthetic Minority Over-sampling Technique
[Chawla et al. 2002]

- It is for **classification** task: generate synthetic data points for feature vector (target variable values are given as the rare events).



- Torgo et al. (2013) extended the **SMOTE** for **Regression** – **SMOTER**.
- Use **relevance function** and user-specified threshold to define relevant (rare) cases and the frequent cases; Use a **weighed average of the target variable values of the two seed examples**, the weights are calculated as an inverse function of the distance of the generated new case to each of the two seed examples.
- Question - **how much oversampling is appropriate?**

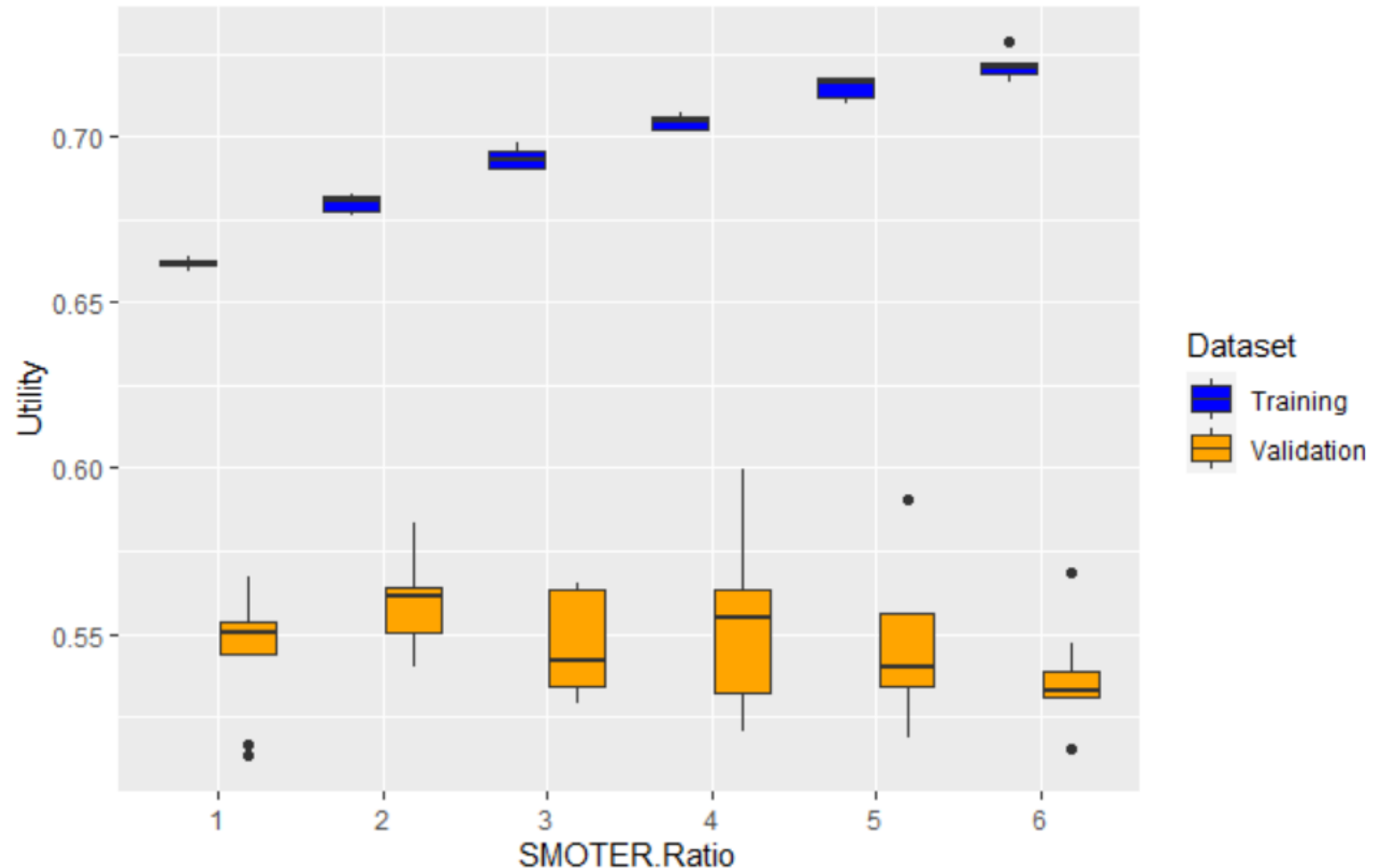
Data pre-processing – SMOTER^R

- To answer the **question about how much oversampling is appropriate**, experiments need to be performed using the **cross-validation** approach.

Simulation example:
5-fold cross-validation to
determine the optimal
oversampling ratio.

(N=600, MCAR=5%, MAR=5%,
MNAR=25%)

In this example, ratio=2 is optimal
(i.e., the minority cases will be
oversampled 2 times as the size in
the available training data). Ratio of
>2 leads to overfitting.



Utility-based Regression (UBR)

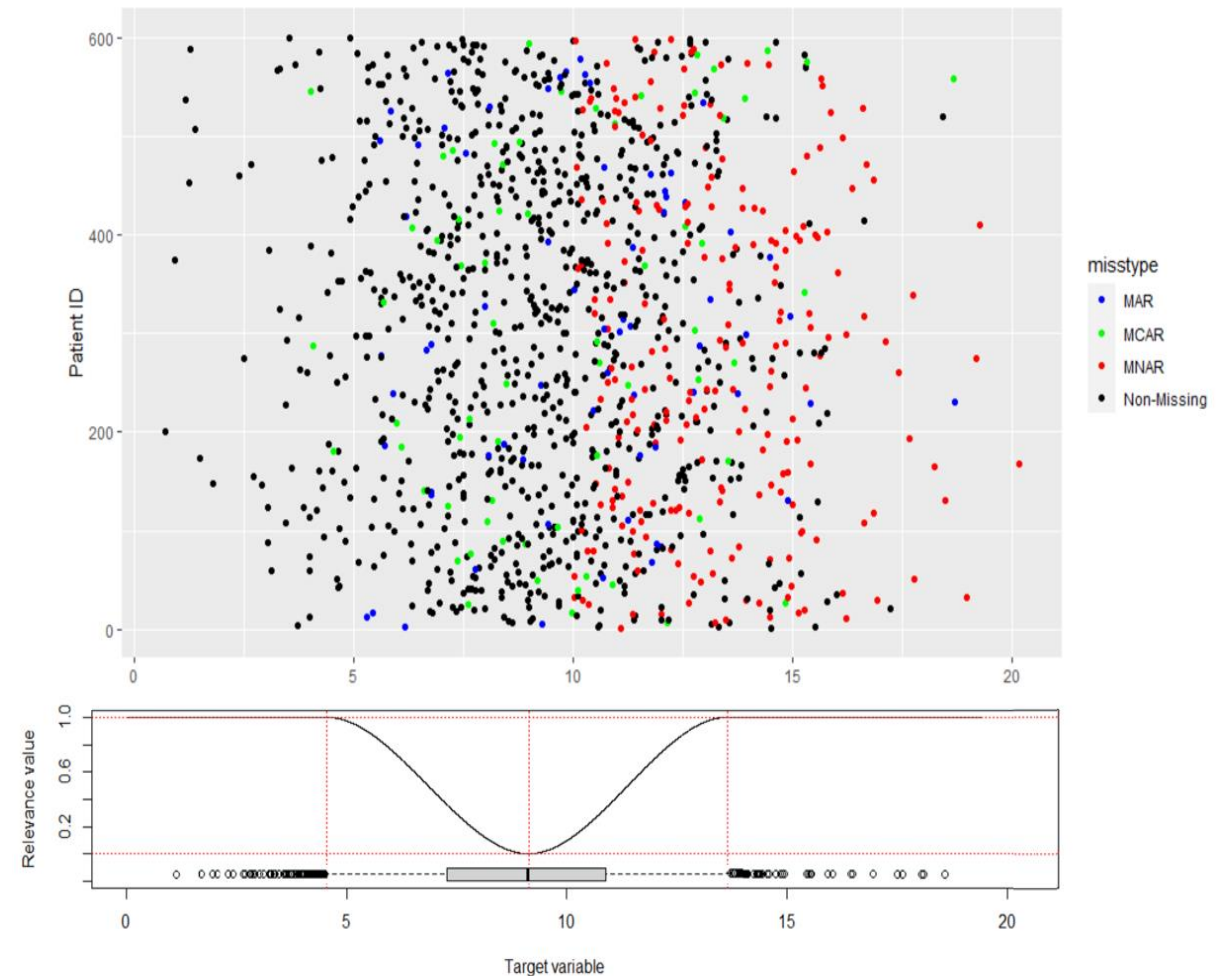
– Why UBR?

