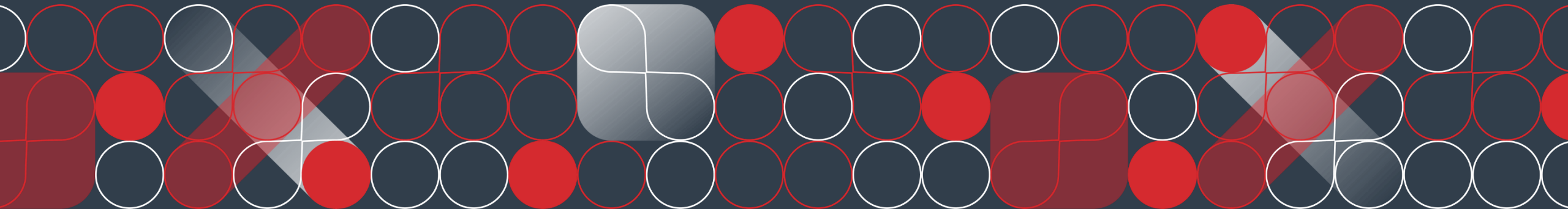
mlcov: R package for Covariate Selection Using Machine Learning

Pharmacometrics in France - 20TH SEPT. 2024

Ibtissem REBAI | Associate scientist
Anna LARGAJOLLI | Director
Floris FAUCHET | Director
Ayman AKIL | Director
Vincent DUVAL | Sr. Director
Mike TALLEY | Sr. Software engineer
James CRAIG | Staff Software engineer



Introduction & Context

Context

- Alternative to Stepwise Covariate Model-building **(SCM)** are gaining a lot of interest since few years in pharmacometrics activities and start to be accepted by regulatory agencies for the selection of covariates [1,2,3]

- The reason is that SCM is one of the most used approaches in covariate selection
  - But suffers from several weaknesses including selection of incorrect covariates and high computational time burden for complex models [4]
    - Stepwise selection of variables in regression is Evil. (freerangestats.info)

- Machine learning **(ML)** tools are increasingly studied to become new alternative to the SCM [5]

[1] Geraldine Ayral et al. "A novel method based on unbiased correlations tests for covariate selection in nonlinear mixed effects models: The COSSAC approach".
[2] Lavielle, M. (2014). Mixed effects models for the population approach: models, tasks, methods and tools..
[3] Amann, L. F., & Wicha, S. G. (2023). Operational characteristics of full random effects modelling ('frem') compared to stepwise covariate modelling ('scm')
[4] Malidi Ahamadi et al. "Operating characteristics of stepwise covariate selection in pharmacometric modeling"
[5] Emeric Sibieude et al. "Fast screening of covariates in population models empowered by machine learning

CERTARA

Context

- Previous simulation work [2] evaluating the performance of the **Boruta** algorithm implemented in R using different classifier (e.g., random forest and **XGboost**) and in combination with other penalized regression methods (i.e., **Lasso** and Elastic net)

- led us to establish a new framework for covariate selection

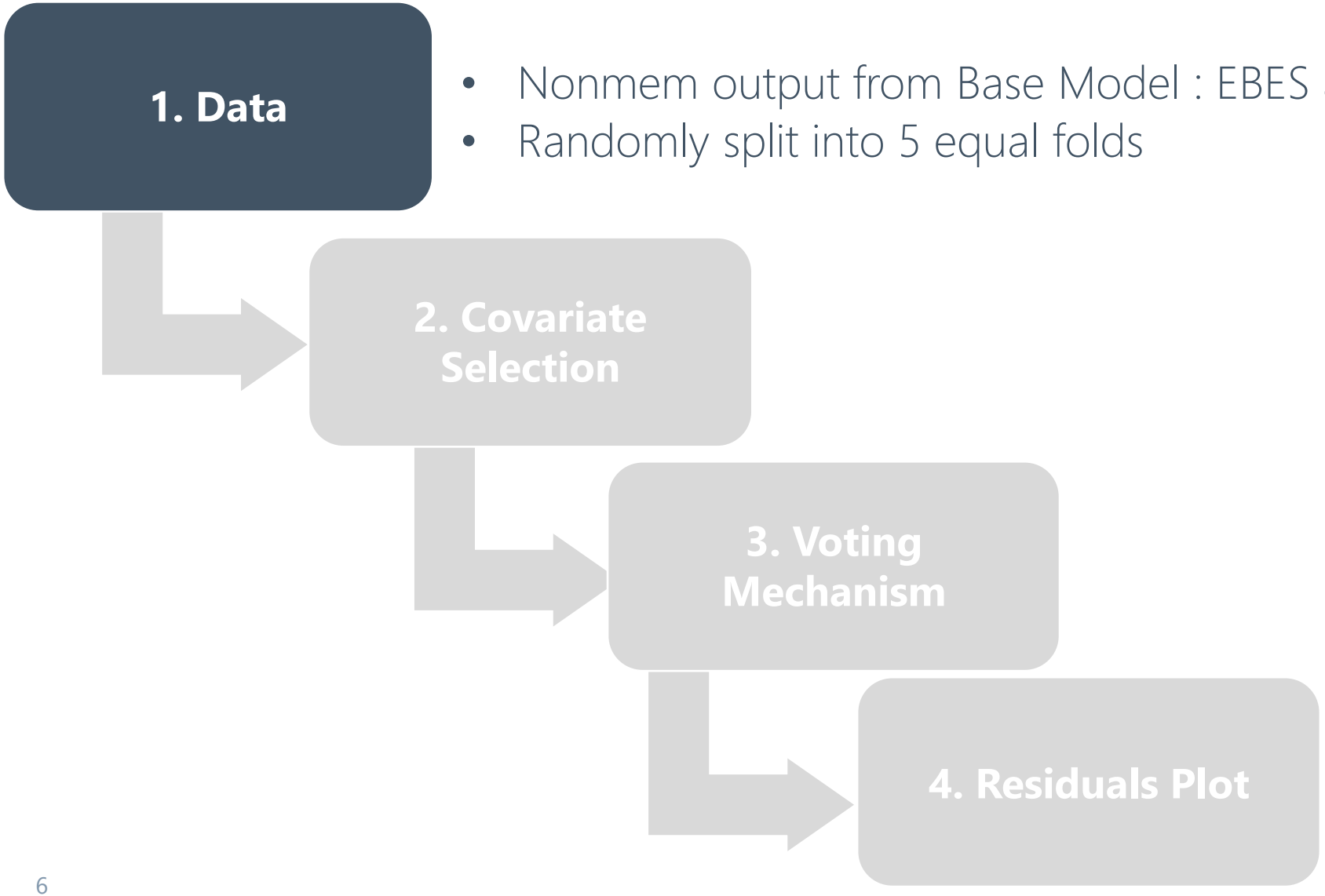
- *mlcov* is an open-source R package at the beta testing phase (https://github.com/certara/mlcov) available for the pharmacometrics community.

[1] Rebai I et al . Evaluation of the Boruta Machine Learning Algorithm for Covariate Selection. 31st PAGE meeting 2023, A Coruna, Spain, June 2023

Methods

Workflow & ML algorithms used in *mlcov*

mlcov R package : Workflow Step 1



- Nonmem output from Base Model : EBES and covariates
- Randomly split into 5 equal folds

Transformation and data splitting

mlcov applied :

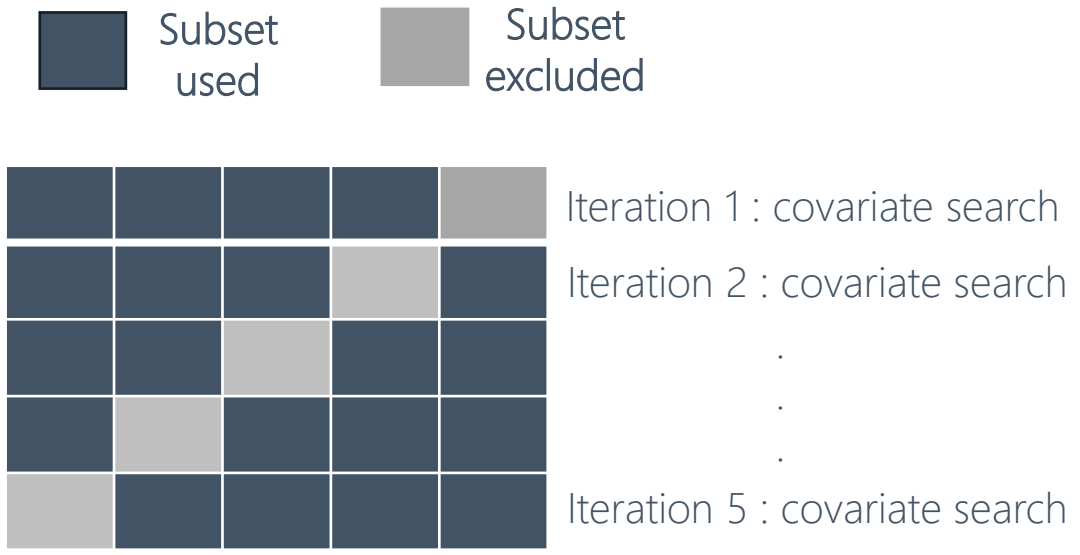
1. Log transformation of individual parameters

