Comparing Modern Machine Learning Methods for Predicting Individual Treatment Effect

Pamela Solano[1] and Thomas Jaki[1,2]

[1]Regensburg University, [2]Cambridge University

July 24, 2024

Motivation

- **not all participants will react equally to an intervention**

Motivation

- **not all participants will react equally to an intervention**
- Characterizing this **heterogeneity in intervention effects** is key to improving patients outcomes.

Objective

to estimate Individual Treatment Effects.

To compare state-of-the-art **Statistical models and Machine Learning algorithms**.

The predicted individual treatment effects **(PITE)** framework

- **PITE is a method to estimate individual treatment effects**.
- PITE consists of the difference between **experimental(E) and control (C) prediction** for each individual [Jaki et al., 2024].

$$\text{PITE}_i = f_E(\boldsymbol{X}_i) - f_C(\boldsymbol{X}_i), \qquad f(\cdot) \text{ is a predictor}$$

The predicted individual treatment effects **(PITE)** framework

challenge

PITE is unobserved.

The outcome is observed for a given patient only under either the **experimental(E)** $f_E(\boldsymbol{x}_i)$ or **control (C) condition** $f_C(\boldsymbol{x}_i)$

a natural question

Which method is the best to estimate PITE?

Can any approach for prediction be applied?

Predictive models should address

- population heterogeneity,

Can any approach for prediction be applied?

Predictive models should address

- population heterogeneity,
- complex structures (linear/non-linear)

Can any approach for prediction be applied?

Predictive models should address

- population heterogeneity,
- complex structures (linear/non-linear)
- and high-dimensionality [Lamont et al., 2016].

Can any approach for prediction be applied?

important!

- The best method to estimate the PITE is not **necessarily** the best fitted model.
- An essential part is **to delineate this heterogeneity** based on the baseline covariates (features).

Regressions approaches applied

Metrics

- Sensitivity (Detect PITE direction)

$$\text{Same Direction}_i = \begin{cases} 1 & \text{if } \text{tPITE}_i \times \text{PITE}_i > 0 \\ -1 & \text{otherwise} \end{cases}$$

true tPITE_i and estimated PITE_i.

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\text{tPITE}_i \times \text{PITE}_i > 0}$$

For each method, calculate the proportion of individuals whose estimated PITE have the same direction as the true PITE.

Metrics

- Risk (Expected Squared Error)

$$\frac{1}{n}\sum_{i=1}^{n}(\text{tPITE}_i - \text{PITE}_i)^2 \times 100$$

Data generation mechanism

Sample size $n = 40, 70, 100, 300, 400, \mathbf{500}, \mathbf{1000}, 1200, 1500$ with allocation ratio 1:1 ($n_C = n_T = n/2$).

$$y = \boldsymbol{X}\beta + t\boldsymbol{Z}\gamma + \epsilon, \quad \epsilon \sim N(0, 1)$$

$t \in (0, 1) \Rightarrow$ benefit $\boldsymbol{Z}\gamma$

Leave-one-out cross-validation for validation

- $\boldsymbol{Z}$ Normal, Linear, Independent

Data generation mechanism

Sample size $n = 40, 70, 100, 300, 400, \mathbf{500}, \mathbf{1000}, 1200, 1500$ with allocation ratio 1:1 ($n_C = n_T = n/2$).

$$y = \boldsymbol{X}\beta + t\boldsymbol{Z}\gamma + \epsilon, \quad \epsilon \sim N(0, 1)$$

$t \in (0, 1) \Rightarrow$ benefit $\boldsymbol{Z}\gamma$

Leave-one-out cross-validation for validation

- $\boldsymbol{Z}$ Normal, Linear, Independent
- $\boldsymbol{Z}$ Normal, Linear, Interactions

Data generation mechanism

Sample size $n = 40, 70, 100, 300, 400, \mathbf{500}, \mathbf{1000}, 1200, 1500$ with allocation ratio 1:1 ($n_C = n_T = n/2$).

$$y = \boldsymbol{X}\beta + t\boldsymbol{Z}\gamma + \epsilon, \quad \epsilon \sim N(0, 1)$$

$t \in (0, 1) \Rightarrow$ benefit $\boldsymbol{Z}\gamma$

Leave-one-out cross-validation for validation

- $\boldsymbol{Z}$ Normal, Linear, Independent
- $\boldsymbol{Z}$ Normal, Linear, Interactions
- $\boldsymbol{Z}$ Normal, Linear, Low correlations **(up to 0.3)**