- **Standard regressions:**
 - The goal is to **minimize the error**, e.g., mean squared error (MSE) and mean absolute error (MAE).
 - **The inadequacy of standard error measures:** they are not suitable in regression problem with non-uniform costs. Their weakness lies in **taking all the prediction errors equally across the range of the target variable.**
- **UBR:**
 - **The goal is to maximize utility**, and that can only be achieved by, **simultaneously, maximizing the relevance and minimizing the error.**
 - **Utility is a function of both the error of the prediction and the relevance (importance) of target variable values.** Utility provides a more reliable estimate of a regression model, as it is **sensitive to the location of values** in the range of target variable.

UBR – Relevance function

- The relevance function $\phi(Y): y \rightarrow [0, 1]$ is a continuous function that expresses the domain-specific importance concerning the target variable y by mapping it into a $[0, 1]$ scale of relevance, where 0 represents the minimum and 1 represents the maximum.
- E.g., **Joint relevance function** for the predicted value (\bar{y}) and true value (y):
$$\phi^p(\hat{y}, y) := (1 - p)\phi(\hat{y}) + p\phi(y)$$
where p is the weight, $p \rightarrow [0, 1]$, e.g., 0.5
- Actual form is **user-defined** according to the specific problem in hand.

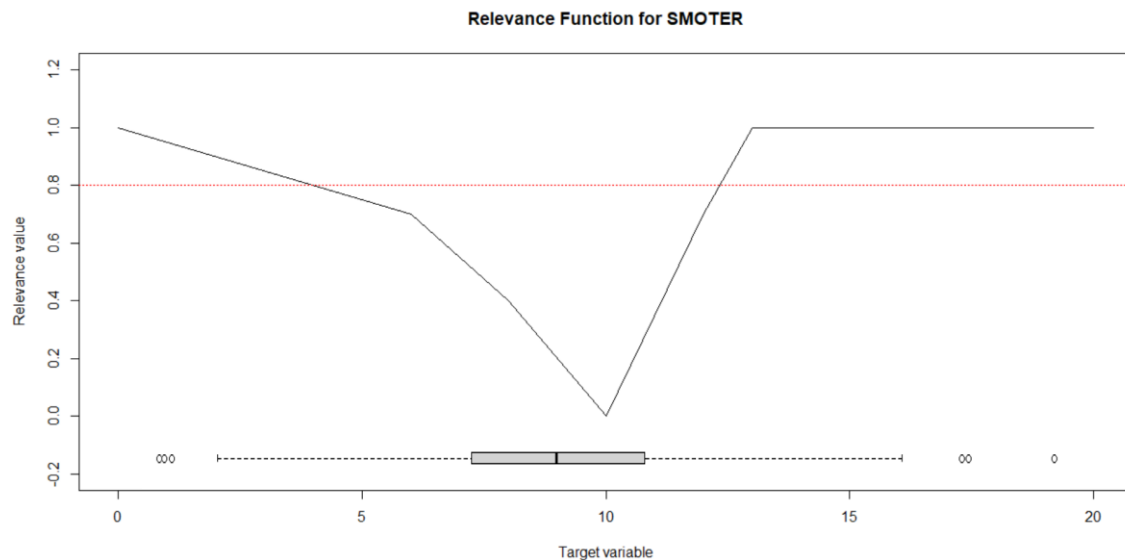
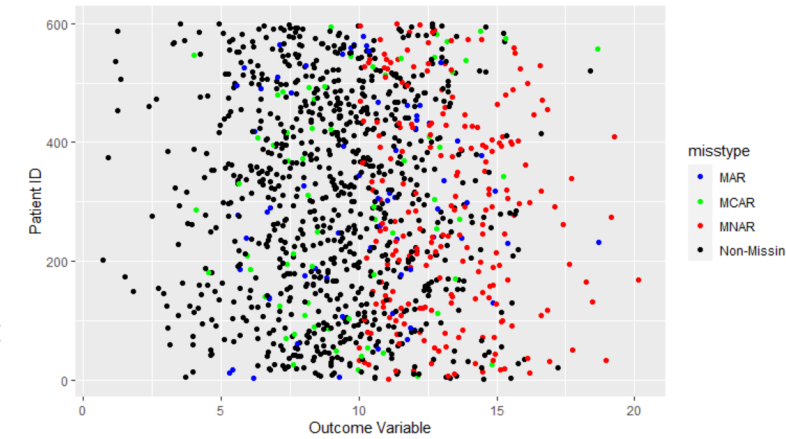
Note: considering MNAR in the area of high values and the MAR spread out across the whole range of target variable, important to assign more relevance for both high and low extreme values, which is also to avoid disproportionately heavy in one tail over the other in the prediction.



Example: assigning more importance to the extreme values according to the boxplot

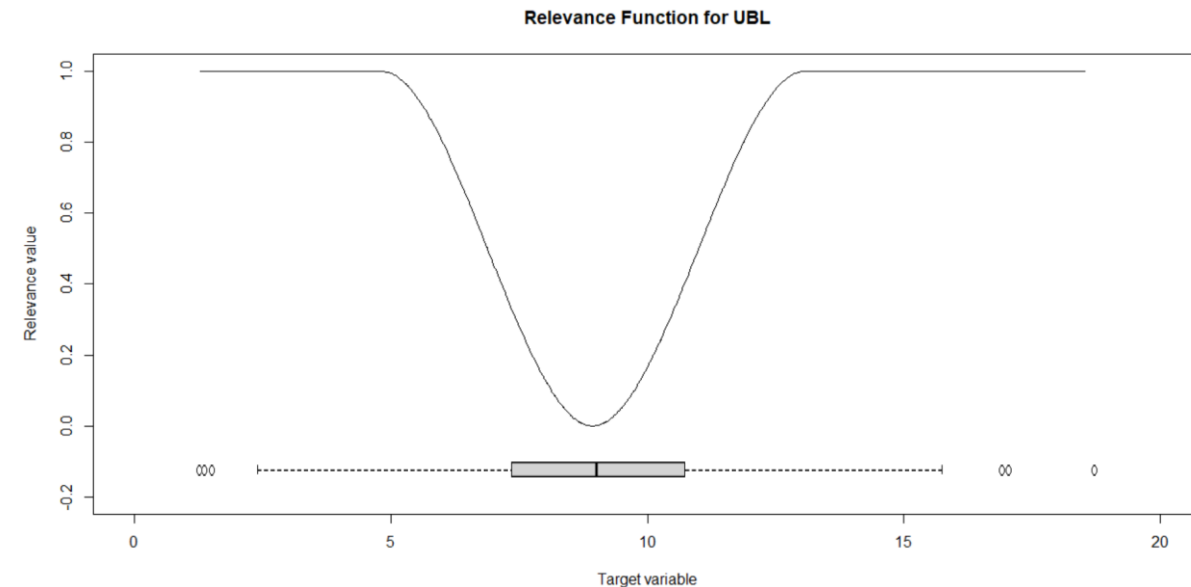
Relevance function

- Actual form of the relevance function is **user-defined**.
- Two ways of defining the relevant function in “UBL”:
 - **method = "range"**, a user-specified matrix should be provided defining the important and unimportant values.
 - **method = "extremes"**, the distribution of the target variable is used to assign more importance to the extreme values according to the boxplot.



Example: user defined relevance, if $\Phi(y) > 0.8$ then oversampling

Important: in missing data problem, MNAR (extreme/rare values) are important, but the frequent cases are also important, as we have MAR.



Example: assigning more importance to the extreme values according to the boxplot (coefficient to be defined by user to specify how far the whiskers extend to the extreme data point).

UBR – Utility-based measures

- The **cost** of a prediction is the product of the joint relevance and the loss (or error) function:

$$c(\bar{y}, y) = \phi(\bar{y}, y) C_{max} L(\bar{y}, y)$$

where $\phi(\bar{y}, y)$ is the **joint relevance function**, C_{max} is the maximum cost that is only assigned when the relevance is maximum (i.e., $\phi(\bar{y}, y) = 1$), $L(\bar{y}, y)$ is the loss (error) function, important to scale the loss function to $[0, 1]$, e.g., percentage-type error.