2. Split dataset into 5 data subsets. 4/5 of data used for covariate search

→ Process repeated 5 times with different subsets at each time

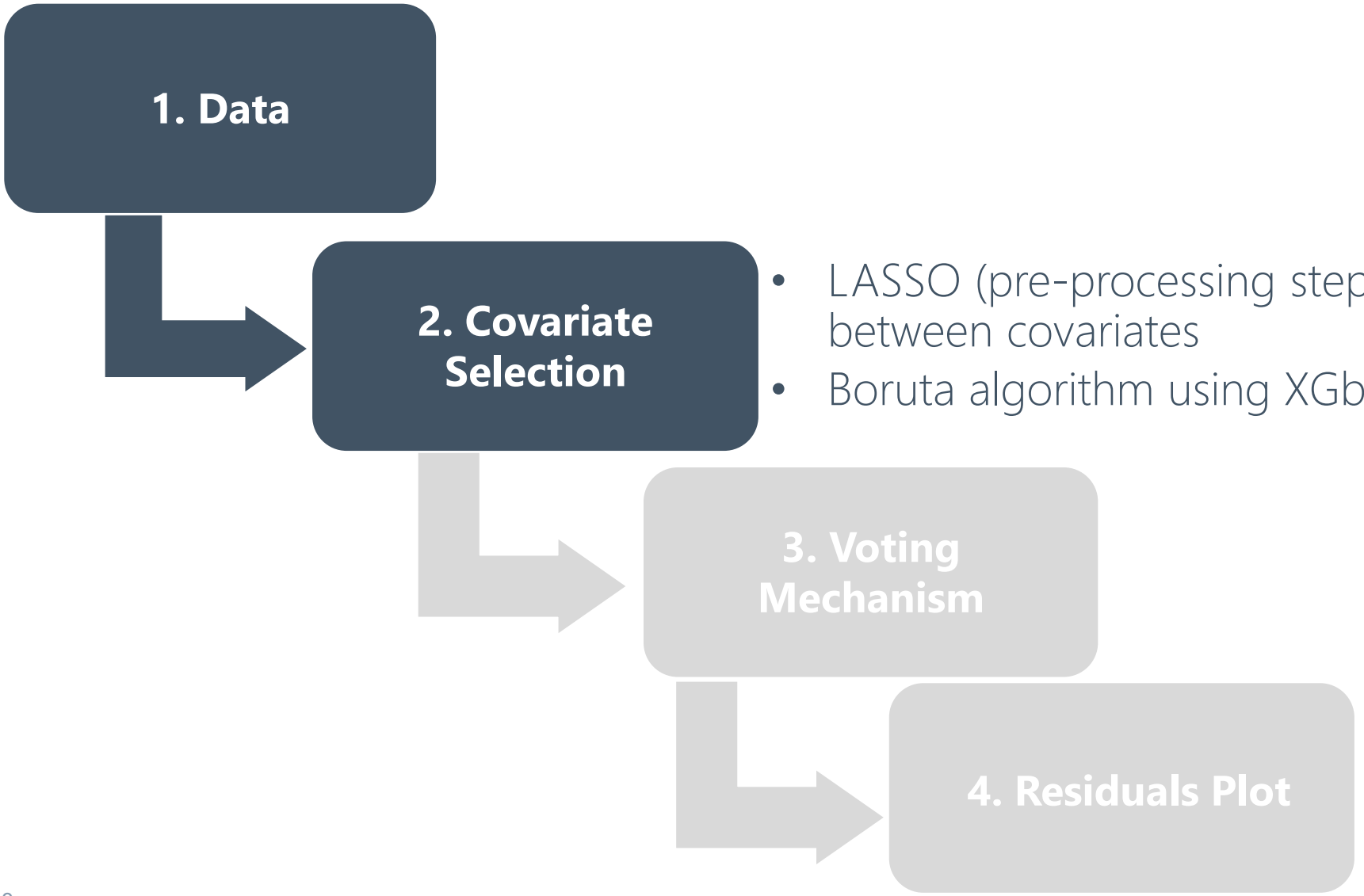
ID	CL	V	KA	WGT	ALB	SEX
1	0.41	5.9	0.30	64.2	4.7	0
2	0.52	10.0	0.58	53.8	4.5	0
3	0.39	5.6	2.70	58.1	4.4	0
4	0.75	14.0	0.19	66.7	4.3	0
5	0.34	5.7	1.10	47.7	4.3	0
6	0.59	11.0	1.50	68.1	4.2	1

splitting



CERTARA

mlcov R package : Workflow Step 2



- LASSO (pre-processing step) to manage collinearity between covariates
- Boruta algorithm using XGboost classifier

CERTARA

Penalized regression method : LASSO [1]

Covariate selection method preferred in the presence of multicollinearity between variables. This method removes variables which have weight set to 0.

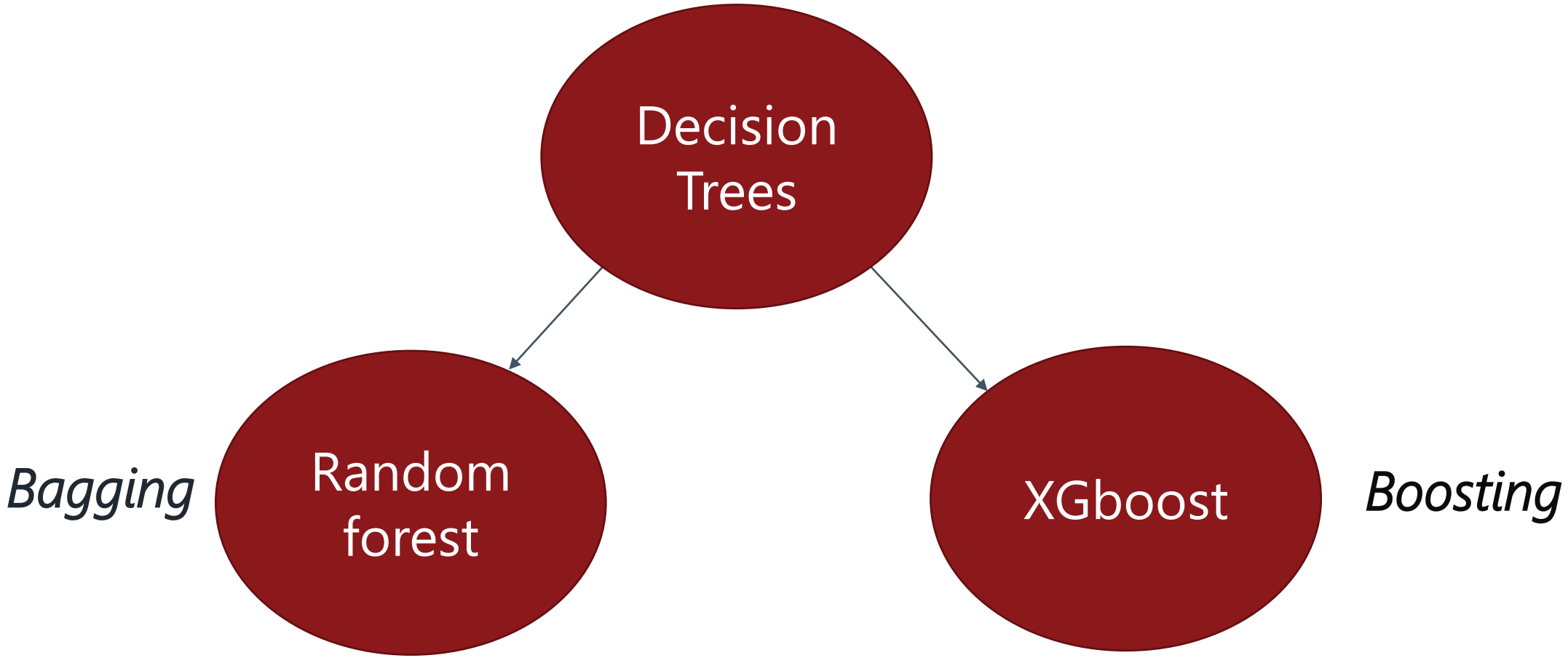
- minimizing the residual sum of squares (RSS) of the linear regression model, by imposing a constraint on the covariate coefficients.

- Objective function = RSS + $\lambda \sum|\beta|$
  - ✓L1 penalty forces some of the coefficients to be exactly equal to zero
  - ✓regularization parameter λ controls the strength of the penalty (optimized by k-fold cross validation)

[1] Robert Tibshirani. "Regression shrinkage and selection via the lasso"

Tree-Based machine learning algorithms

→ supervised learning models that address classification or regression problems by constructing a tree-like structure to make predictions.