Data generation mechanism

Sample size $n = 40, 70, 100, 300, 400, \mathbf{500}, \mathbf{1000}, 1200, 1500$ with allocation ratio 1:1 ($n_C = n_T = n/2$).

$$y = \boldsymbol{X}\beta + t\boldsymbol{Z}\gamma + \epsilon, \quad \epsilon \sim N(0, 1)$$

$t \in (0, 1) \Rightarrow$ benefit $\boldsymbol{Z}\gamma$

Leave-one-out cross-validation for validation

- $\boldsymbol{Z}$ Normal, Linear, Independent
- $\boldsymbol{Z}$ Normal, Linear, Interactions
- $\boldsymbol{Z}$ Normal, Linear, Low correlations **(up to 0.3)**
- $\boldsymbol{Z}$ Normal, Linear, High correlations **(up to 0.5)**

Data generation mechanism

Sample size $n = 40, 70, 100, 300, 400, \mathbf{500}, \mathbf{1000}, 1200, 1500$ with allocation ratio 1:1 ($n_C = n_T = n/2$).

$$y = \boldsymbol{X}\beta + t\boldsymbol{Z}\gamma + \epsilon, \quad \epsilon \sim N(0, 1)$$

$t \in (0, 1) \Rightarrow$ benefit $\boldsymbol{Z}\gamma$

Leave-one-out cross-validation for validation

- $\boldsymbol{Z}$ Normal, Linear, Independent
- $\boldsymbol{Z}$ Normal, Linear, Interactions
- $\boldsymbol{Z}$ Normal, Linear, Low correlations **(up to 0.3)**
- $\boldsymbol{Z}$ Normal, Linear, High correlations **(up to 0.5)**
- $\boldsymbol{Z} \sim U[0.1, 0.5]$, Non-Linear. Benefit: $Z_1\gamma_1/(Z_2 + \gamma_2)$

Data generation mechanism

Sample size $n = 40, 70, 100, 300, 400, \mathbf{500}, \mathbf{1000}, 1200, 1500$ with allocation ratio 1:1 ($n_C = n_T = n/2$).

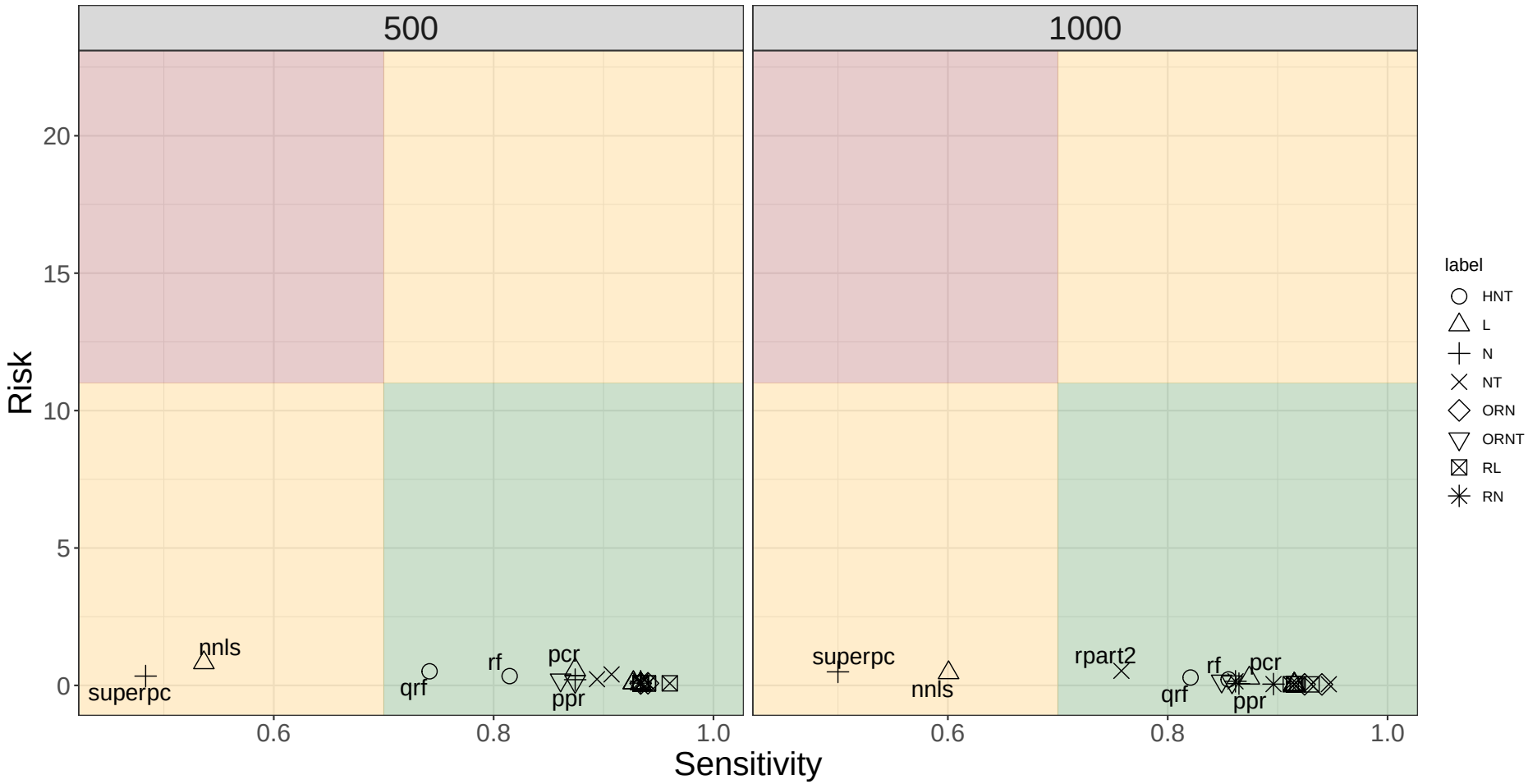
$$y = \boldsymbol{X}\beta + t\boldsymbol{Z}\gamma + \epsilon, \quad \epsilon \sim N(0, 1)$$