- The **benefit** of a prediction is the product of the relevance of true value and the complementary of the loss

$$b(\bar{y}, y) = \phi(y) B_{max} (1 - L(\bar{y}, y))$$

where B_{max} is a maximum reward, $\phi(y)$ is the relevance function of true values

- The **utility** of a prediction is the net balance between its benefit and cost:

$$U(\bar{y}, y) = b(\bar{y}, y) - c(\bar{y}, y)$$



UBR – Optimize the utility and prediction

- In optimization process, for each case, the **maximum integral of the product of the conditional probability density function and the utility surface** will be determined. The optimal prediction for a case q is given by:

$$y^*(q) = \operatorname{argmax}[z] \int pdf(y|q).U(y, z)dy$$

- where $pdf(y|q)$ is the **conditional probability density** estimation for case q , and $U(y, z)$ is the utility evaluated on the true value y and predicted value z .
- The optimization process uses the Quantile Regression Forest (QRF) to estimate the conditional probability density (Meinshausen, 2006).
 - In QRF, trees are grown as in the standard random forests (RF) algorithm. RF provides the conditional means only. QRF keeps the value of all observations in a particular node, and assesses the conditional distribution (probability density).
- **Final prediction is the conditional means take target variable utility and conditional probability into account.**
- **Tool:** R package “UBL” – Branco et al., 2017

Simulation study – data generation

- **Simplify the longitudinal data problem, look at the cross-sectional of longitudinal data**
- **Target variable and covariates are normal distribution; missing data indicators are binary variables** (MCAR, MAR, MNAR)
- **Simultaneously generating correlated normal and binary data**
- Suppose that X and Y follow a bivariate normal distribution with a correlation of ρ_{XY} . If X is dichotomized to produce X_D , then the resulting correlation between X_D and Y can be given as δ_{XDY} (point-biserial correlation, Demirtas and Dogana, 2012):

$$\delta_{XDY} = \rho_{XY} \left(\frac{h}{\sqrt{pq}} \right)$$

where p and q are the proportions of the observations above and below the point of dichotomization, respectively, and h is the PDF of the normal curve at the same point.

Tool: R package “BinNor”

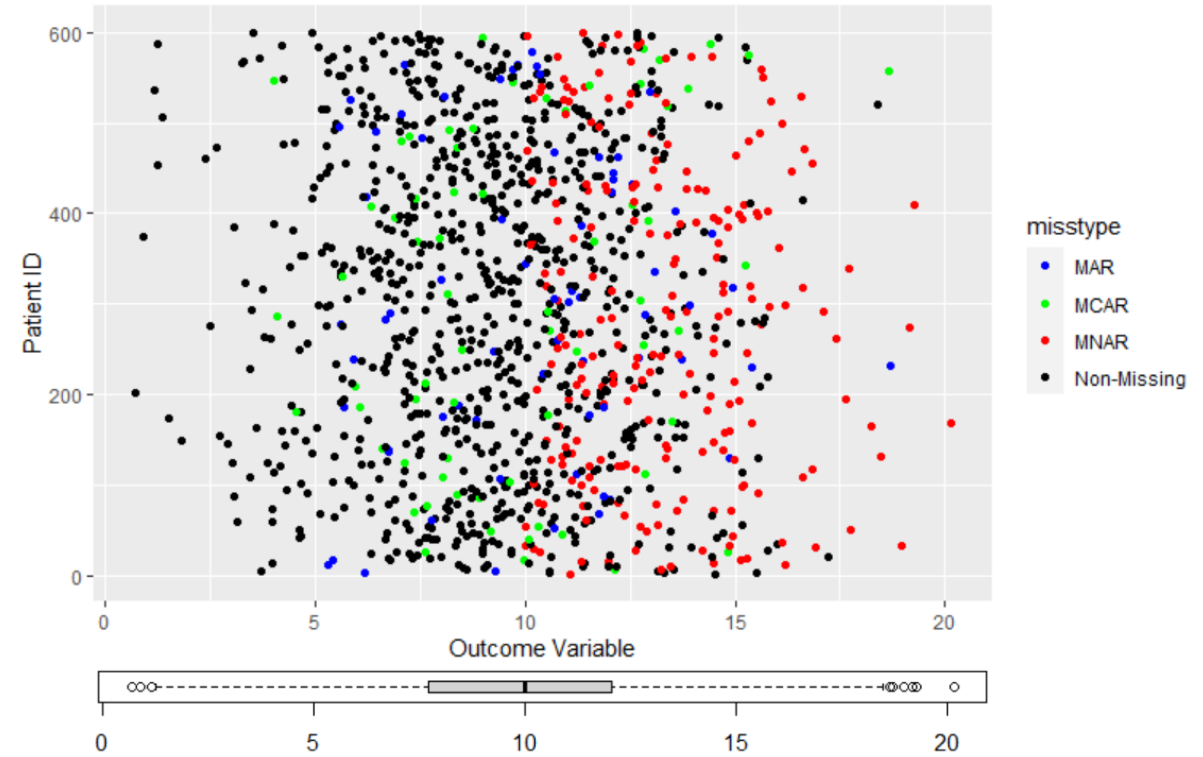
Correlation matrix in simulation

	MCAR	MAR	MNAR	OUT-COM	COV1	COV2	COV3	COV4	COV5	COV6	COV7
MCAR	1	0	0	0	0	0	0	0	0	0	0
MAR	0	1	0	0	0.4	0	0	0	0	0	0
MNAR	0	0	1	0.5	0	0.2	0.2	0.2	0.2	0.2	0.2
OUTCOME	0	0	0.5	1	0	0.5	0.5	0.5	0.5	0.5	0.5
COV1	0	0.4	0	0	1	0	0	0	0	0	0
COV2	0	0	0.2	0.5	0	1	0.2	0.2	0.2	0.2	0.2
COV3	0	0	0.2	0.5	0	0.2	1	0.2	0.2	0.2	0.2
COV4	0	0	0.2	0.5	0	0.2	0.2	1	0.2	0.2	0.2
COV5	0	0	0.2	0.5	0	0.2	0.2	0.2	1	0.2	0.2
COV6	0	0	0.2	0.5	0	0.2	0.2	0.2	0.2	1	0.2
COV7	0	0	0.2	0.5	0	0.2	0.2	0.2	0.2	0.2	1

Simulation study – data generation

- One **outcome variable** ($N(10,10)$), **7 covariates** ($N(10,10)$) and **3 missingness indicator variables (Binary)** are generated simultaneously given the correlation matrix.
- The **outcome** variable is correlated with MNAR flag and 2-7th covariate (Corr.Coeff.=0.5 medium to high level correlation)
- **MCAR flag** is independent from the outcome and the co-variates (actual Corr.Coeff.=0.02 to facilitate the data generation, $P=5\%$)
- **MAR flag** is correlated with the first covariate only (Corr.Coeff.=0.4) and independent from the outcome and the other covariates ($P=5\%$)
- **MNAR flag** is positive correlated with outcome (Corr.Coeff.=0.5) and the 2-7th covariate (Corr.Coeff.=0.2 weak to medium level correlation), $P=25\%$
- The **first covariate** is correlated with MAR flag only (Corr.Coeff.=0.4)
- The **2-7th covariates** are correlated with the outcome (Corr.Coeff.=0.5), therefore they are also correlated with each other (Corr.Coeff.=0.2).

