Innovative methods for predicting clinical immunogenicity with high dimensional data

Philippe Broët

Université Paris-Saclay (France)
&
CHU Sainte-Justine & Université de Montréal (Canada, Québec)

Lisboa, February 2019

Outline

- Introduction
- Predict or Explain with high dimensional data
- Methodological challenges of clinical prediction
- Classical and innovatives approaches

  Compound predictor and regularized methods;
  Neural networks; Regression trees & Random Forests
- Prediction of immunogenicity with Random Forests
- Conclusion

What is the main focus of the talk?

- An objective: clinical Prediction
  - To predict clinical immunogenicity of a drug based on patient-related high-dimensional data (genomic/patient's genetic makeup).

  - To build a rule based on what we observed from clinical studies for future risk prediction.

- A Framework for achieving this objective: Statistical predictive models
  - We consider that what is observed Y (outcome) can be modeled as the combination of systematic known factors X (predictors) and random effects ε.

  - To provide prediction rules with good prediction accuracy.

Prediction is not Explanation

- **Prediction study**

  - We want firstly to accurately predict an outcome or that something will (not) occur and secondly explain/interpret why it will (not) happen
  - Main focus: To predict (predictive accuracy)
  - Data >> Model

- **Explanatory study**

  - We want firstly to test some hypotheses about the disease process (mainly an hypothesis regarding a relationship between a phenotype and a bio-clinical factor).
  - Main focus: To explain (understand)
  - Test (Hypothesis Testing) & Estimation (Quantify the relationship)
  - Model >> Data

Prediction is not Explanation (2)

- New shift towards prediction
  - During the 20th century with the tremendous development of modern statistics, the statistical methodology has been oriented towards explanatory goal (testing or estimation approaches).
  - The new century is heading towards predictive inference.
- Prediction vs. Explanation
  - It is often (wrongly) assumed that models with high explanatory power (Cox model) inherently possess predictive power. Explanatory models are not well-designed for interactions, non-linearity.
  - Evaluation criteria are different: Explanatory power $\neq$ Prediction accuracy (p-value $\neq$ predictive values).
  - Prediction and interpretability are usually in conflict.
- Clinical applications require a good balance between prediction accuracy and interpretability

What do we need for prediction studies?

- Main (predictive) objective
- Selected population
- Outcome: continuous (prediction), discrete (discrimination), time-to-event
- Explanatory variables (predictors): Use to construct the rule
- Prediction horizon (e.g. 1-year)
- Predictive model (ML approaches)
- Strategies 2-steps: Developmental step (build and evaluate - accuracy measure) & External validation step (generalizability)

Prediction with High Dimensional data

Data can be summarized as Matrix: n= subjects (row) p: factors (col)

- Matrix of the predictors ($X_{n\times p}$)
- Vector of the outcomes (phenotype) $Y_{n\times 1}$

$$X = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ \dots & x_{ij} & \dots \\ x_{n1} & \dots & x_{np} \end{bmatrix}$$

$$Y = \begin{bmatrix} y_{11} & \dots & y_{1p} \end{bmatrix}$$

High dimensionality: Mainly fat matrices: $p \ggg n$

$$X = \begin{bmatrix} x_{11} & & & \dots & \dots & x_{1p} \\ \dots & \dots & \dots & x_{ij} & \dots & \dots \\ x_{n1} & & & \dots & \dots & x_{np} \end{bmatrix}$$

Prediction with High Dimensional data (2)

- Prediction objective: Build a decision rule $X \curvearrowright_{\Phi} \tilde{Y} = \Phi(X)$

  - Use the predictors (X), Define the rule (Φ), Evaluate the accuracy of the rule (difference between $\tilde{Y}$ and Y)

  - Supervised methods (prediction or classification)

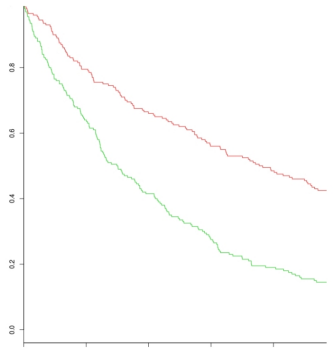
- Challenges: Fat matrices $p \ggg n$
  - Reduction of the dimensionality; Selection of the variables.
  - Various frameworks: Regularization, random forests, neural networks.