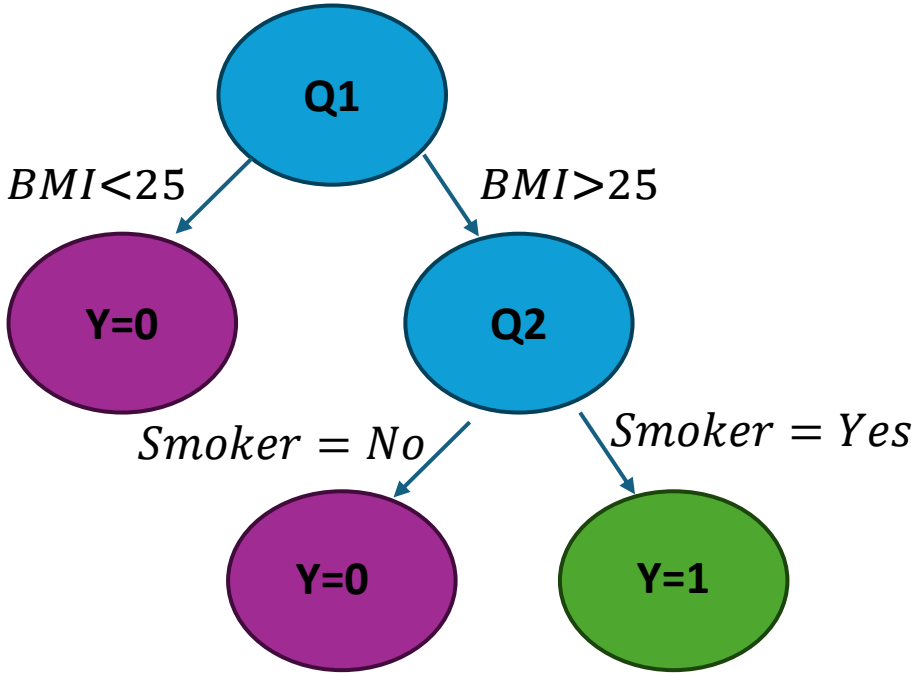
CERTARA

Decision trees

It works by splitting data into branches based on decision rules, forming a tree-like structure.

Example: observed variables are X = (BMI, SMOKER)

→ *goal : predict the variable Y which is 1 if the individual has a risk of developing a certain pathology, 0 if the individual has no risk based on results to question*

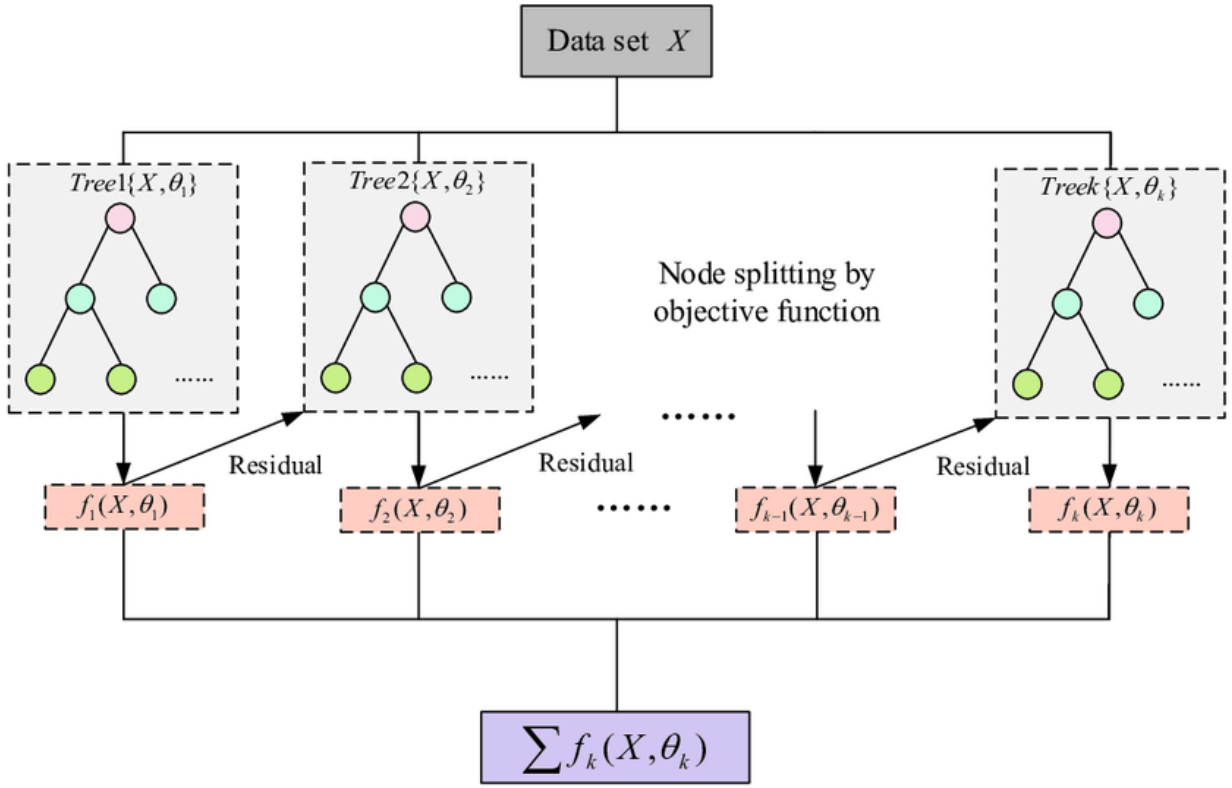


CERTARA

XGboost

Creates a series of decision trees and combines them to form a ML model
→ By fitting a decision tree to the residuals of the previous tree, allowing the model to iteratively improve its predictions.

1. Initialize the model
2. Fit additional decision trees
3. Combine the decision trees
4. Make predictions
5. Select the most important covariates based on their **importance score**

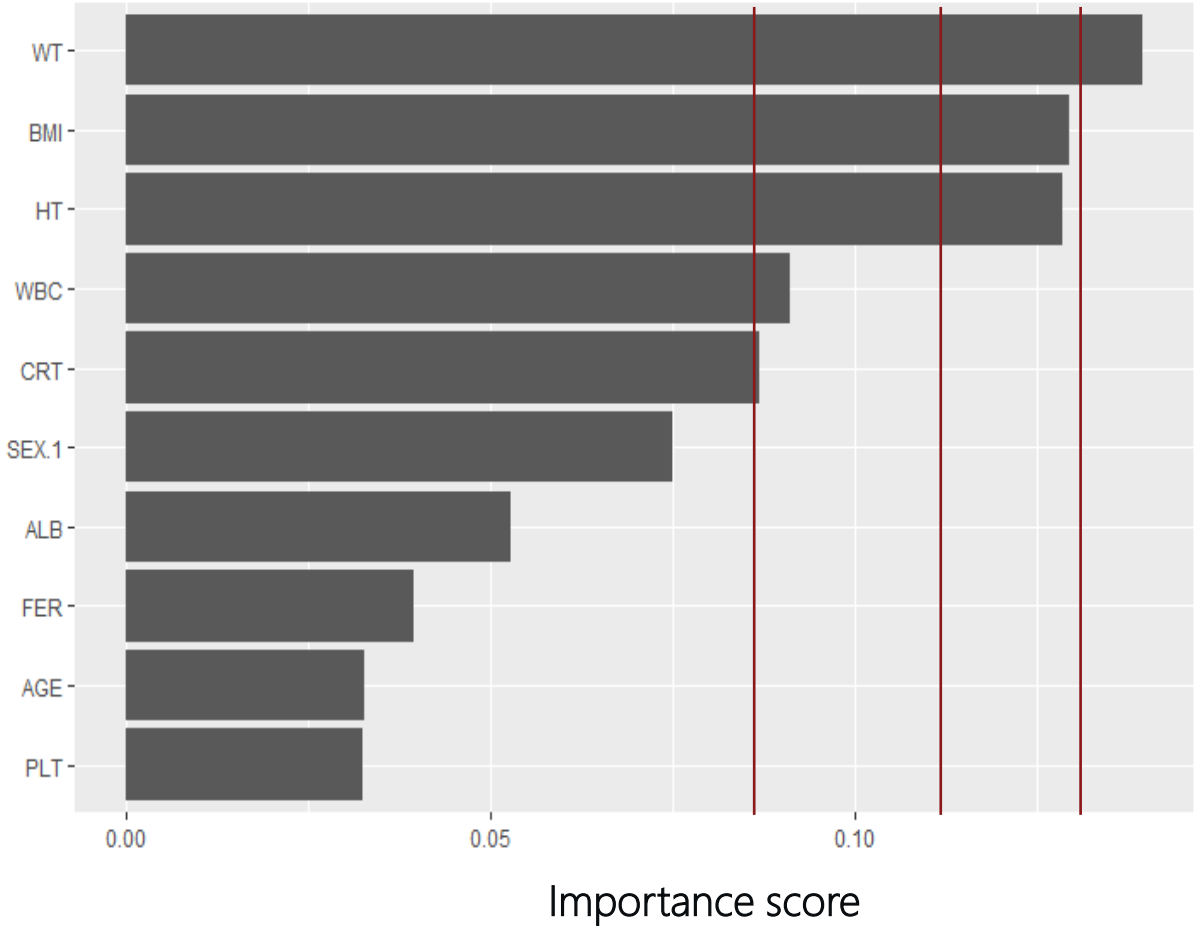


https://www.researchgate.net/publication/345327934_Degradation_state_recognition_of_piston_pump_based_on_ICEEMDAN_and_XGBoost#pf6

CERTARA

Importance Score

Indicates how useful or valuable each covariates was in the construction of the boosted decision trees within the model



What is the threshold to select covariate?