Problem!!! The estimation of mean based on non-missing data (black data points) is biased, i.e., underestimated

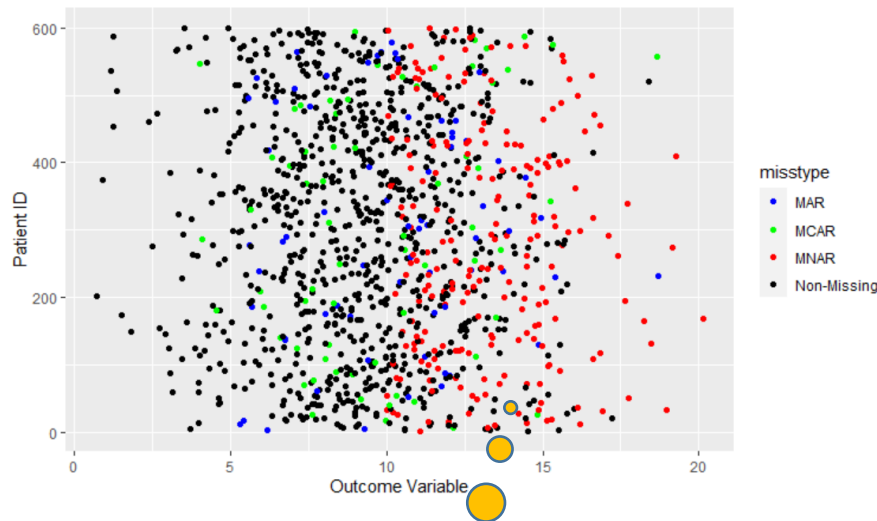


Simulation study – performance measure

- The target variable follow $N(10,10)$, consider the MNAR in one tail (i.e., high values tend to be missing).
- Imputation method:
 - **UBL with SMOTER (proposed method)**, UBL without SMOTER
 - QRF with SMOTER, QRF without SMOTER
 - RF with SMOTER, RF without SMOTER
 - MI without SMOTER
- Performance measure:
 - **Calculate the mean** of imputed target variable (non-missing data + imputed data) by different methods, and **compare** with the mean of true values (i.e., complete outcome variable before set the missing values). Close to the true mean is better.
 - **One sample t-test** ($H_0: \mu=10$): test the imputed data (by different methods separately using t-test (the larger p-value is better)).
 - **Simple linear regression** of imputed value vs. true value, compare the intercept (close to 0 is better) and slope (close to 1 is better).

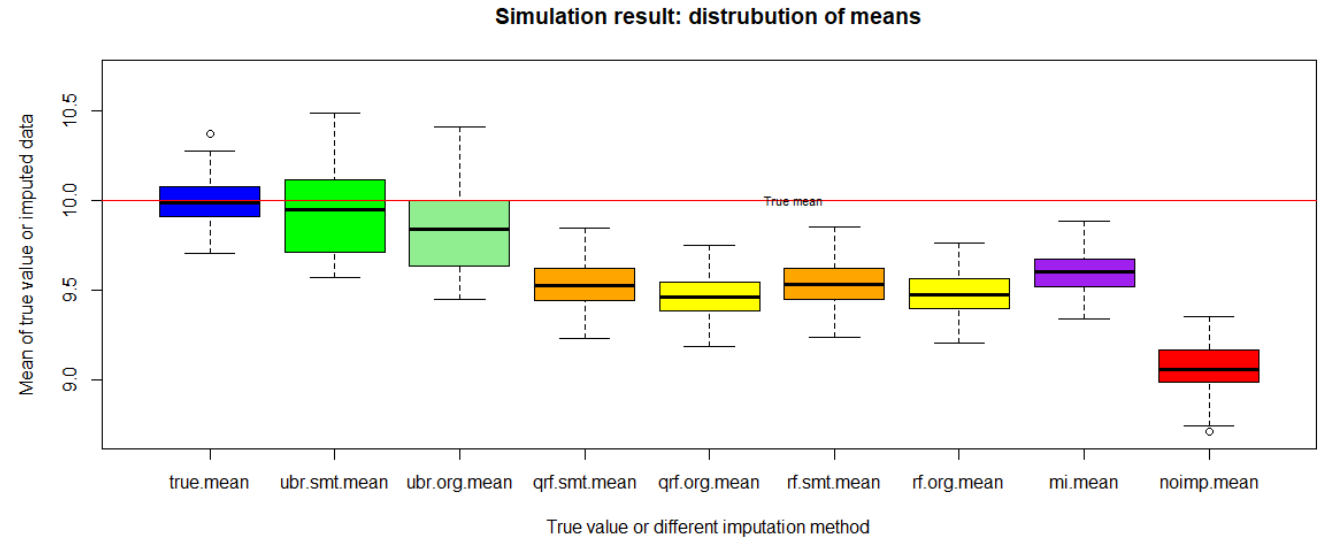
Results – performance evaluation in terms of bias and variance

from 100 replications (N=600,
MCAR=5%, MAR=5%, MNAR=25%)

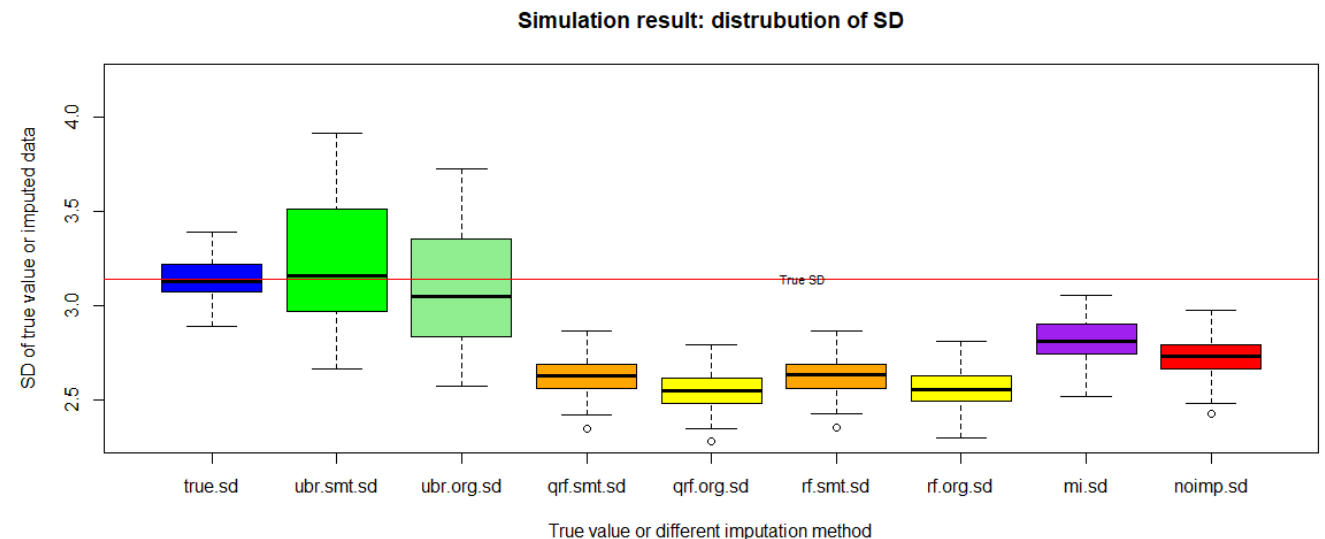


**Problem: the estimation of
mean based on non-missing
data (black data points) is
biased, i.e., underestimated**

**Compare the mean of true and imputed values (by different method):
close to 10 is better**



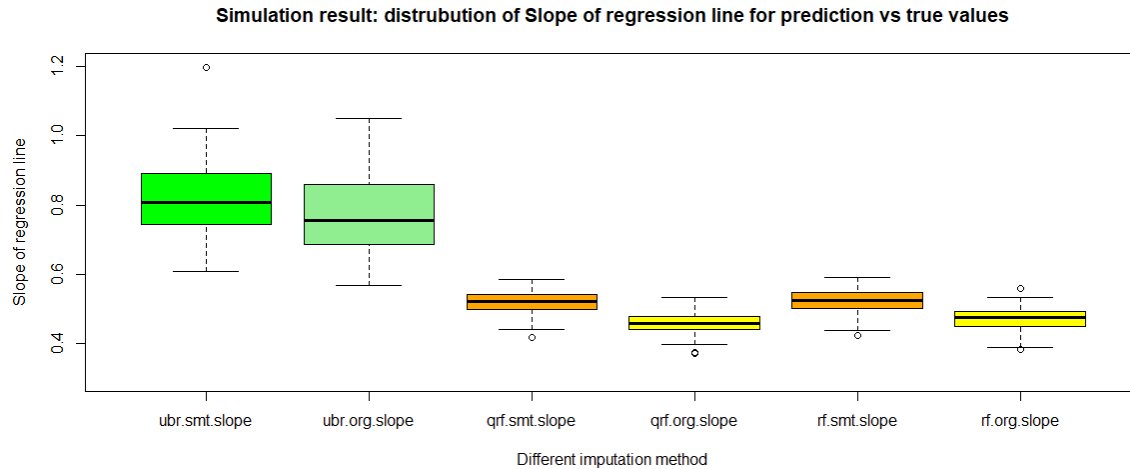
**Compare the standard deviation of true and imputed values (by different
method): close to sqrt(10) is better**