What are the specificity of clinical immunogenicity

- Aim & challenges
  - To predict the immunogenicity of a drug (ADA)
  - Take into account that immunogenicity is a dynamic event which is monitored within a window of time.
- Time-to-event data
  - Take into account for censored data (incomplete information)
  If a patient is free of ADA at 6 months but lost to follow-up after $\Longrightarrow$ Use the information of being free of ADA up to 6 months
  - The information is $X = min(T, C)$ and $\delta = 1$ if $X = T$ and $\delta = 0$ otherwise (censored information)

Time-to-event analysis

- Survival modeling
  - Survival function:



$$S(t) = Pr(T > t)$$

  - **Hazard risk**:

$$\lambda(t) = \lim_{\delta t \to 0} \frac{Pr(t<T<t+\delta t | T>t)}{\delta t}$$

  - **Cox model** (multiplicative):

$$\lambda(t; X) = \lambda_0(t) \times e^{\Phi(X)}$$

  - Other models: Accelerated failure times,...

- Predictive tools adapted to time-to-event data
  - Compound predictor and regularized methods
  - Neural networks
  - Regression trees & Random Forests

Compound predictor and regularized methods

- **Compound predictor**
  Mostly relying on Cox model.
  Linear combination of weighted genomic measurements (risk score)

$$\Phi(X) = \sum_{i=1}^{p} w_i X_i$$
$$\text{with } \lambda(t; X) = \lambda_0(t) \times e^{\Phi(X)}$$

  - If $\Phi(X)$ increases then the probability of the event increases
  - Ad-hoc separation bad/good prognostic groups
  - Two-steps: Feature selection (most significant) and prediction

- **Regularized predictor**
  - Selection and prediction in one-step (Lasso and Elastic-net)

Compound predictor and regularized methods (2)

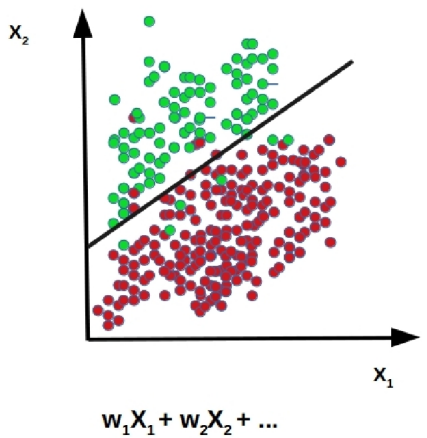
- Advantages

  Straightforward to understand, to communicate (weights; parameters); to explain. $S = (0.251) * gene_1 + (-0.232) * gene_2 + ... + (0.08) * gene_k$