$t \in (0, 1) \Rightarrow$ benefit $\boldsymbol{Z}\gamma$

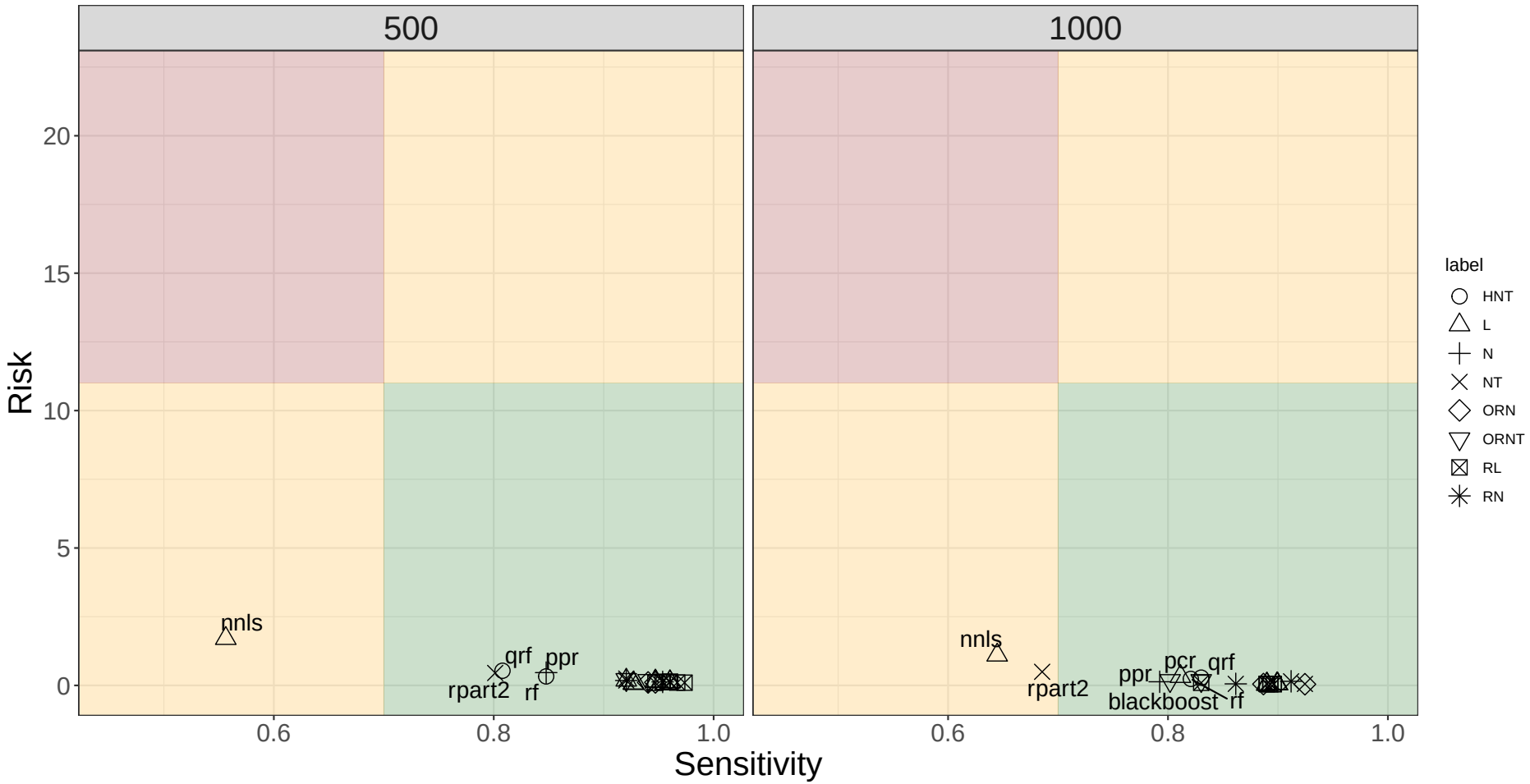
Leave-one-out cross-validation for validation