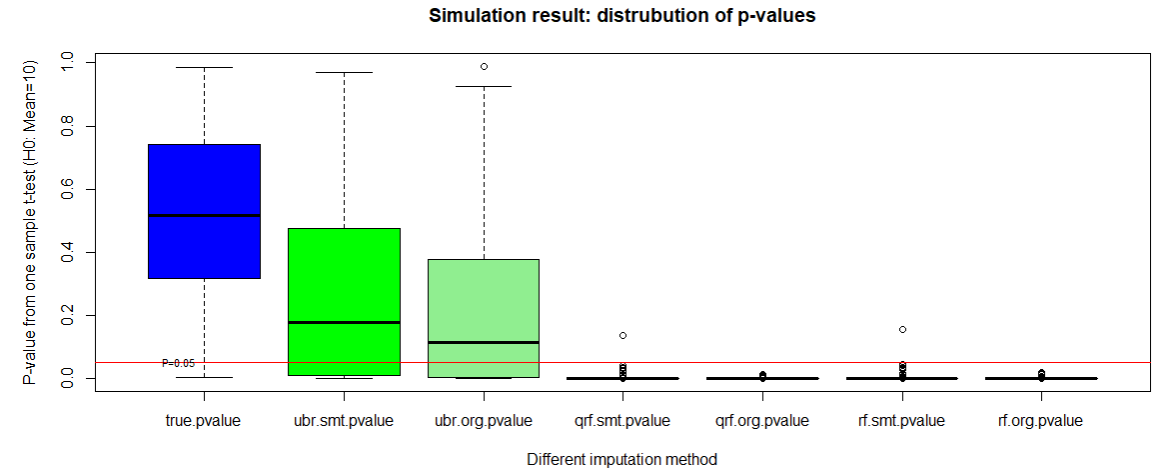
Results – performance evaluation in terms of goodness of prediction

from 100 replications (N=600, MCAR=5%,
MAR=5%, MNAR=25%)

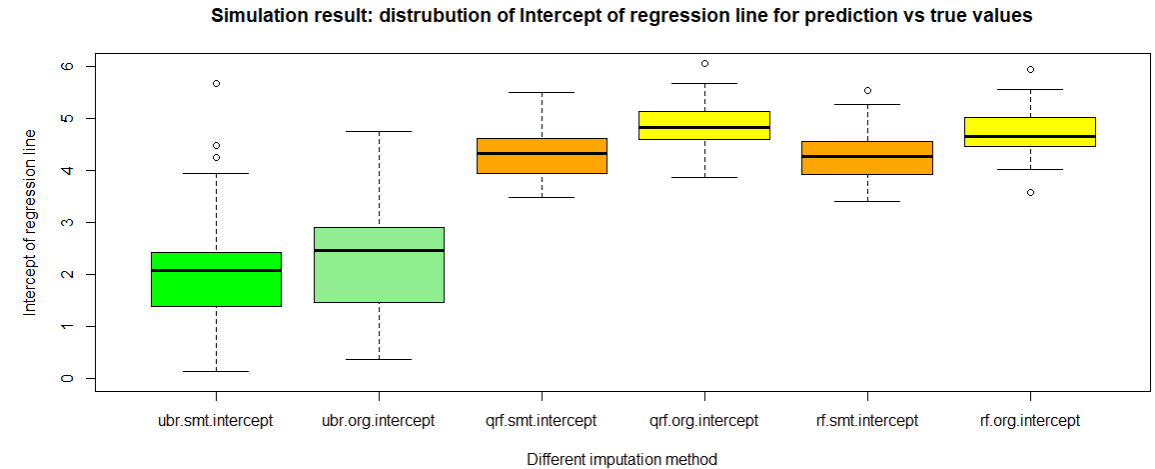
Compare the slop of the regression line for true and imputed
values (by different method): close to 1 is better



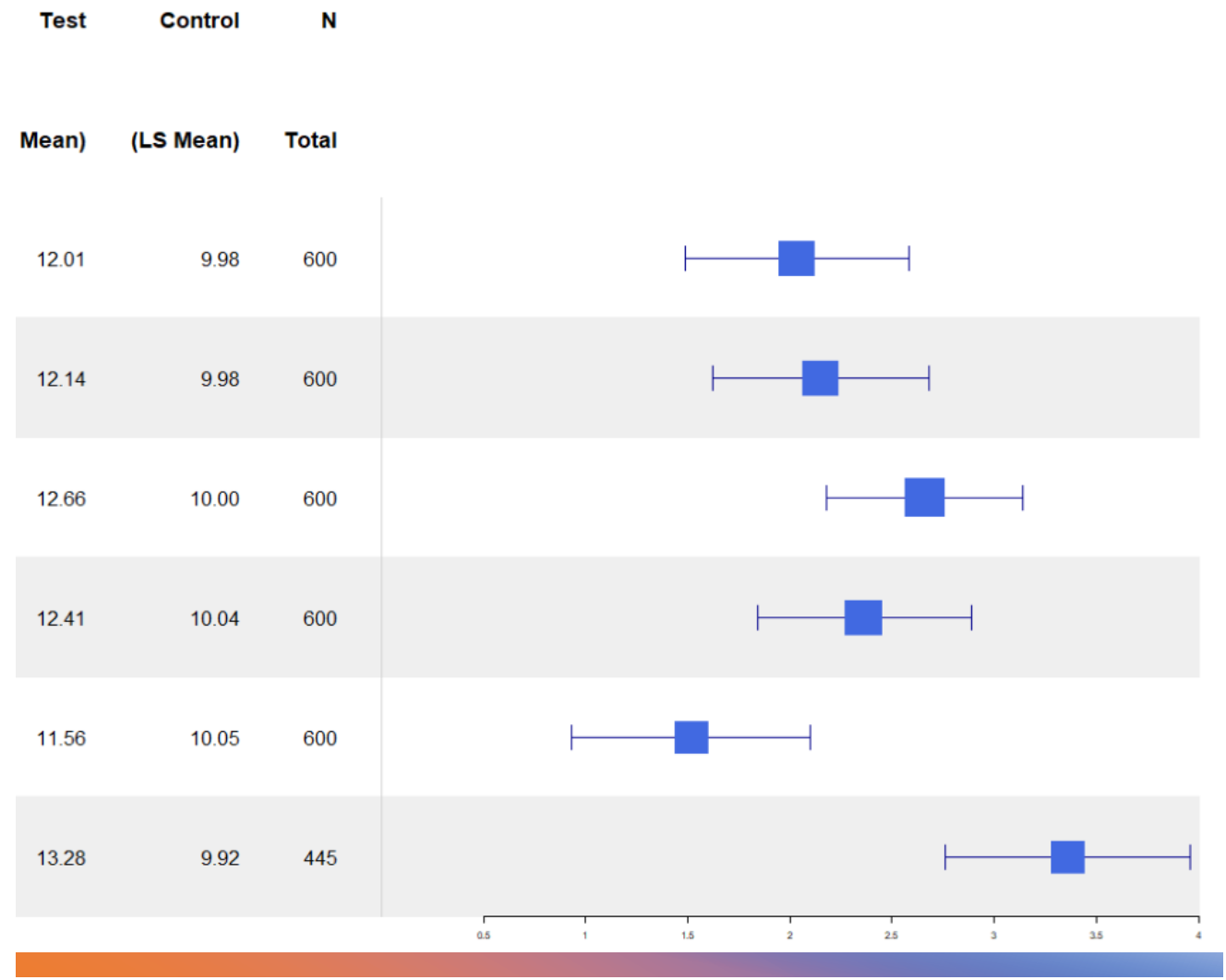
Compare the p-values of t-test for the imputed data ($H_0: \text{mean} = 10$)
by different method



Compare the intercept of the regression line for true and imputed
values (by different method): close to 0 is better



- **Practicals**
 - Relevance function
 - SMOTER
 - UBR for prediction
- R package “UBL”



UBL – Remarks

- We aim to handle the **realistic missing data scenarios** in clinical trials with continuous outcome variable.
 - **We treat MNAR as imbalanced learning task.**
 - We propose a **hybrid imbalanced learning approach that combines UBR with SMOTER**. The UBR takes both the prediction error and relevance of the target variable value into account such that the areas been assigned high relevance get more focus in the learning process. SMOTER over-samples the rare cases.
- Overall, our hybrid approach provides plausible prediction for all the MCAR, MAR and MNAR data and reduced the bias of missing data in the aggregated estimation.
- We encourage the integration of utility-based learning strategies for handling of missing data in the analysis of clinical trials.
- **Limitations** of this research include:
 - The use of some specific technical elements, such as QRF and SMOTER, is based on our current knowledge in this domain, and this can be further improved once new and better methods emerge;
 - To demonstrate the basic idea of UBR, we look at the cross-sectional data only. However, in practice, missing data problem is more often in the longitudinal studies. Therefore, from practical point of view, an extension of the UBR in the longitudinal setting is necessary.