  Regularization avoid overfitting

- Drawbacks

  Simplified model: Linear model

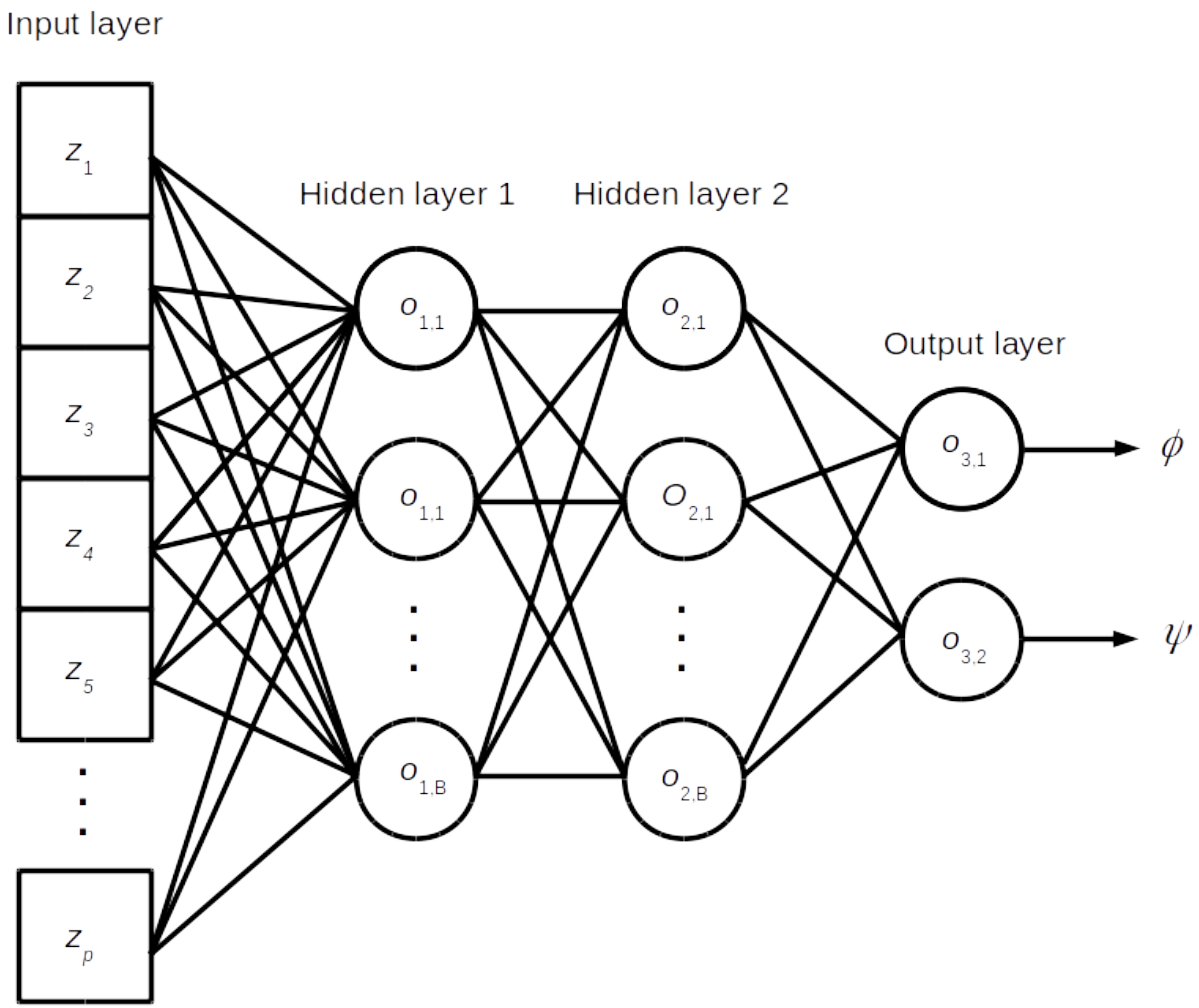


Performs poorly when there are complex non-linear relationships (e.g. gene*gene interaction)

Machine learning (ML) approaches

- Compound (regularized) predictors have been widely used in the last decade for time-to-event prediction.
- Well-known for not being suited for coping with complex higher-order interactions with high-dimensional data.
- ML approaches
  - Artificial Neural Networks
  - Random forests

Artificial neural networks



Artificial neural networks (2)

- Advantages
  - Powerful nonlinear modelization
  - High predictive accuracy

Artificial neural networks (3)