BORUTA

CERTARA

BORUTA : Feature selection algorithm [1]

Identify significant covariates by comparing observed one to random ones created

1. The first idea: shadow features

	age	height	weight	shadow_age	shadow_height	shadow_weight
0	25	182	75	51	176	75
1	32	176	71	32	182	71
2	47	174	78	47	168	78
3	51	168	72	25	181	72
4	62	181	86	62	174	86

- Duplicates the set of explanatory variables and shuffle the values in each column: **shadow features**

- Trains our **XGboost** classifier: Ensure an idea of the **importance** for each of the covariates (original and shadow)

- Take the importance of each original covariates and compare it with a threshold → **threshold** defined as the highest covariates importance recorded among the shadow features

	age	height	weight	shadow_age	shadow_height	shadow_weight
feature importance %	39	19	8	11	14	9
hits	1	1	0	-	-	-

[1] Miron B Kursa et al. "Boruta–a system for feature selection". (2010

CERTARA

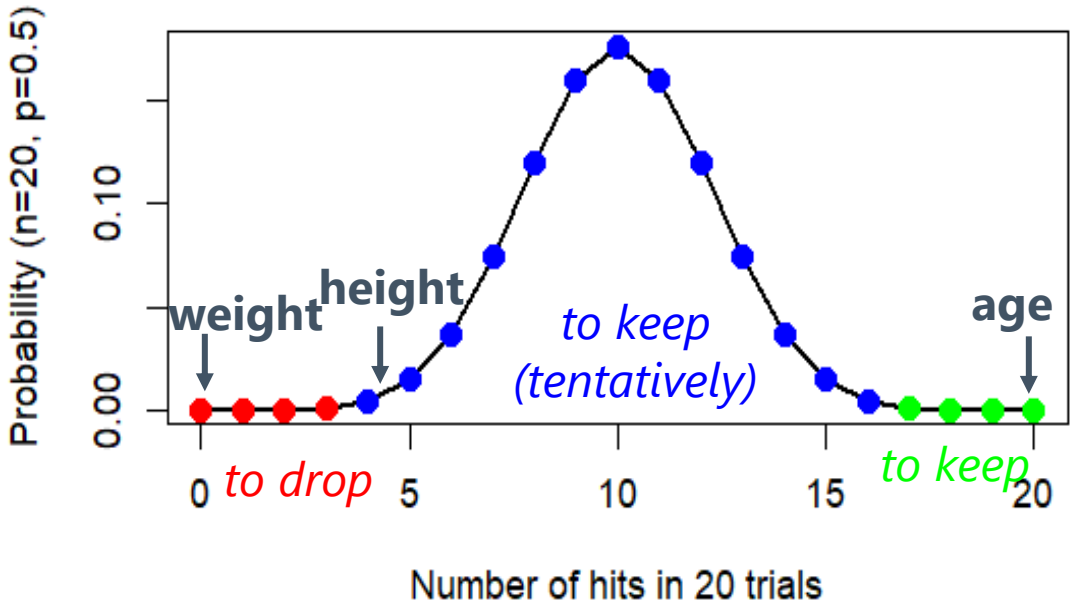
BORUTA : Feature selection algorithm

2. The second idea: binomial distribution

- As often happens in machine learning, the key is **iteration**

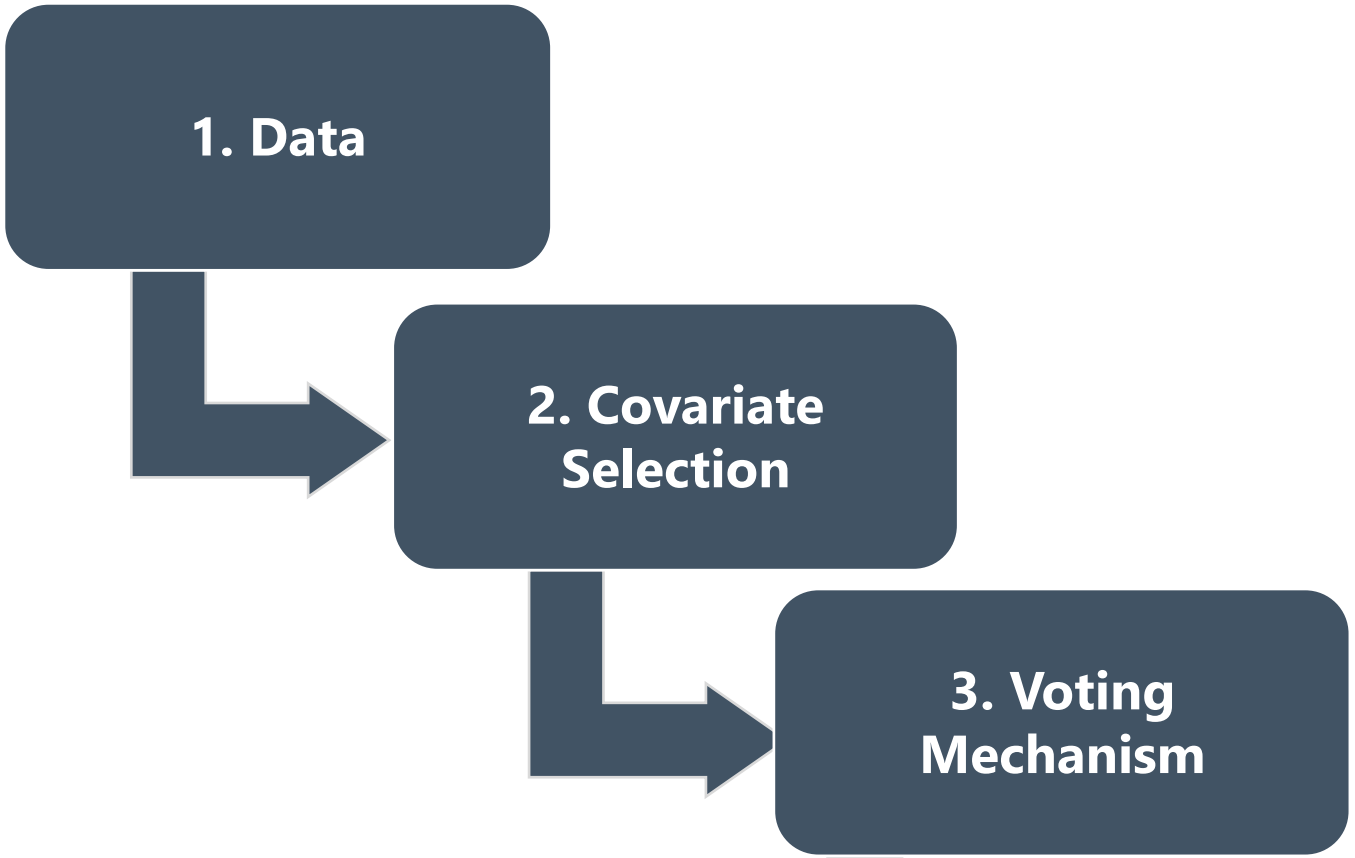
	age	height	weight
hits (in 20 trials)	20	4	0

- What is the **probability** that we shall keep a covariate?