Key References

- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons, Inc
- Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–215. <https://doi.org/10.1214/ss/1009213726>.
- Breiman, L. Random Forests, *Mach Learn* 45 (1) (2001) 5–32, doi: 10.1023/A:1010933404324 .
- European Medicines Agency. (2011). Guideline on missing data in confirmatory clinical trials. (11–12). EMA.
- National Research Council of the National Academies. (2010). The prevention and treatment of missing data in clinical trials. (53–54). National Academies Press.
- Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88, 125–134.
- Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Journal Biometrika*, 81(3), 471–483. <https://doi.org/10.2307/2337120>.
- Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1113–1121.
- Chawla, N. V., Bowyer, K.W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(2002), 321–357. <https://doi.org/10.1613/jair.953>
- Ribeiro, R.P. (2011) *Utility-based Regression*, Dep. Computer Science, Faculty of Sciences - University of Porto, 2011 PhD thesis .
- Torgo, L., Ribeiro, R.P. (2007). Utility-Based Regression. 597–604. 10.1007/978-3-540-74976-9_63.
- Torgo, L., Ribeiro, R.P., Pfahringer, B., Branco, P. (2013) Smote for regression, in: *Progress in Artificial Intelligence*, Springer, 2013, pp. 378–389. pages .
- Branco, P., Ribeiro, R.P., Torgo, L. (2017). UBL: an R package for utility-based learning.
- Meinshausen, N. (2006) Quantile Regression Forests, *J. Mach. Learn. Res.* 7
- Meinshausen, N. (2017). Quantile regression forests, a R package available at <https://cran.r-project.org/package=quantregforest> .
- Liaw, A., Wiener, M. (2018) in: Package “randomForest”: Breiman and Cutler’s random Forests for Classification and Regression, 4, R Development Core Team, 2018, pp. 6–10 .
- Hastie, T., Tibshirani, R., Friedman, J. (2008) *The Elements of Statistical Learning* (2nd ed). <https://web.stanford.edu/~hastie/ElemStatLearn/>.
- Haliduola, N.H., Bretz, F., Mansmann, U. (2022) Missing data imputation in clinical trials using recurrent neural network facilitated by clustering and oversampling, *Biometrical J.* 64 (5) (2022) 863–882, doi: 10.1002/bimj.2020 0 0393.
- Haliduola, N.H., Bretz, F., Mansmann, U. (2022) Missing data imputation using utility-based regression and sampling approaches. *Comput Methods Programs Biomed.* 2022;226: 107172.

Thanks!

Comments or
Questions?



Backup slides

Pros. and Cons. of the strategies in imbalanced learning

	Data pre-processing	Special-purpose learning methods	Prediction post-processing
Goal	Weighing the data space before applying any learning algorithm	change existing algorithms to provide a better fit to the imbalanced distribution	change the predictions after applying any learning algorithm
Advantages	any standard learning algorithm can then be used	very effective in the contexts for which they were designed; more comprehensible to the user	any standard learning algorithm can be used
Disadvantages	difficult to decide the optimal distribution (a perfect balance does not always provide the optimal results); the strategies applied may severely increase/decrease the total number of examples	difficult task because it requires a deep knowledge of both the learning algorithm and the target domain; often unavailable cost-benefit matrix; difficulties of using an already adapted method in a different learning system	the models do not reflect the user preferences; the model's interpretability is meaningless as they were obtained optimizing a loss function that is not in accordance with the user preference

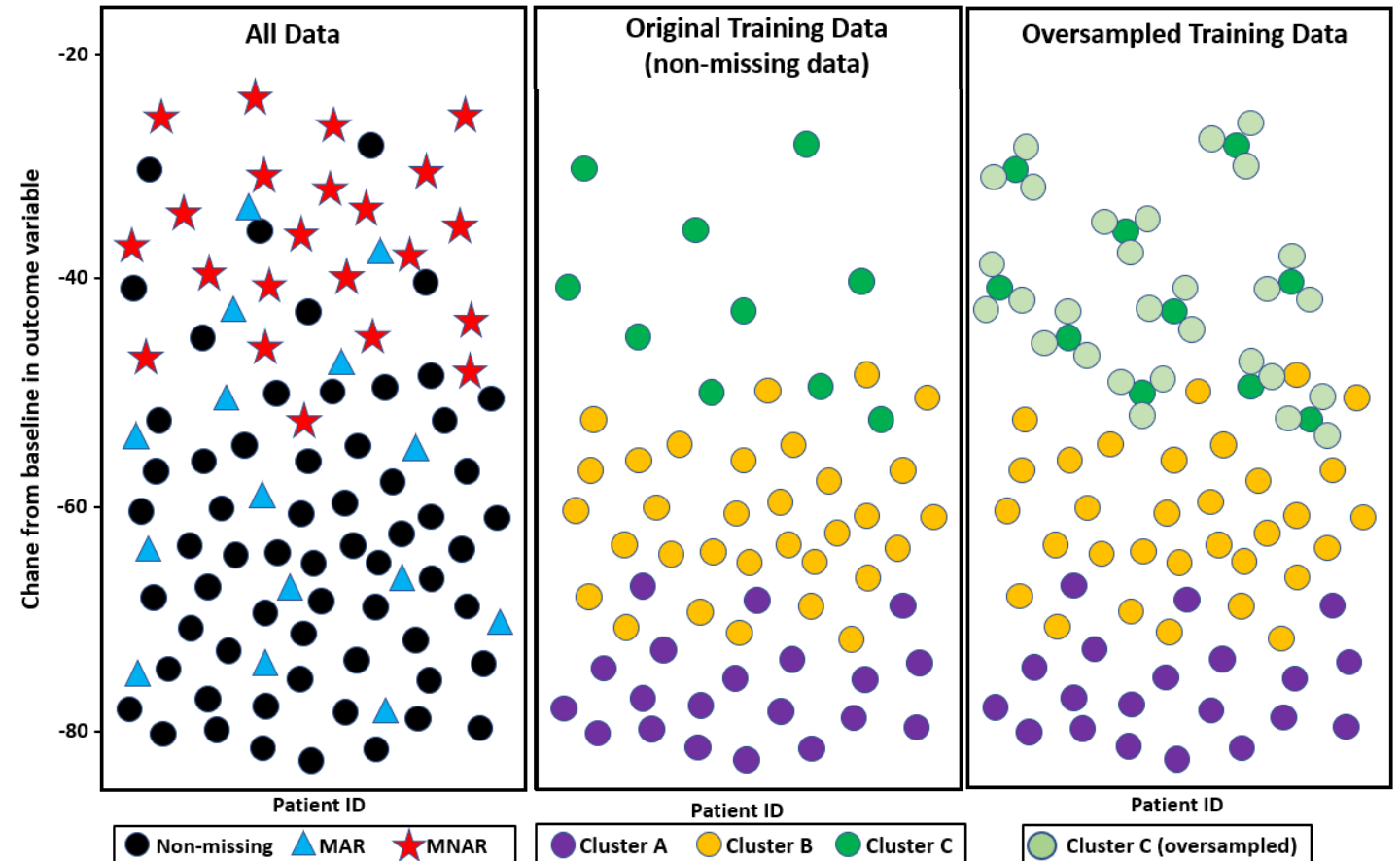
[Branco, Torgo and Ribeiro, 2016]