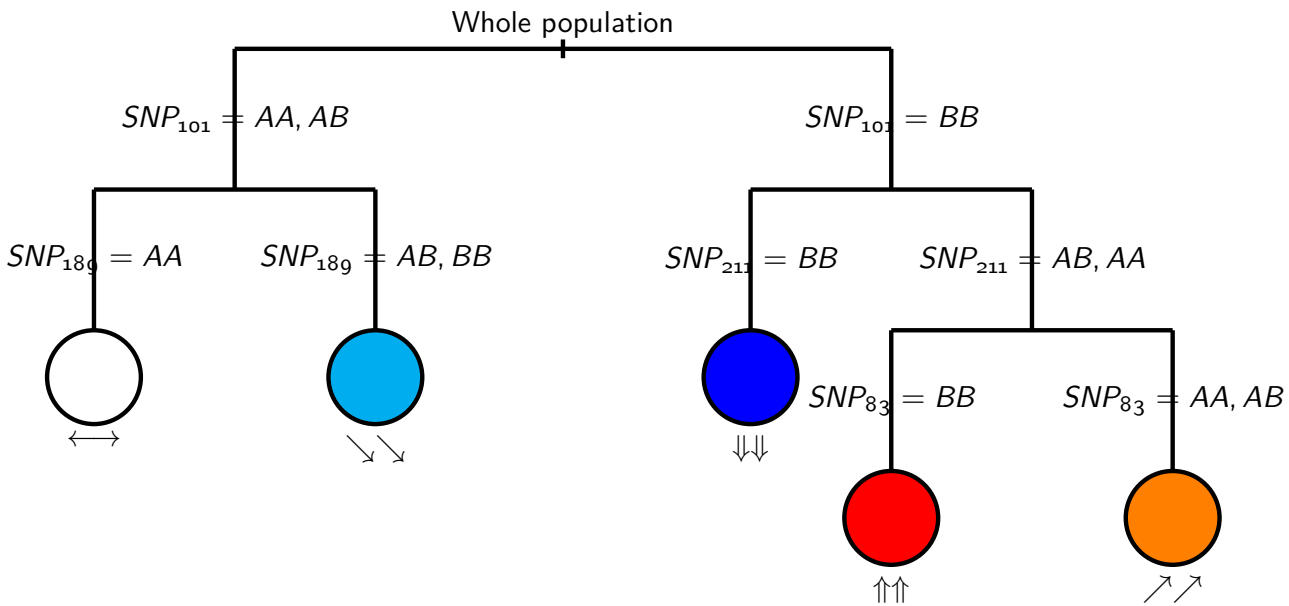
- Drawbacks

  - Low level of interpretability (black-box).
  - Require a very large amount of data (training samples).
  - Use mainly for $n >> p$.
  - Computationally intensive to train. Require expertise to tune the architecture and hyperparameters.
  - Prone to overfitting.
  - Need more research works for being use in clinical medicine with time-to-event outcomes.

Tree-based models & Random Forests

- Tree based model **(recursive partitioning methodology)** for taking into account gene $\times$ gene interaction

$\Longrightarrow$ Main principle: To decompose the data space (explanatory variables) recursively into more homogeneous areas (with respect to the main outcome) in a tree-structured fashion.



Tree-based models (2)

- Advantages

  - Powerful nonlinear modelization
  - High level of interpretability (set of rules)
  - Can be extended for time-to-event data: Survival trees (create homogeneous groups with respect to the probability of the occurence of ADA).

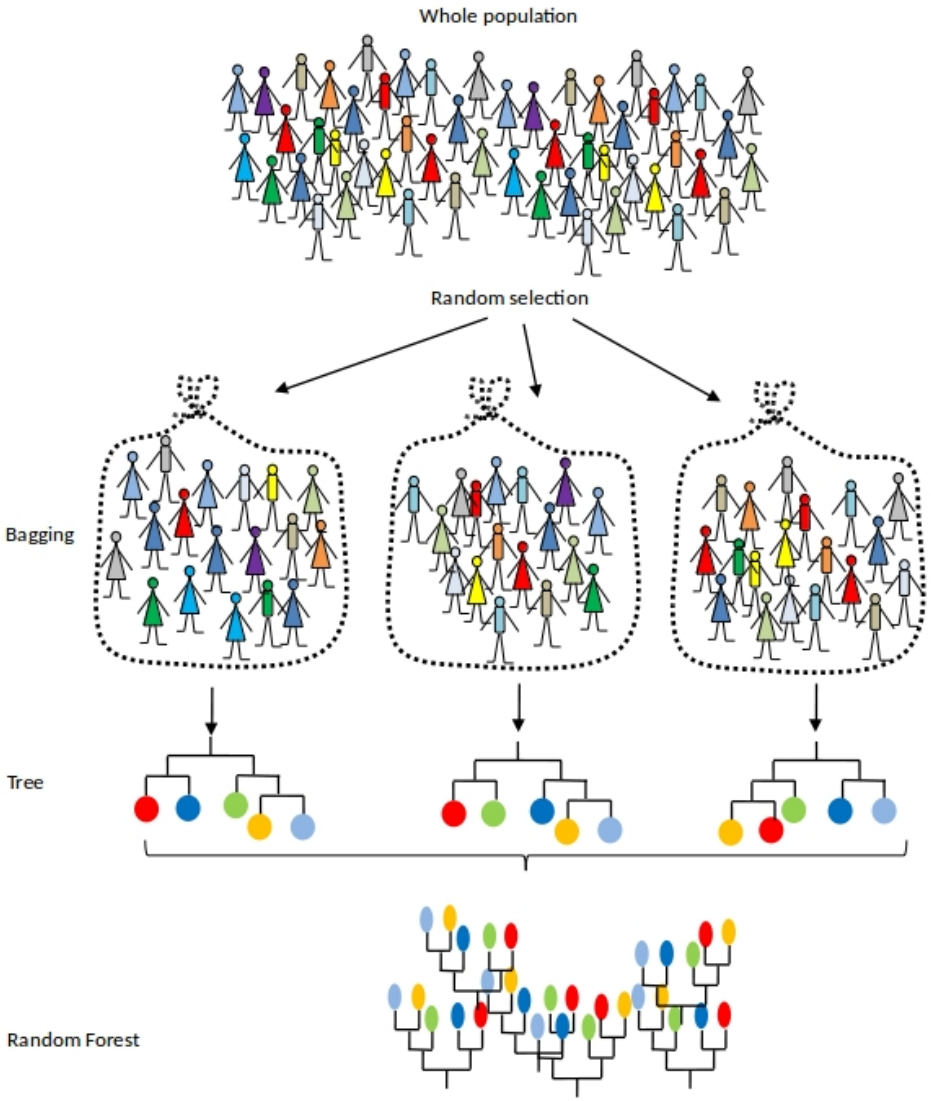
Tree-based models (3)