CERTARA

mlcov R package : Workflow Step 3

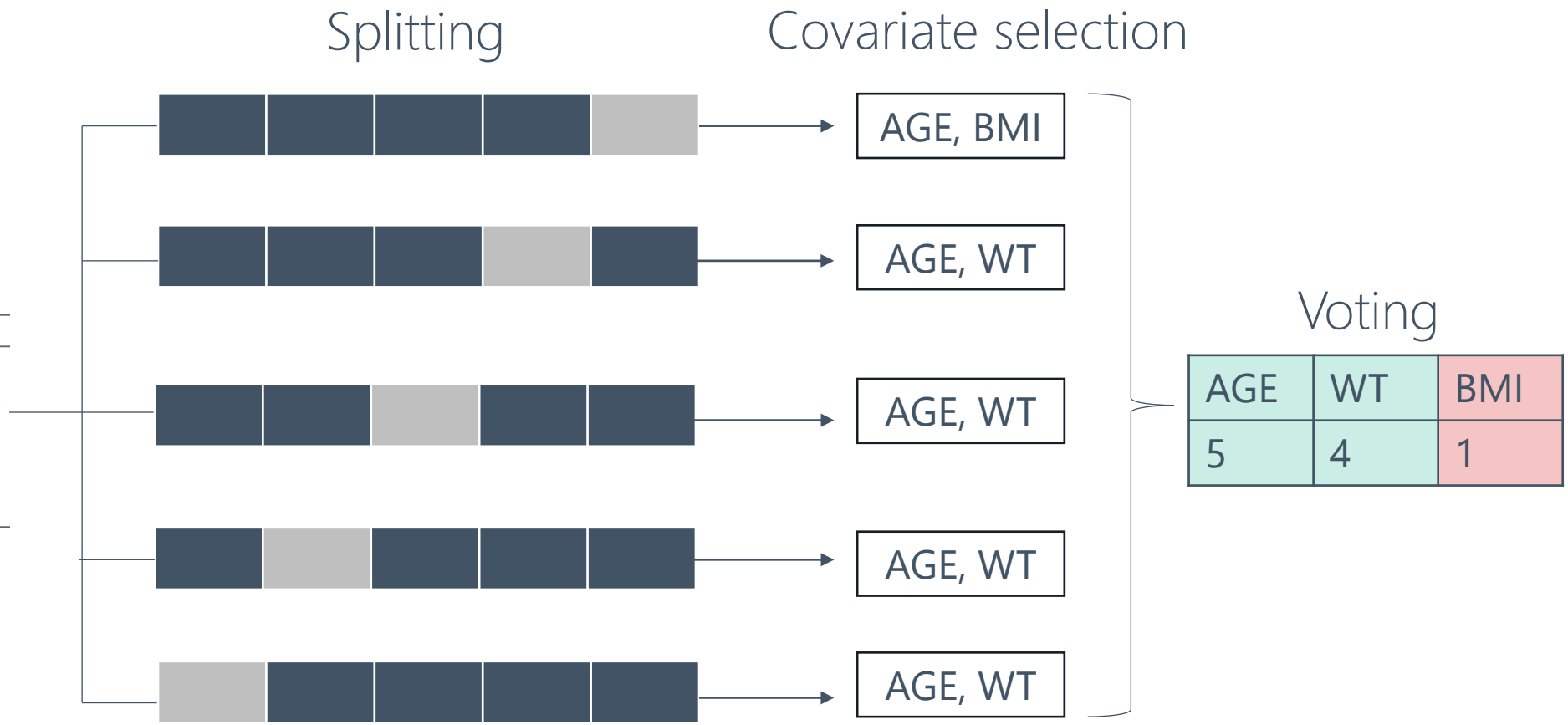


- Counted covariates selected in 5 folds
- Covariates chosen ≥ 2times

CERTARA

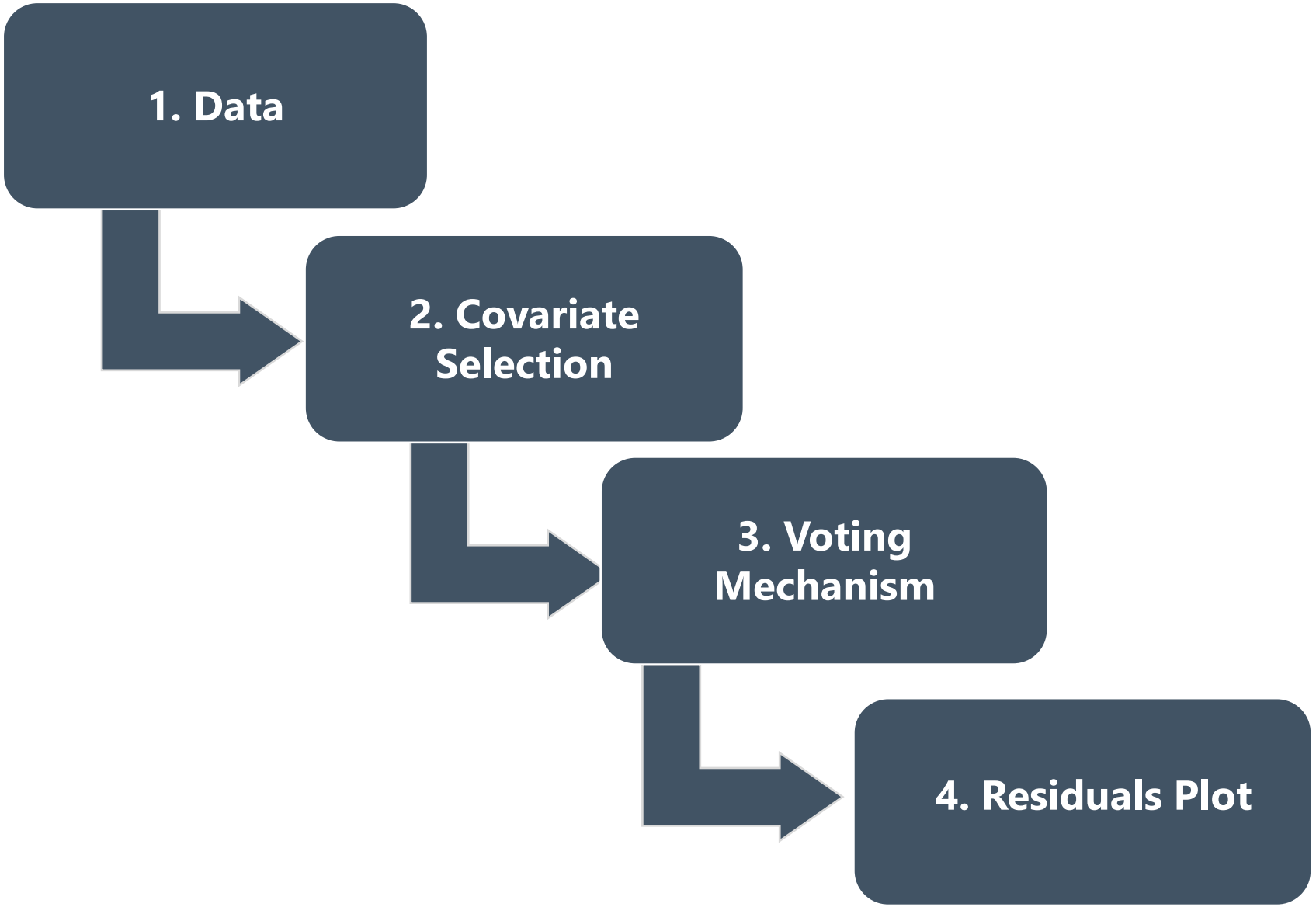
Voting mechanism

ID	CL	V	KA	WGT	ALB	SEX
1	0.41	5.9	0.30	64.2	4.7	0
2	0.52	10.0	0.58	53.8	4.5	0
3	0.39	5.6	2.70	58.1	4.4	0
4	0.75	14.0	0.19	66.7	4.3	0
5	0.34	5.7	1.10	47.7	4.3	0
6	0.59	11.0	1.50	68.1	4.2	1



CERTARA

mlcov R package : Workflow Step 4



- reveals potential overlooked trends with the unselected covariates.

CERTARA

Residuals Plot

ID	CL	V	KA	WGT	ALB	SEX
1	0.41	5.9	0.30	64.2	4.7	0
2	0.52	10.0	0.58	53.8	4.5	0
3	0.39	5.6	2.70	58.1	4.4	0
4	0.75	14.0	0.19	66.7	4.3	1
.	0.34	5.7	1.10	47.7	4.3	0
200	0.59	11.0	1.50	68.1	4.2	1