- $\boldsymbol{Z}$ Normal, Linear, Independent
- $\boldsymbol{Z}$ Normal, Linear, Interactions
- $\boldsymbol{Z}$ Normal, Linear, Low correlations **(up to 0.3)**
- $\boldsymbol{Z}$ Normal, Linear, High correlations **(up to 0.5)**
- $\boldsymbol{Z} \sim U[0.1, 0.5]$, Non-Linear. Benefit: $Z_1\gamma_1/(Z_2 + \gamma_2)$
- $\boldsymbol{Z} \sim U[0.1, 0.5]$, Non-Linear. Benefit: $\frac{\log(Z_1)\gamma_0}{{Z_1}^{\gamma_2} - \gamma_3\sqrt{Z_2 + 2}}$

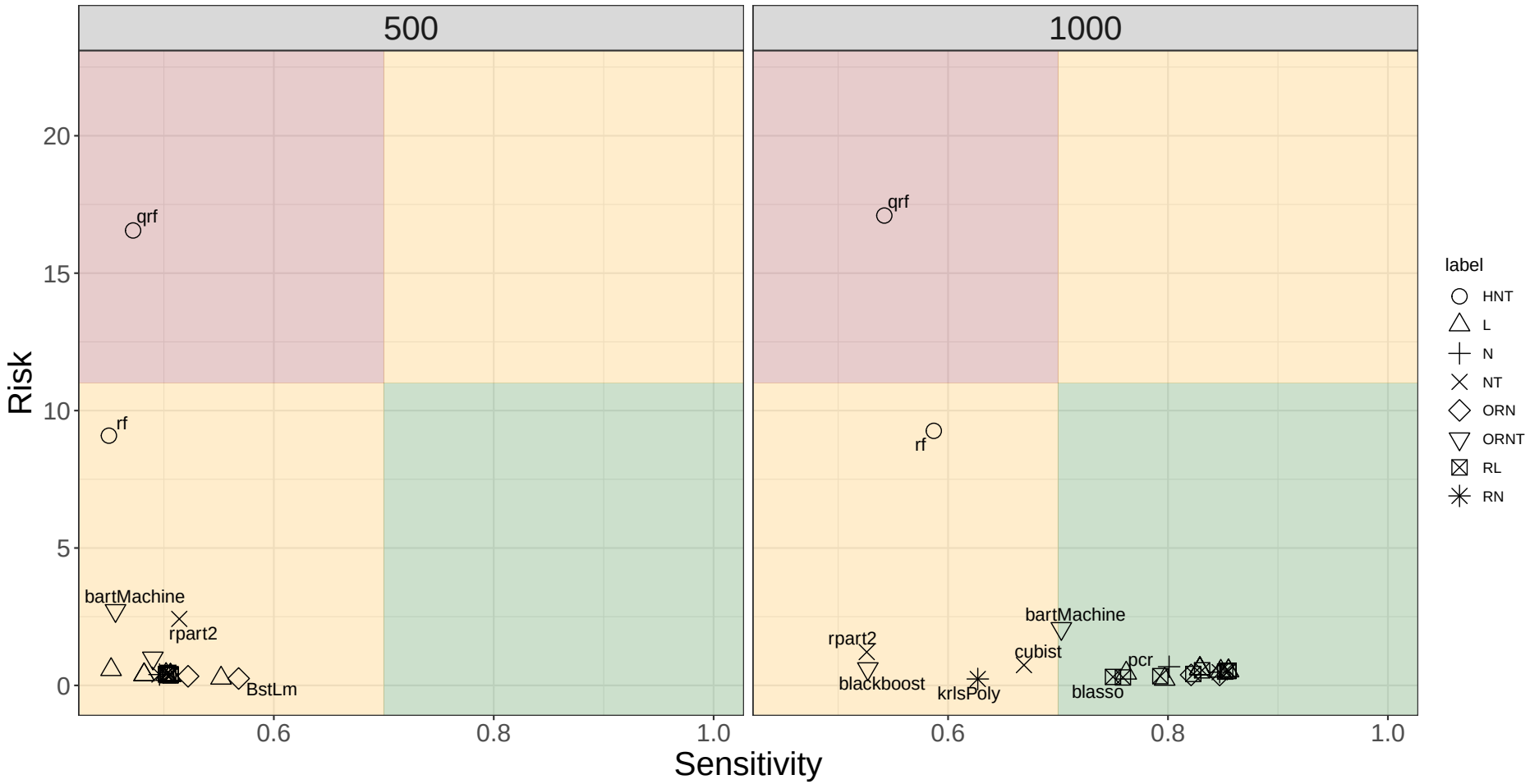