- Drawbacks
  - Prone to instability.
  - A small change in the data set can result in a very different series of splits
  - Variable selection and prediction somewhat precarious !

- Solution: $\Rightarrow$ Create multiple predictors (with bootstrapped samples) and Aggregate the predictions (**Random Forets**)

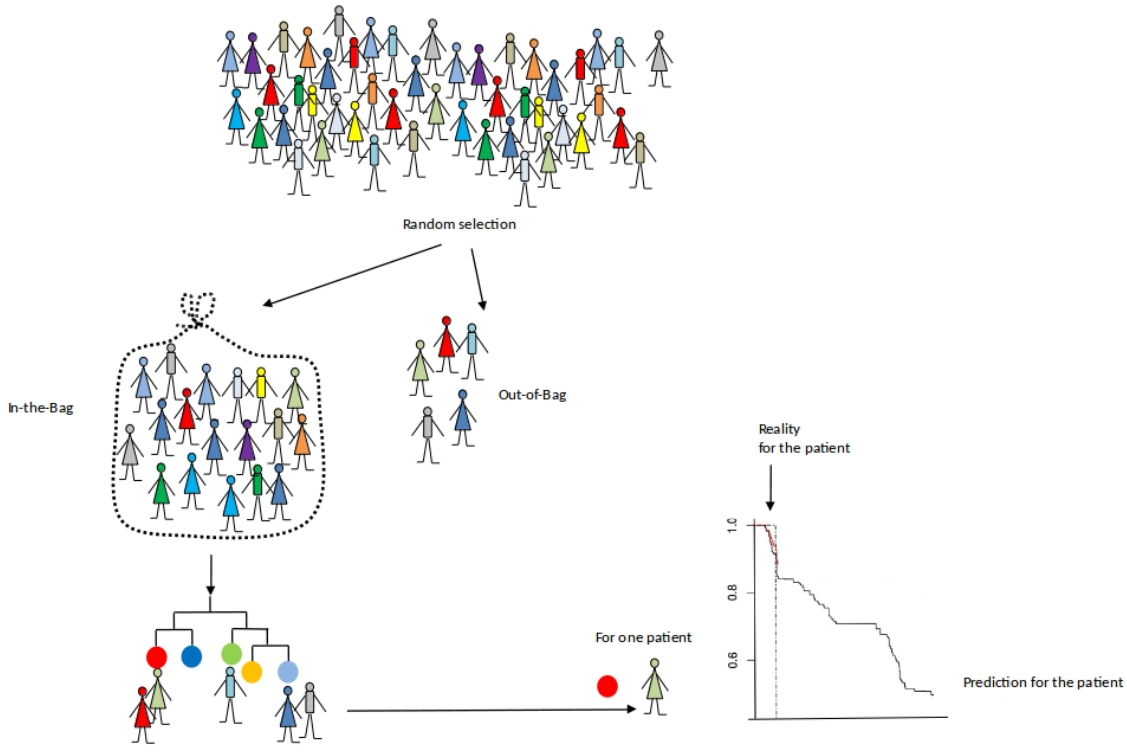
Random Forests

- Main principle: Build many trees (with bootstrapped samples)