Train Test Split

Covariates selected after MVE

ID	ALB	SEX
1	4.7	0
4	4.3	1
3	4.4	0
.	4.3	0

X train

CL
0.41
0.75
0.39
0.34

y train

Used training

Covariates selected after MVE

ID	ALB	SEX
2	4.5	0
200	4.2	1

X test

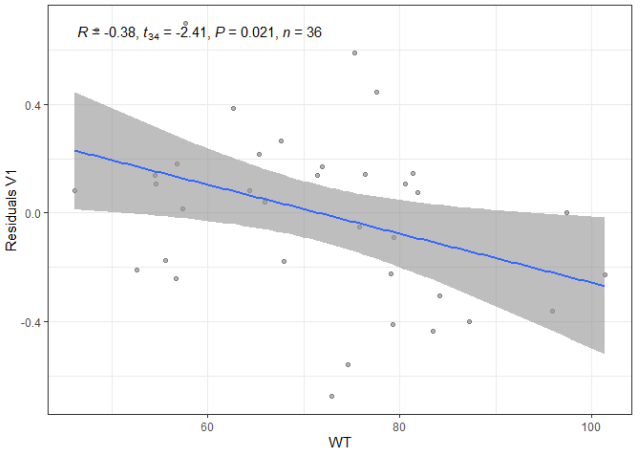
CL
0.52
0.59

y test

Used testing

ML Model

Residuals vs Non-selected covariate

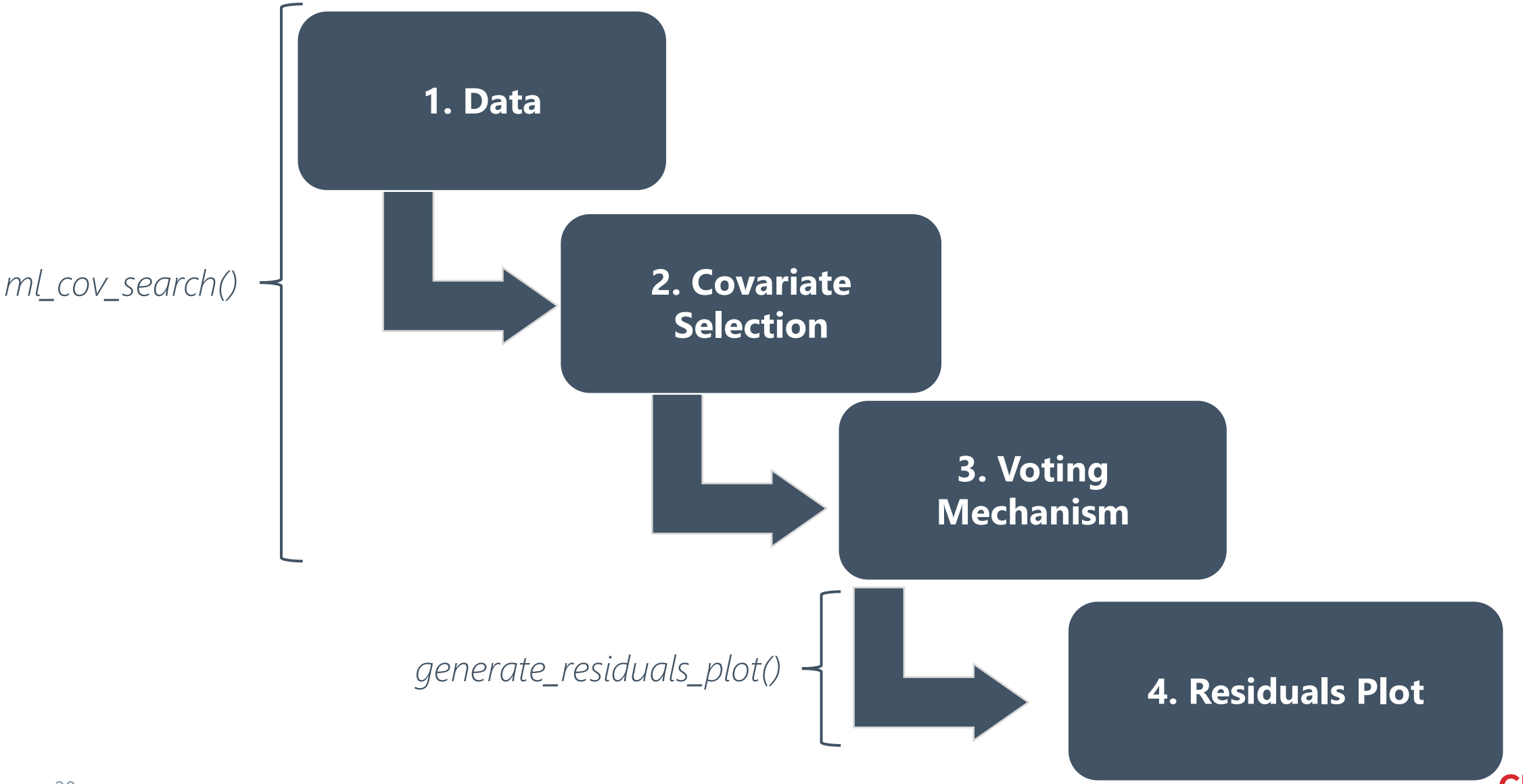


Process repeated 10 times changing train and test data

If absolute majority of p-values are significant : covariate is included in the list of selection

CERTARA

mlcov R package : Functions



CERTARA

Results - Clinical Study Data

Clinical Study Data

Objective : compare the *mlcov* R package and the traditional SCM methodology with clinical study data. Results of both approaches were compared with respect to covariates identified as clinically relevant.

Data :

- one-compartment model developed on Phase 2/ Phase 3 data including N=1957 patients
- Rich PK data
- 14 covariates relationships tested for both SCM and *mlcov*
- Settings of SCM : Power for Continuous, Additive for Categorical

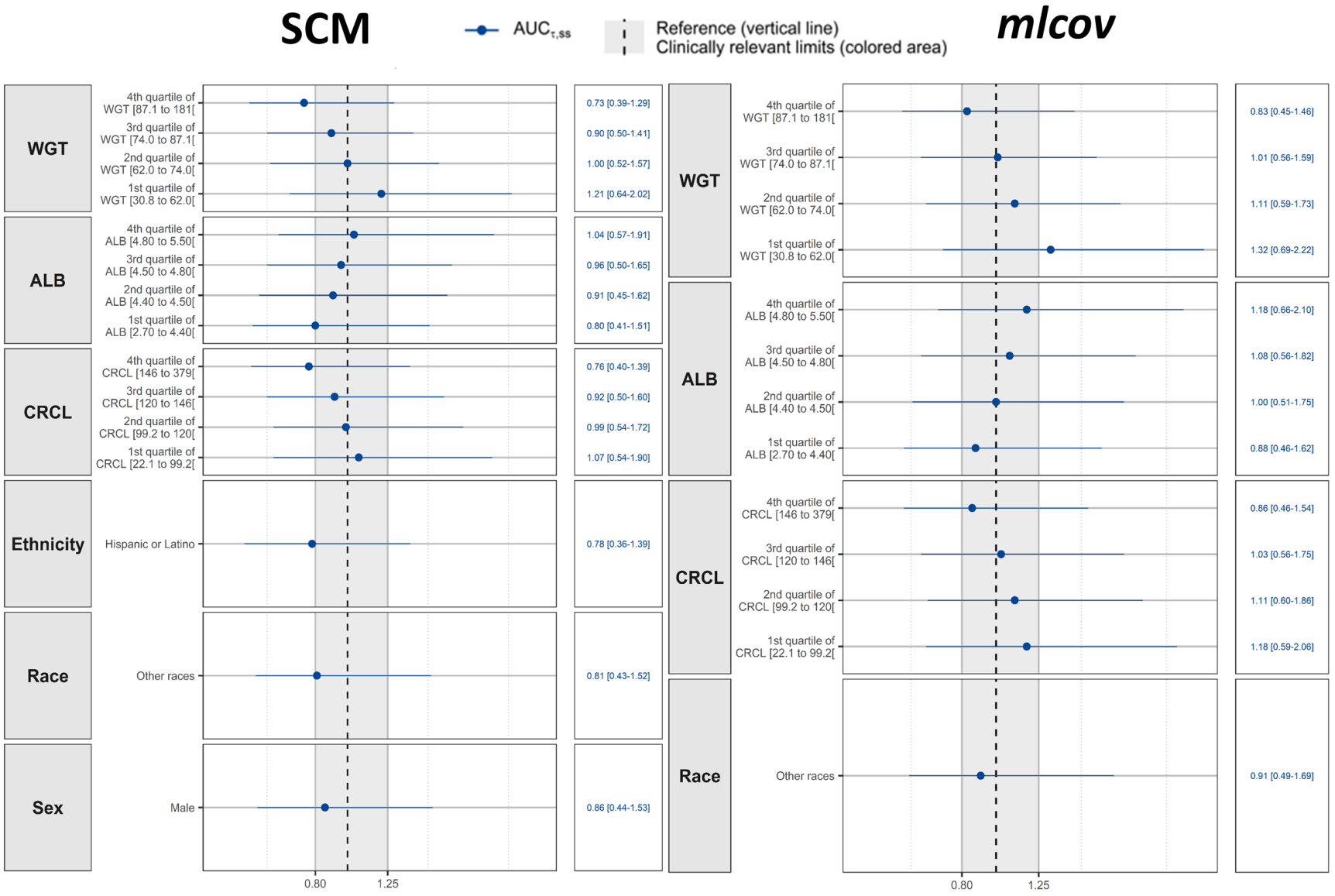
Parameters	Covariates tested
CL/F	WT, albumin, creatinine clearance (CRCL), sex, race, ethnicity (ETHN)
V/F	WT, albumin (ALB), sex, race, ETHN
Ka	age, formulation (FORM), device

Result SCM vs *mlcov*

PROJECT	SCM	*mlcov*	Shrinkage
Data 1	**CL ~ ALB, CRCL, ETHNIC, RACE2, SEXF, WGT** **KA ~ DEVICE** **V ~ ALB, WGT**	**CL ~ ALB, CRCL, RACE2, WGT** **KA ~** **V ~ ALB, WGT**	CL **26%** Ka **69%** V **6%**
Number of covariate selected	9	6	
Covariate rejected by user	1	0	
Execution time	13h	5min	

- Sex and Ethnicity not selected by *mlcov*, likely due to their correlations with bodyweight and race
- Estimation step in software needed after covariate selection

Comparison with Multivariate forest plots



- Covariates unselected by *mlcov* (Sex and Ethnicity) showed no clinical relevance
- Similar trends are observed between both approaches resulting in same conclusions on the clinical relevance of the covariates.

CERTARA

Additional Clinical Study Data

PROJECT		SCM	*mlcov*	Shrinkage
Data 2	Details	**PK, 2 CPT Model, N= 206 participants**		
	Result	CL ~ SEX V1 ~ SEX **30min**	CL ~ **BWT** V1 ~ SEX **1min 30s**	CL **9%** V1 **36%**
Data 3	Details	**PK, Parent Metabolite Model, N= 79 participants**		
	Result	F1 ~ SEX **8h**	F1 ~ SEX **40s**	F1 **2%**
Data 4	Details	**PK, 2 CPT TMDD Model, N= 2796 participants**		
	Result	CL ~ ALBI, COMB, ECOGBIN, LDH, **CRCLI, SEXN, WT** V1 ~ SEXN, WT **9days**	CL ~ ALBI, COMB, ECCOGBIN, LDH, **REGN** V1 ~ SEXN, WT **10min**	CL **15%** V1 **20%**
Data 5	Details	**PK, 2 CPT TMDD Model, N= 1663 participants**		
	Result	CL ~ ALBI, COMB, PIND, WT, **SEXN** V1 ~ SEXN, WT **2days**	CL ~ ALBI, COMB, PIND, WT V1 ~ SEXN, WT **8min**	CL **19%** V1 **25%**

Discussion & Perspectives

- The covariate selection process can become efficient and user friendly by using Machine Learning framework algorithms as implemented in the *mlcov* package

Limitations :

- *mlcov* still in development phase
- Doesn't handle time varying covariate and missing covariates imputation is needed
- Categorial covariates seems to be more challenging to be selected
- LASSO could remove covariate of interest
- Doesn't handle correlation between parameter

Next steps:

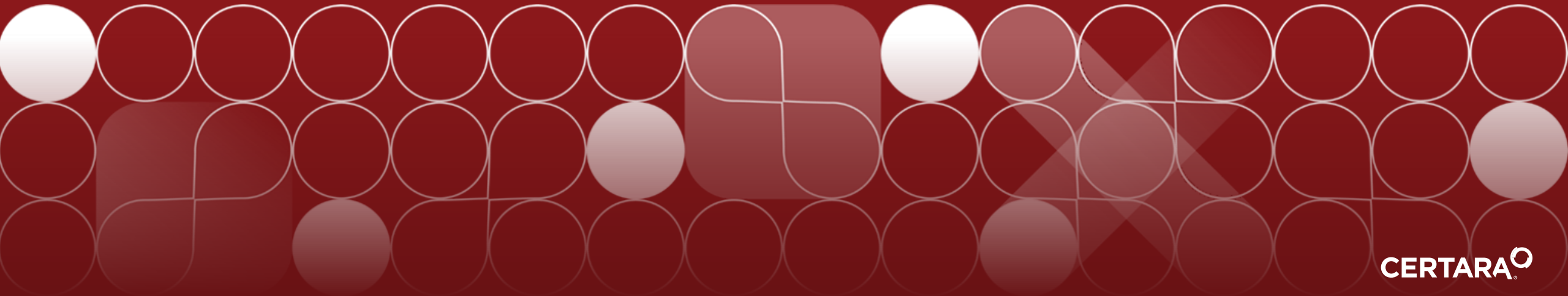
- Continue to test *mlcov* on clinical studies data
- try to get more acceptance by submitting the results to agencies

CERTARA®

Accelerating Medicines, Together

Thank you for your attention.