Linear, no Interaction



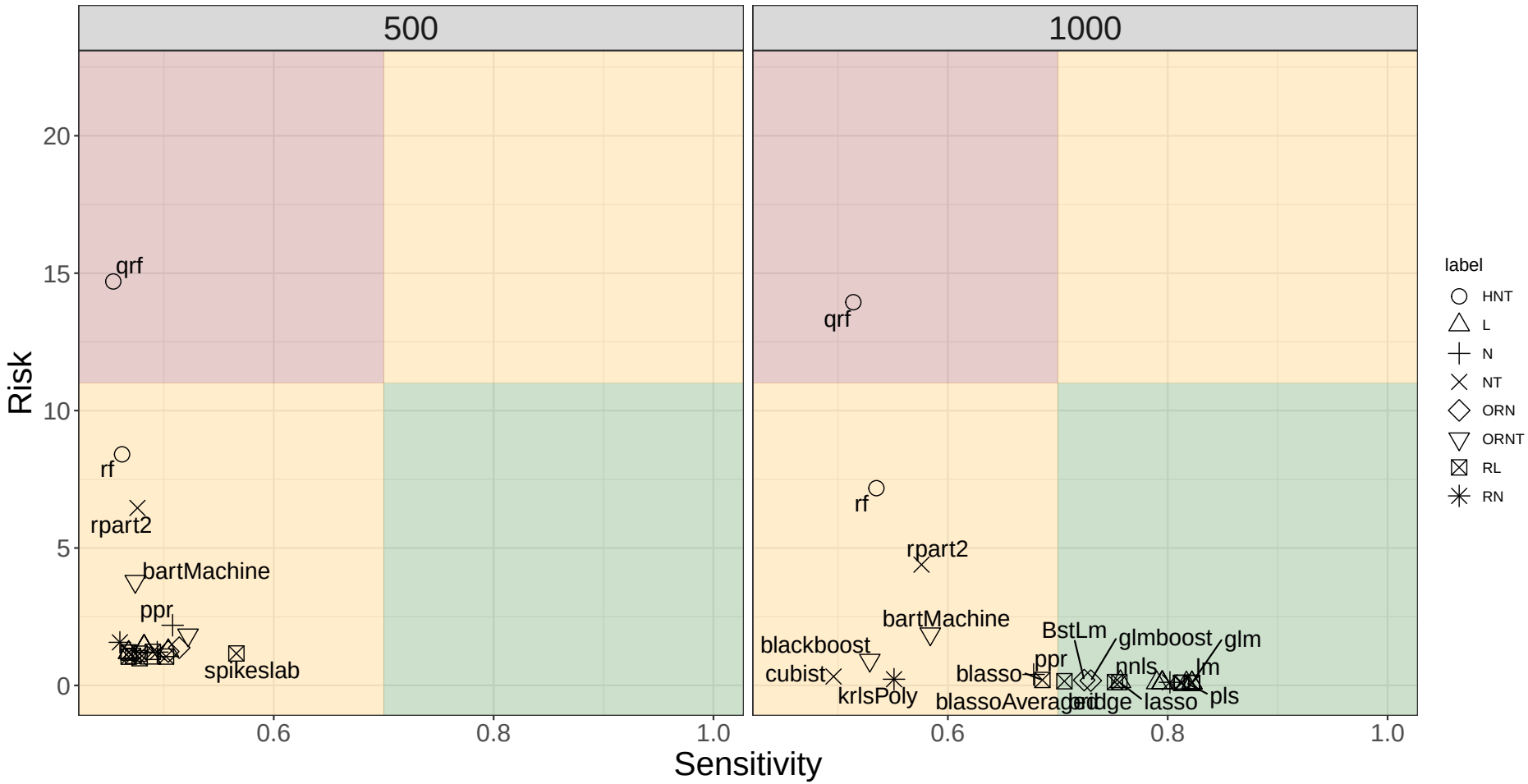
Linear, with Interaction



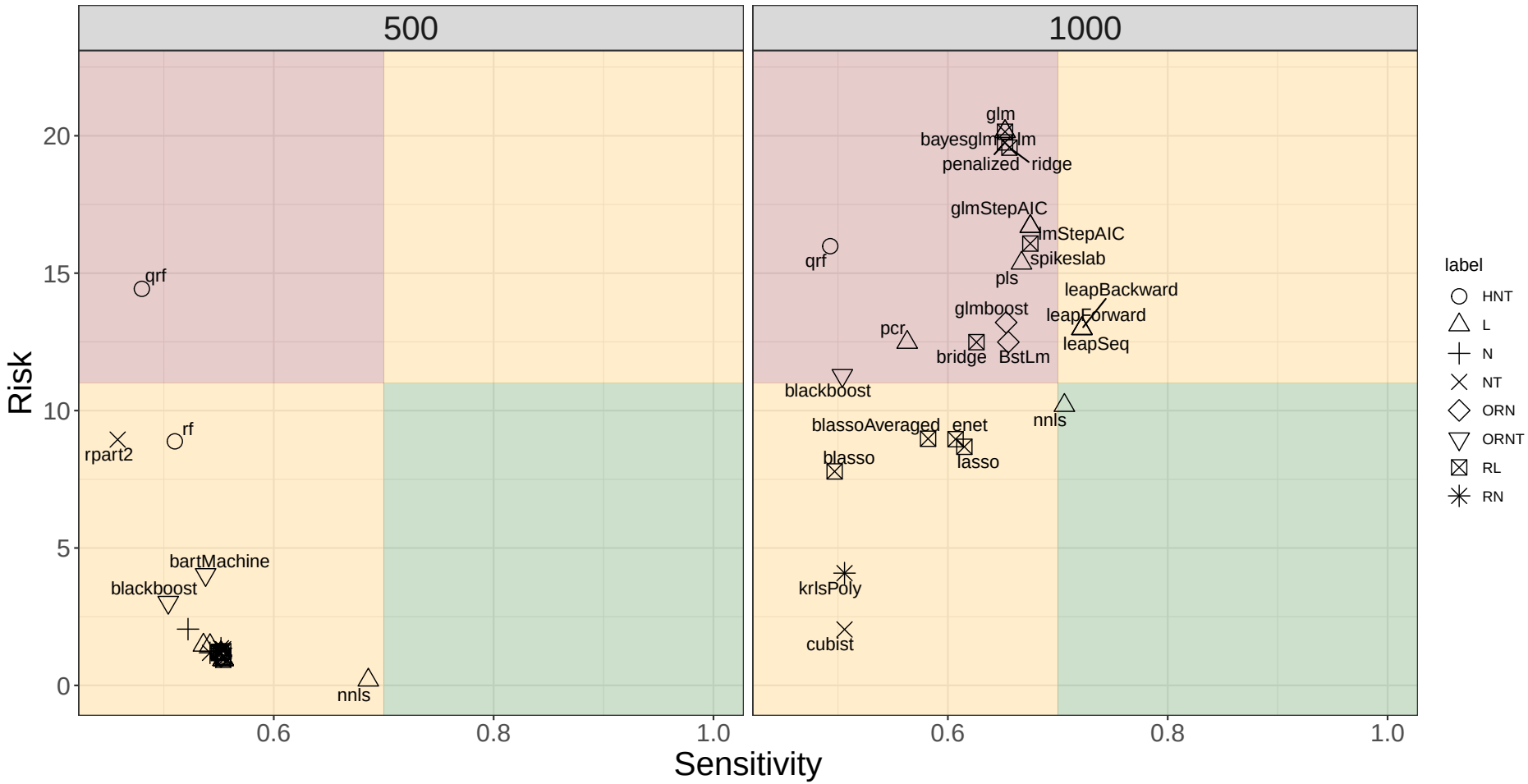
Linear, low Correlation