Random Forests (2)

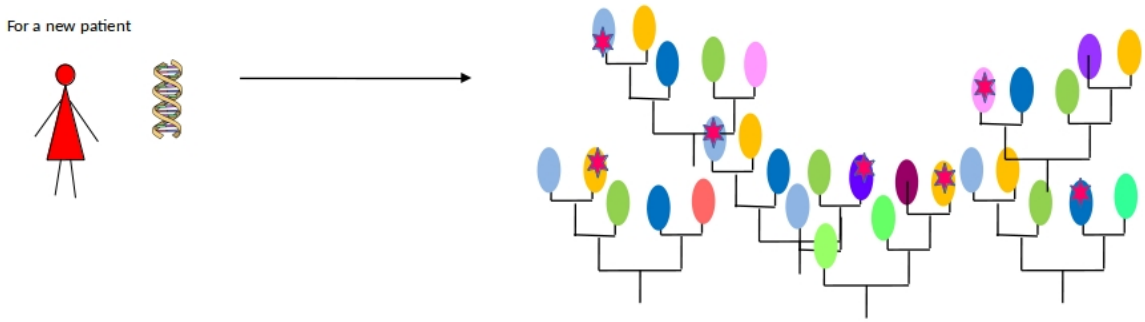
- **Provide accuracy estimation:** The **Brier Score** is a summary of the prediction error by integrating over time

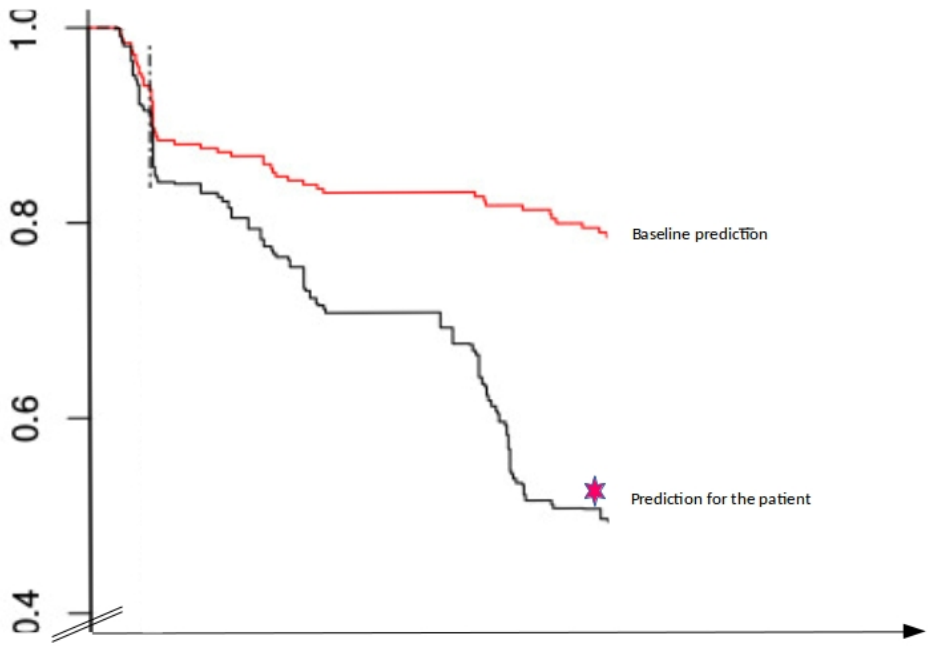


$\Longrightarrow$ The final prediction of a forest is the average of the predictions of the trees

Random Forests (3)