RNN (+ Clustering + OS) in longitudinal data

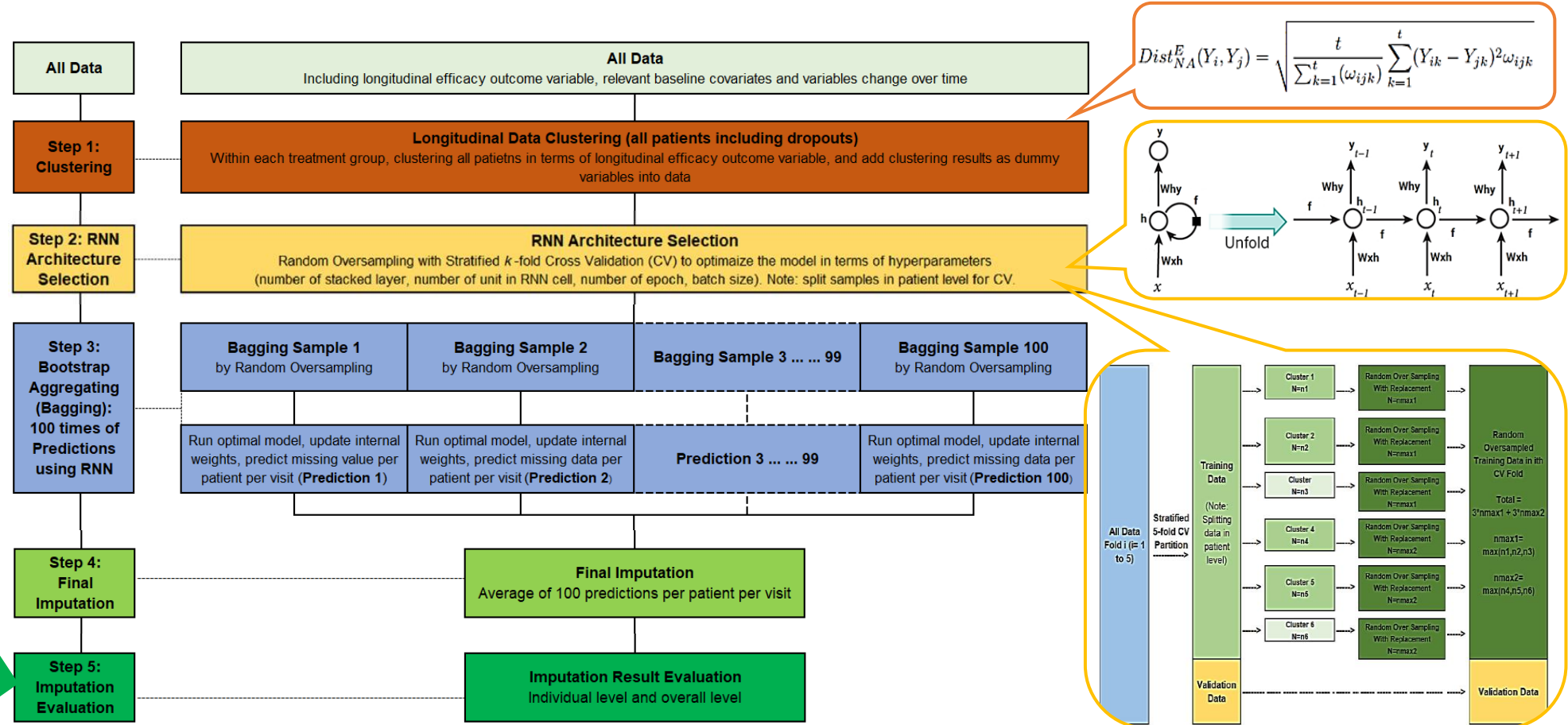
- Main idea

- The problem of MNAR is seen as an **imbalanced machine learning** task, i.e., to **oversample the minority cases** (Weiss, 2013) to **compensate for the MNAR** data in certain area. It can also handle the MAR data by looking for accurate individual information.
- To be able to use the oversampling in longitudinal continuous variable, **clustering through k-mean algorithm** (Gower, 1971) needs to be performed the first.
- **Recurrent neural networks** (RNN) model to fit the longitudinal trajectories. RNN can learn from the past to predict the future outcomes.

Illustration of clustering and oversampling for target variable



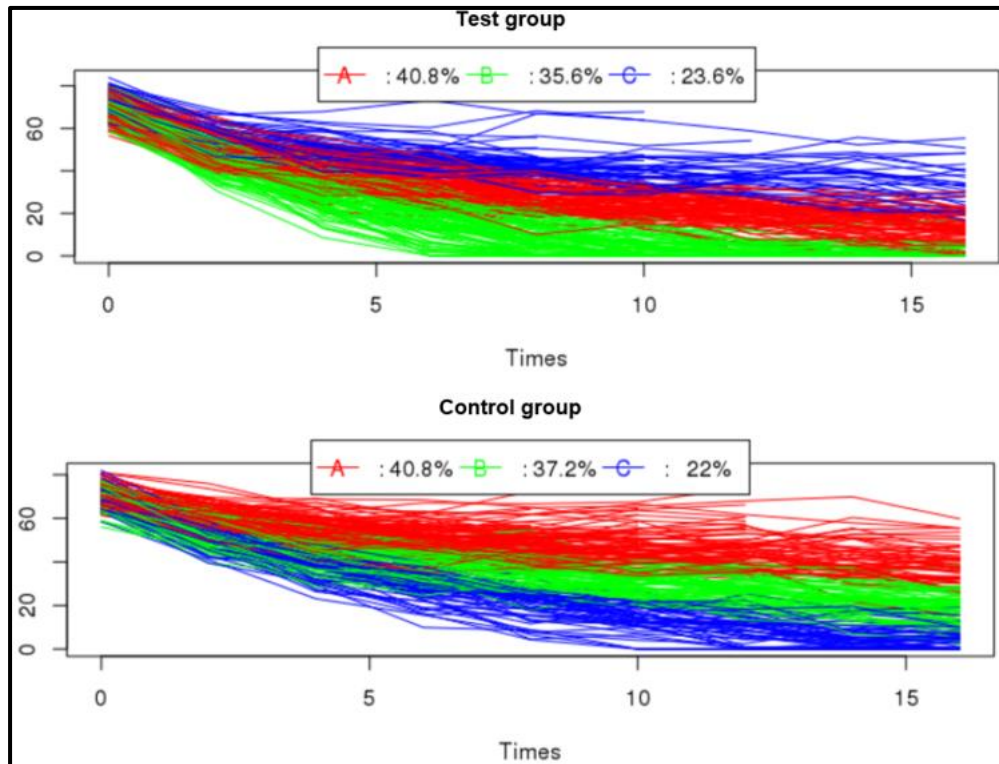
Overall workflow



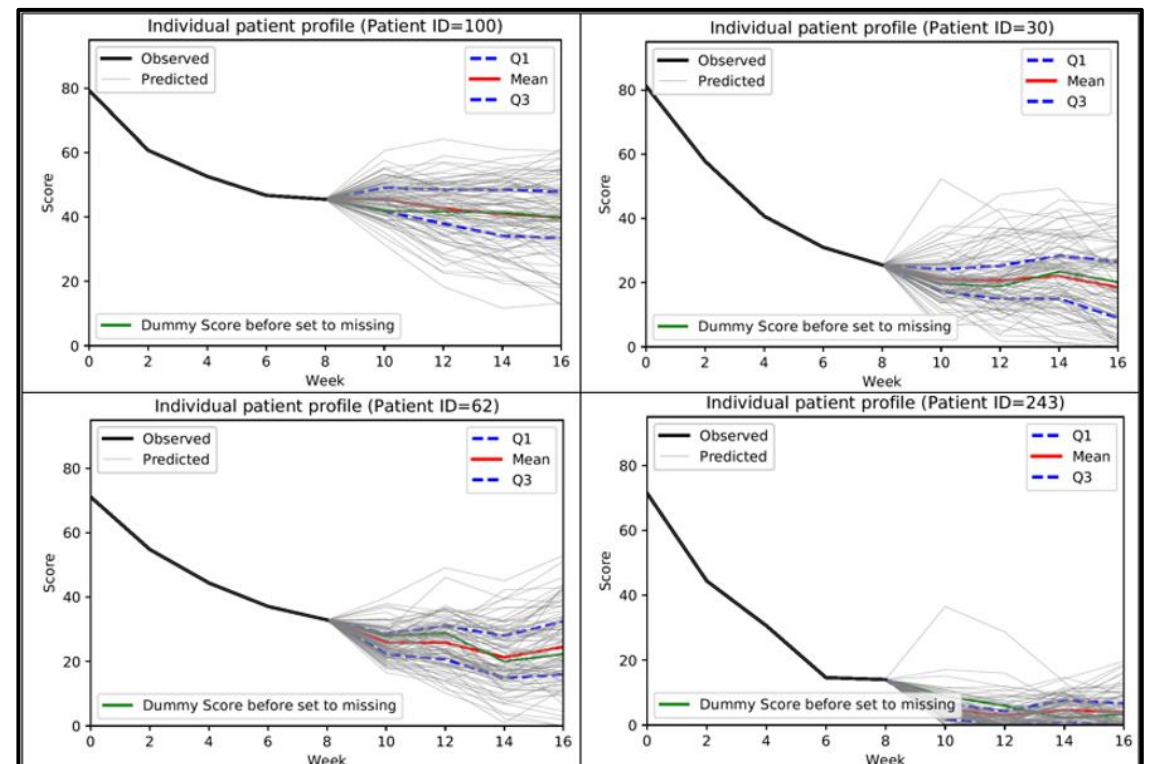
Imputation results are evaluated by comparing the estimated treatment effect with the **classic methods** for MNAR data - **joint modeling** of the response and the missingness processes: PMM (Little, 1993), SM (Little, 1995) and SPM (Little, 1995).