Contact : ibtissem.rebai@certara.com

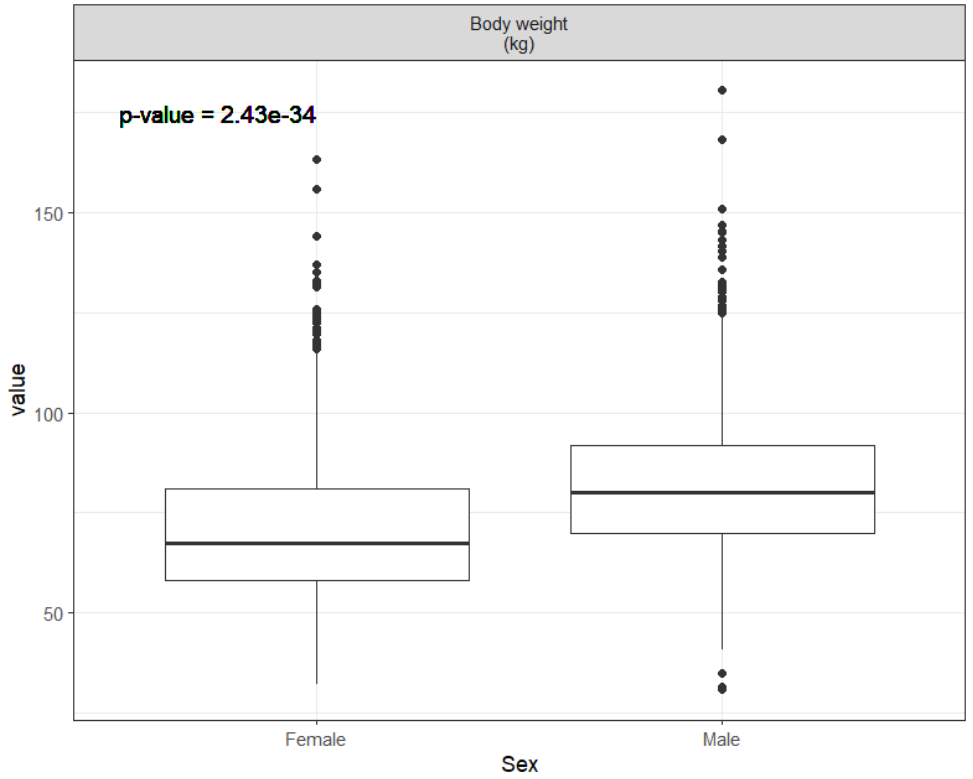


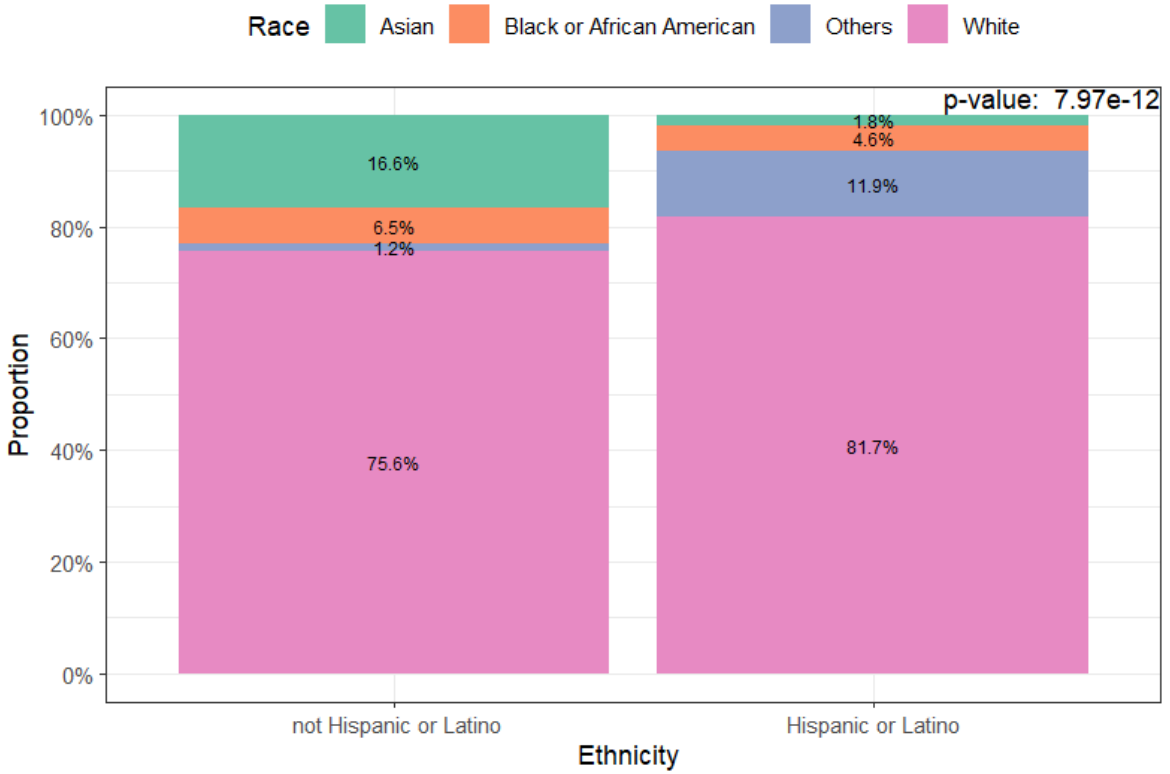
mlcov R package : Implementation

```
devtools::install_github("certara/mlcov")
library(mlcov)

result <- ml_cov_search(data = read.csv('Base_model_Outputs.csv',header = T), #NONMEM output (EBEs+cov)
                        pop_param =  c("CL","V1"),
                        cov_continuous = c("WGT","CRCL","AGE"),
                        cov_factors = c("SEX","ETHNIC","RACE"))


generate_residuals_plot(data = read.csv('Base_model_Outputs.csv',header = T),
                        result = result,
                        pop_param = c('CL'))
```