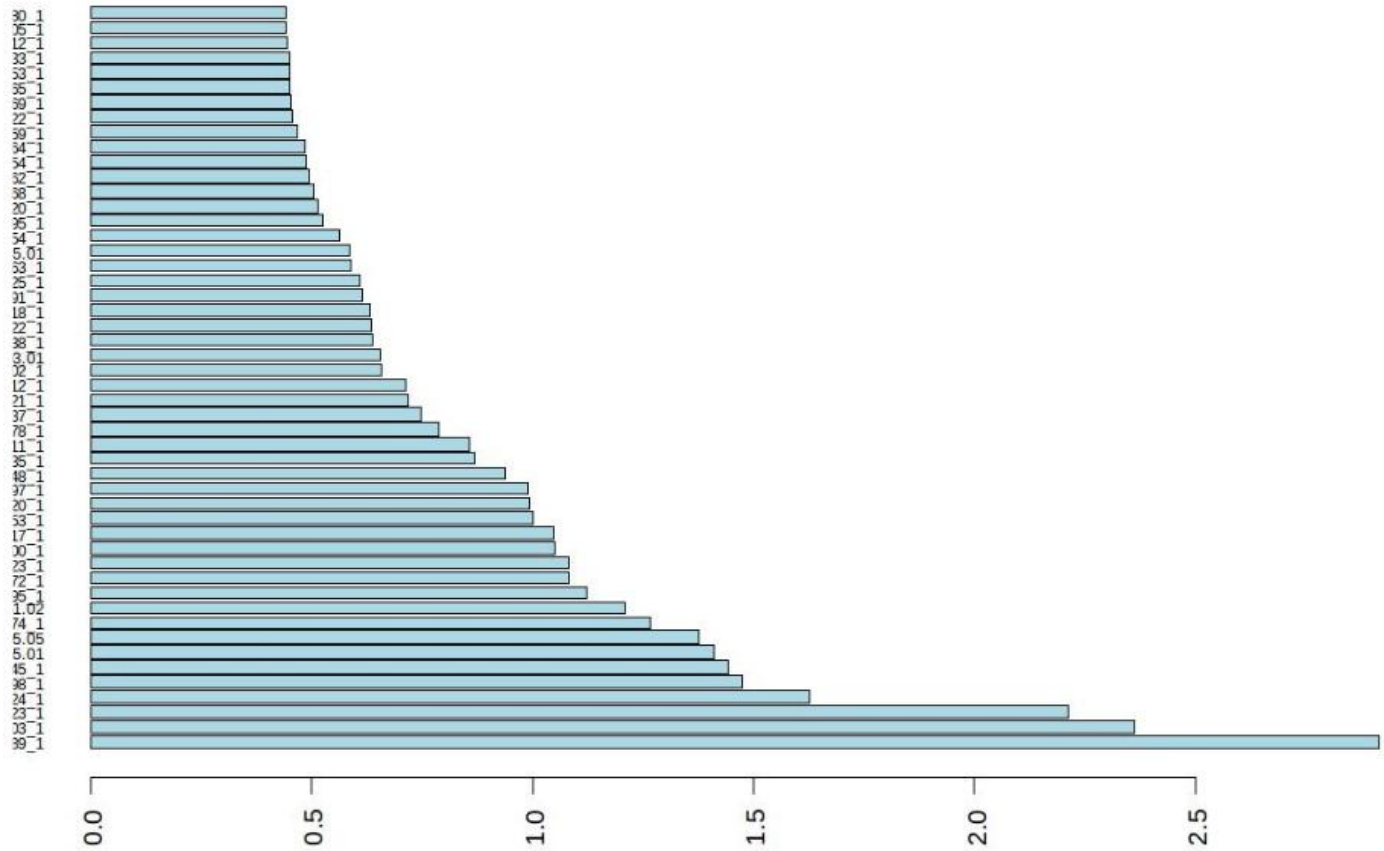
- **Prediction use:** Bagged survival trees can be used for immunogenicity prediction of a new patient





Selection

- **Provide interpretative tools:** The loss of interpretability associated with the forest is compensated by a ranking of variable importance (selection).



Conclusion

- Clinical prediction for imunogenicity

- Prediction is a hot topic
- When high dimensional data with complex interactions are expected ML approches can be used
- Prediction for clinical use should strike a good balance between accuracy and interpretability.
- RF offers a good trade-off.
- NN are newcomers and still black-box.

- Limitations and Extensions

-There is no one-size-fits-all predictive tool.
- Methods should be tailored to specific problems.

Thank you for your attention