Results of clustering and individual prediction

-- simulation study

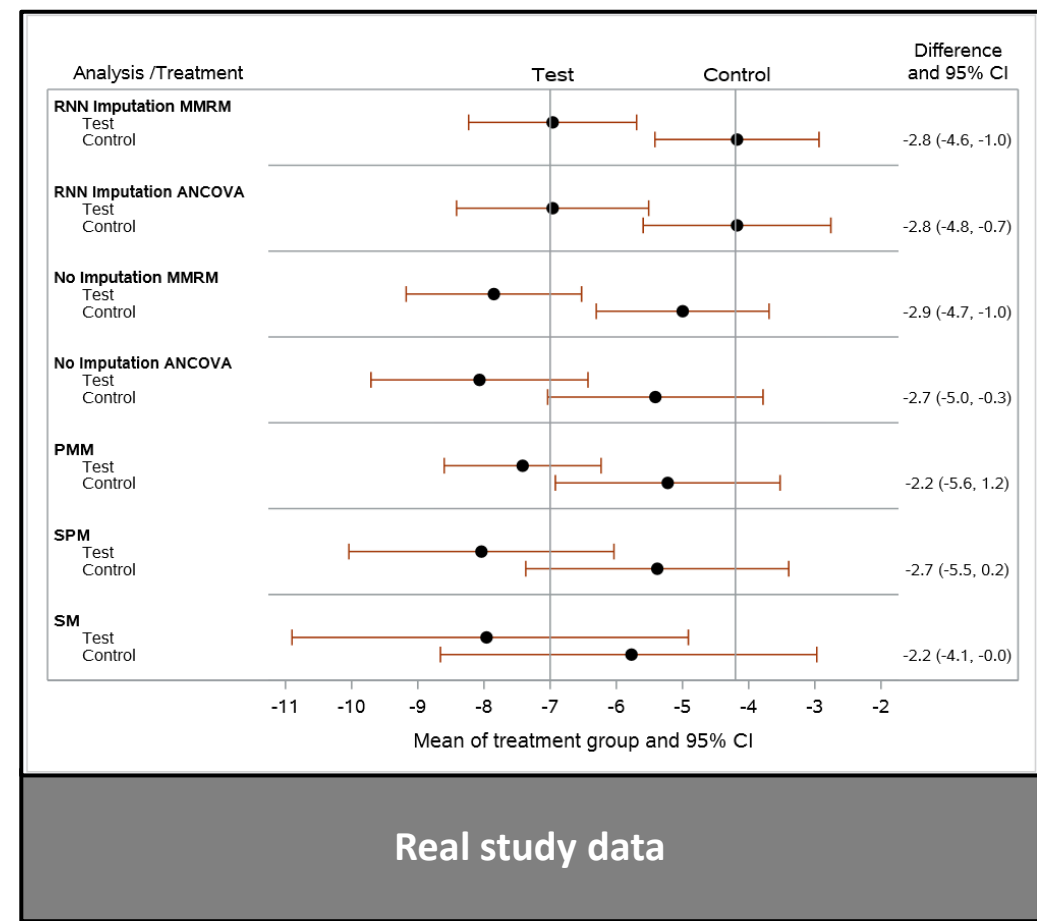
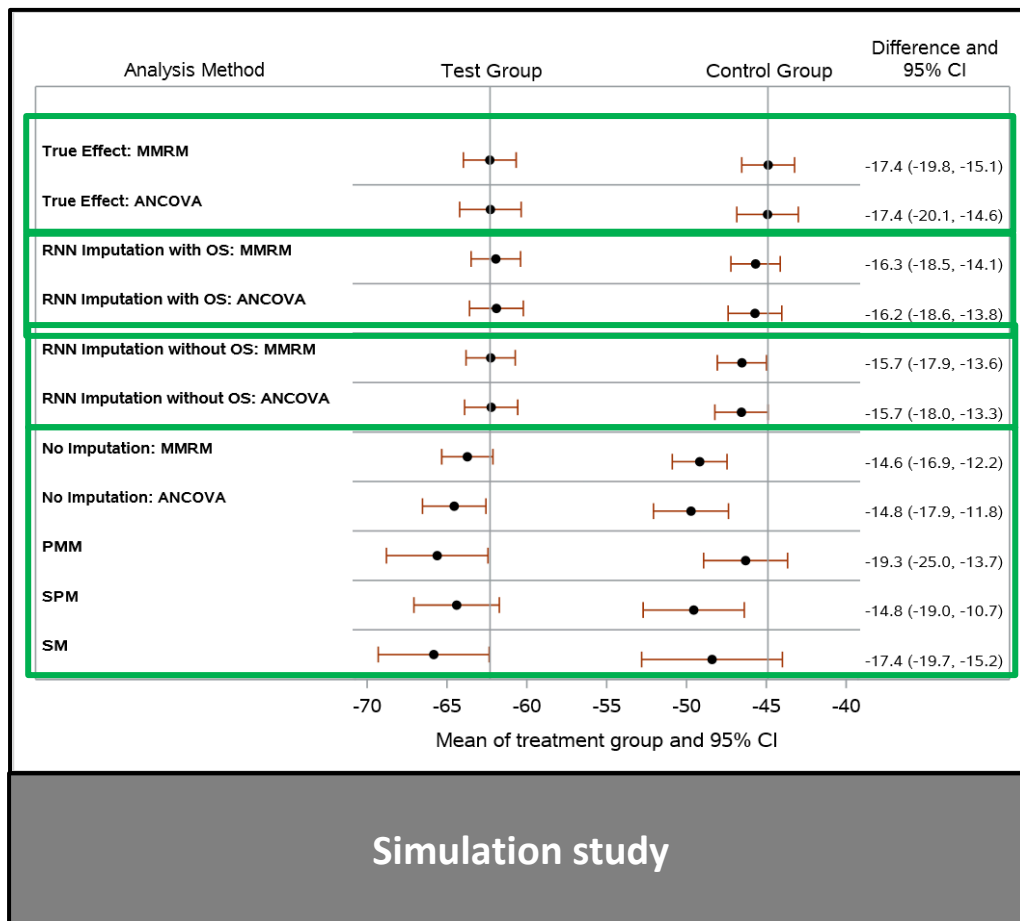


Longitudinal clustering result



Examples of individual prediction

Results of analysis using different methods



Discussions

- The computational approach **comprises necessary components** to handle the problem.
- The ***k*-mean trajectory clustering is a crucial step** in the proposed method.
- **Balancing the classes is key** to improve the prediction accuracy for the MNAR data.
- The commonly used analysis methods did not perform as well as the proposed method when both MAR and MNAR exist in one dataset.
- Therefore, this paper **offers an opportunity to encourage the integration of ML strategies for handling of missing data in the analysis of clinical trials**.
- The **limitation of this paper**: monotonous missing patterns, continuous longitudinal outcome, three cluster approach with a quite standard metric.