- Estimation step in software needed after covariate selection

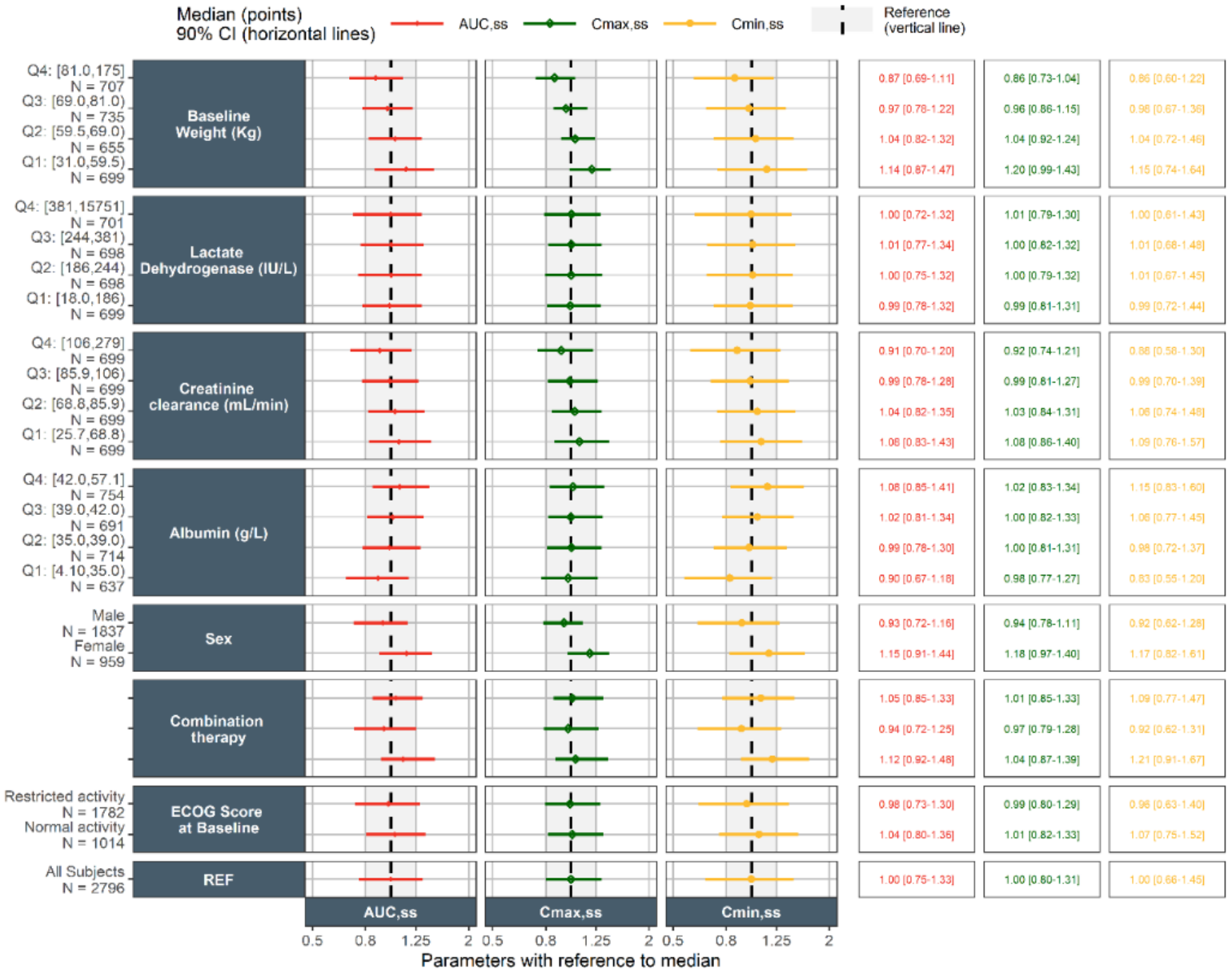
Correlation plot Data 1





CERTARA

Multivariate Forest Plot Data 3

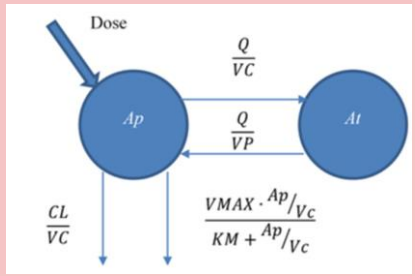


CERTARA

Previous work – simulation study

Simulations

One TMDD model



Scenarios	Covariate relationships added
Scenario 1	No covariate
Scenario 2	CL ~ WT
Scenario 3	V1 ~ WT
Scenario 4	V1 ~ AGE
Scenario 5	V1 ~ SEX
Scenario 6	V1 ~ WT, AGE
Scenario 7	V1 ~ WT, AGE ; CL ~ WT
Scenario 8	V1~ WT, AGE; CL ~ WT, SEX
Scenario 9	V1~ WT, SEX; CL ~ WT, AGE

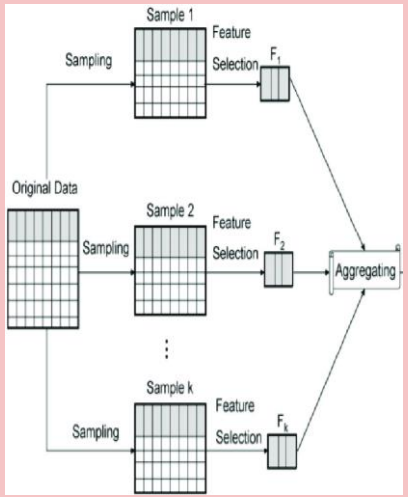
X 100 datasets

Machine Learning Framework

Feature selection algorithm:

1. **Boruta** using **XGboost/Random forest**
2. **Lasso** **+** Boruta using XGboost

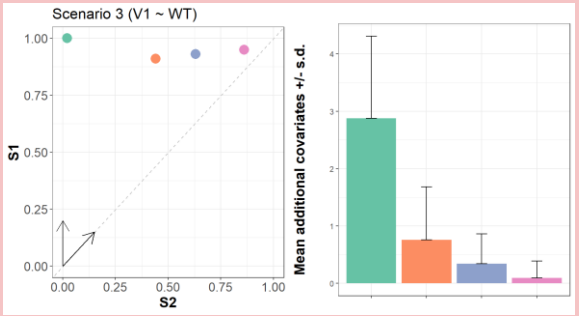
3. **+ Majority Voting Ensemble**



Covariate search assessment:

- **Type 1 error :** false positive rate in scenario 1
- **S1 :** correct covariates with additional ones
- **S2 :** exclusively the correct covariates

Results



CERTARA

Back-up slides

Feature importance in Gradient Boosting

- Mean Decrease Accuracy and Mean Decrease Impurity are two measures used to determine the importance of each feature in a dataset.
  → used to rank the importance of each feature, with higher values indicating more important features.

- Mean Decrease Accuracy (MDA)
  → Measures decrease in model accuracy when a feature is removed.
  → Higher MDA value indicates a more important feature.

- Mean Decrease Impurity (MDI)
  → Measures decrease in node impurity in decision tree when a feature is removed.
  → Higher MDI value indicates a more important feature.

- Impurity in Decision Trees
→ Impurity refers to the degree of "mixing" of different classes within a node.
→ An impure node contains a mixture of different classes.
→ A pure node contains only instances of a single class.

- Goal : Split data to make resulting nodes as pure as possible with respect to the target variable. Ensures accurate predictions on new data.

- Measures of Impurity : Gini Impurity and Entropy
  → used to evaluate the best split in decision trees, aiming to make nodes as pure as possible

Back-up slides

Gradient Boosting

- gradient descent algorithm : iterative optimization algorithm
  → to find the minimum of a cost function by iteratively adjusting the model parameters in the direction of steepest descent of the gradient of the cost function (difference between the predicted values (F(X)) and the actual target values (y))

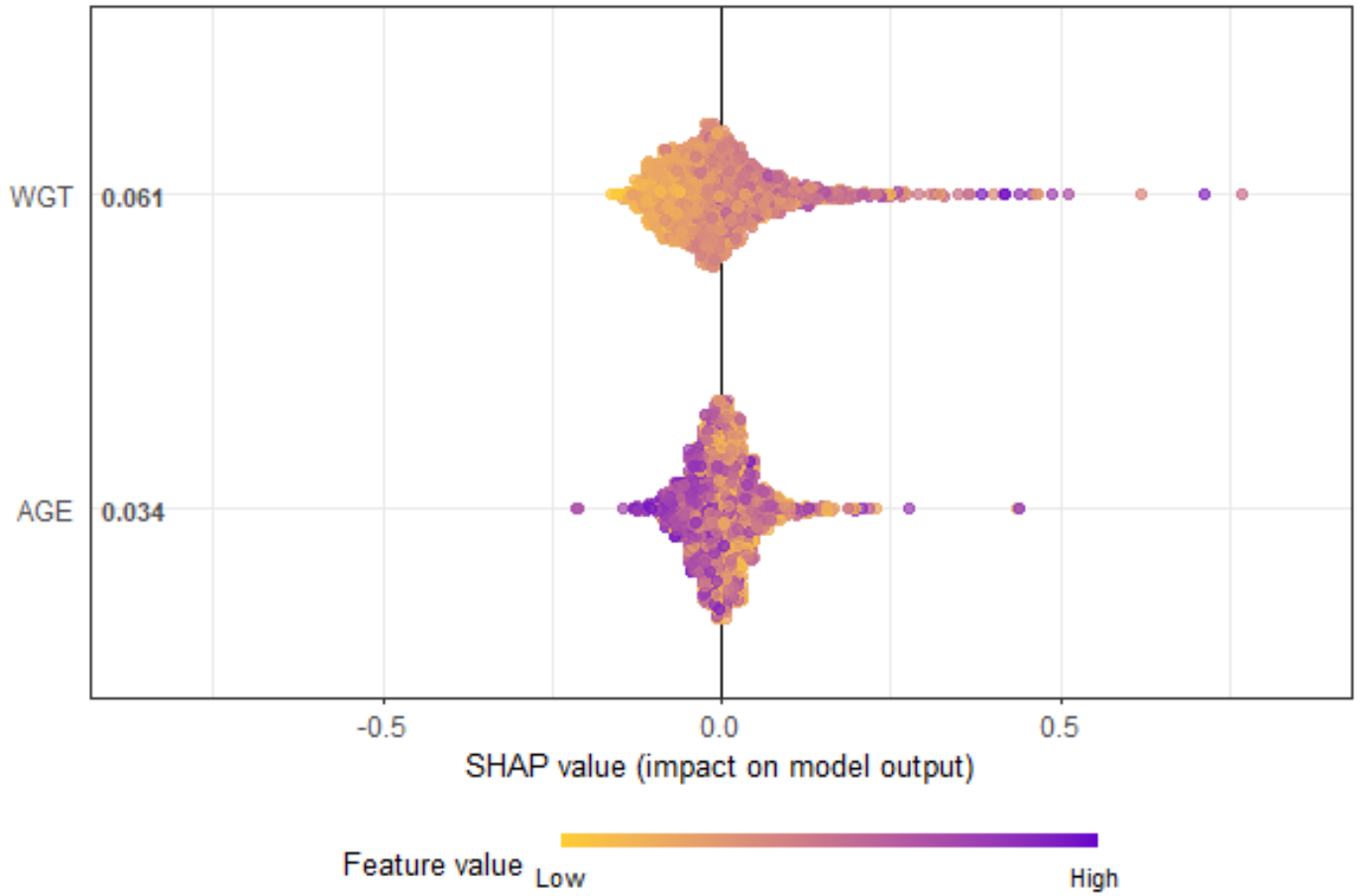
Weighted sum of decision trees in a gradient boosting machine

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) + \frac{\lambda}{2} \sum_{k=1}^{K} \omega_k^2$$

SHAP plot

how each selected covariate affects the model predictions ?



- y-axis indicates the covariate being analyzed
- x-axis shows the SHAP values, representing the impact of WT and AGE on the model's output (CL).
- color distribution suggests that low and high values of the covariate have differing impacts on the predictions

CERTARA