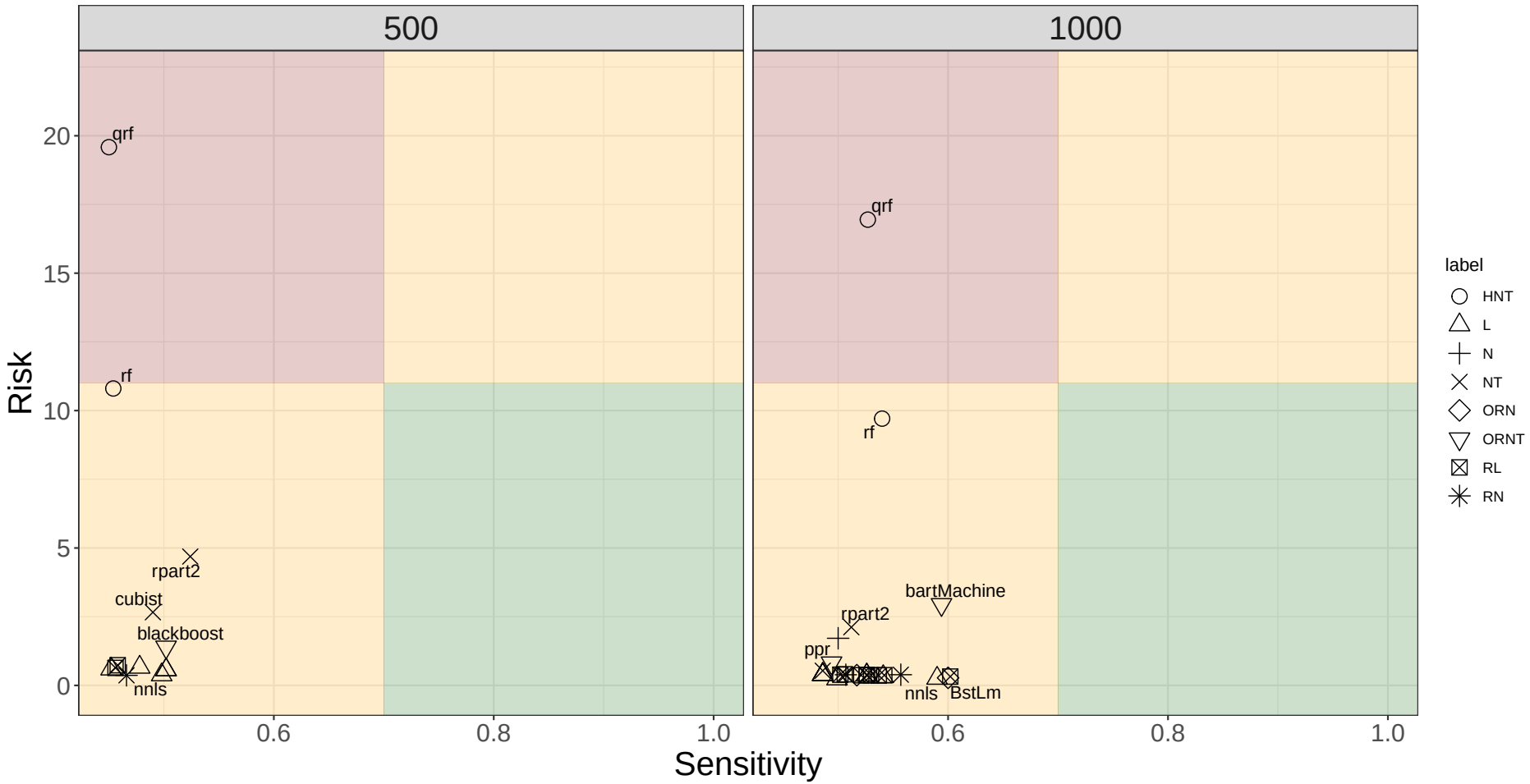
Linear, high Correlation



No-Linear, Simple

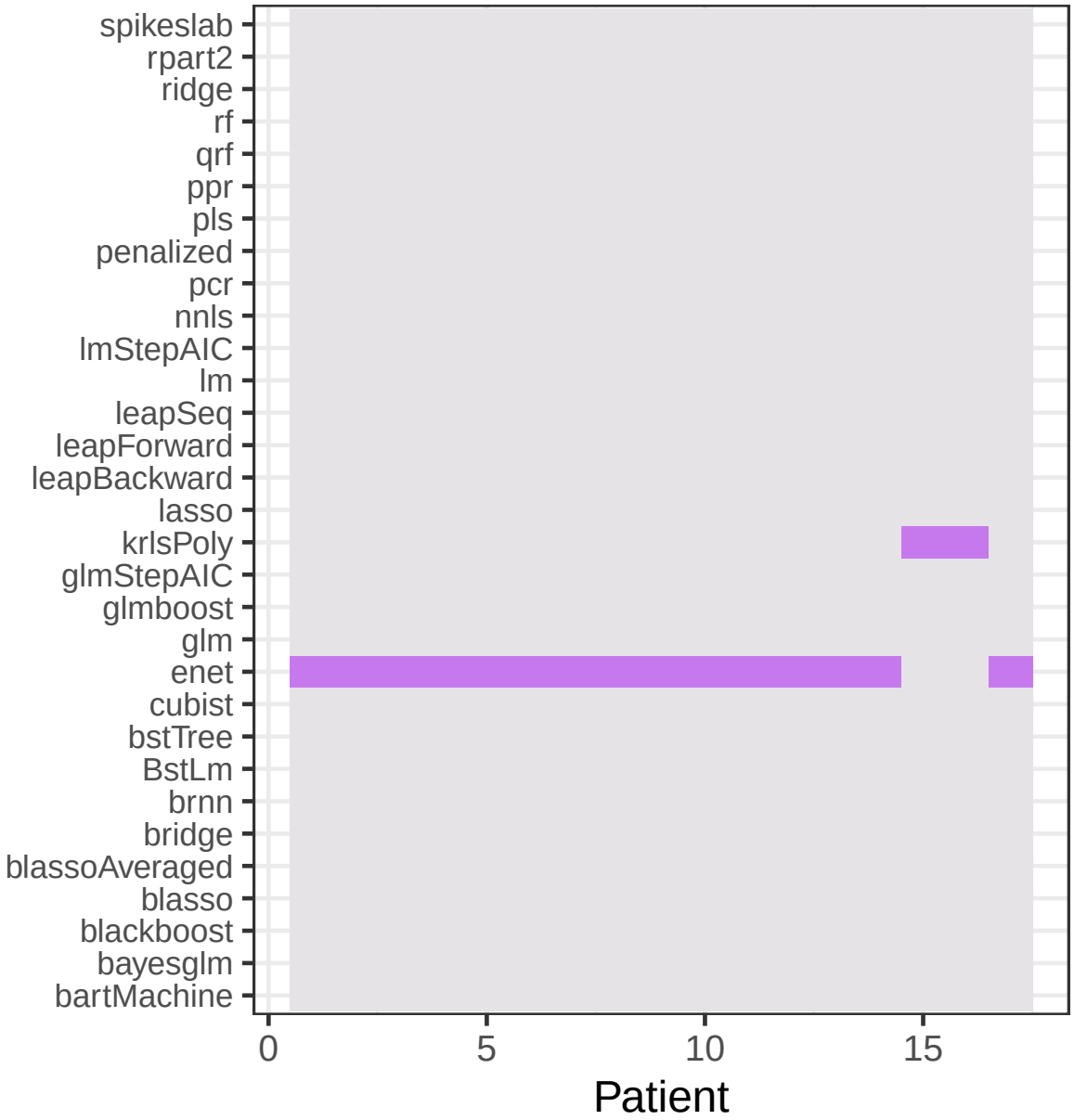


No-Linear, Complex



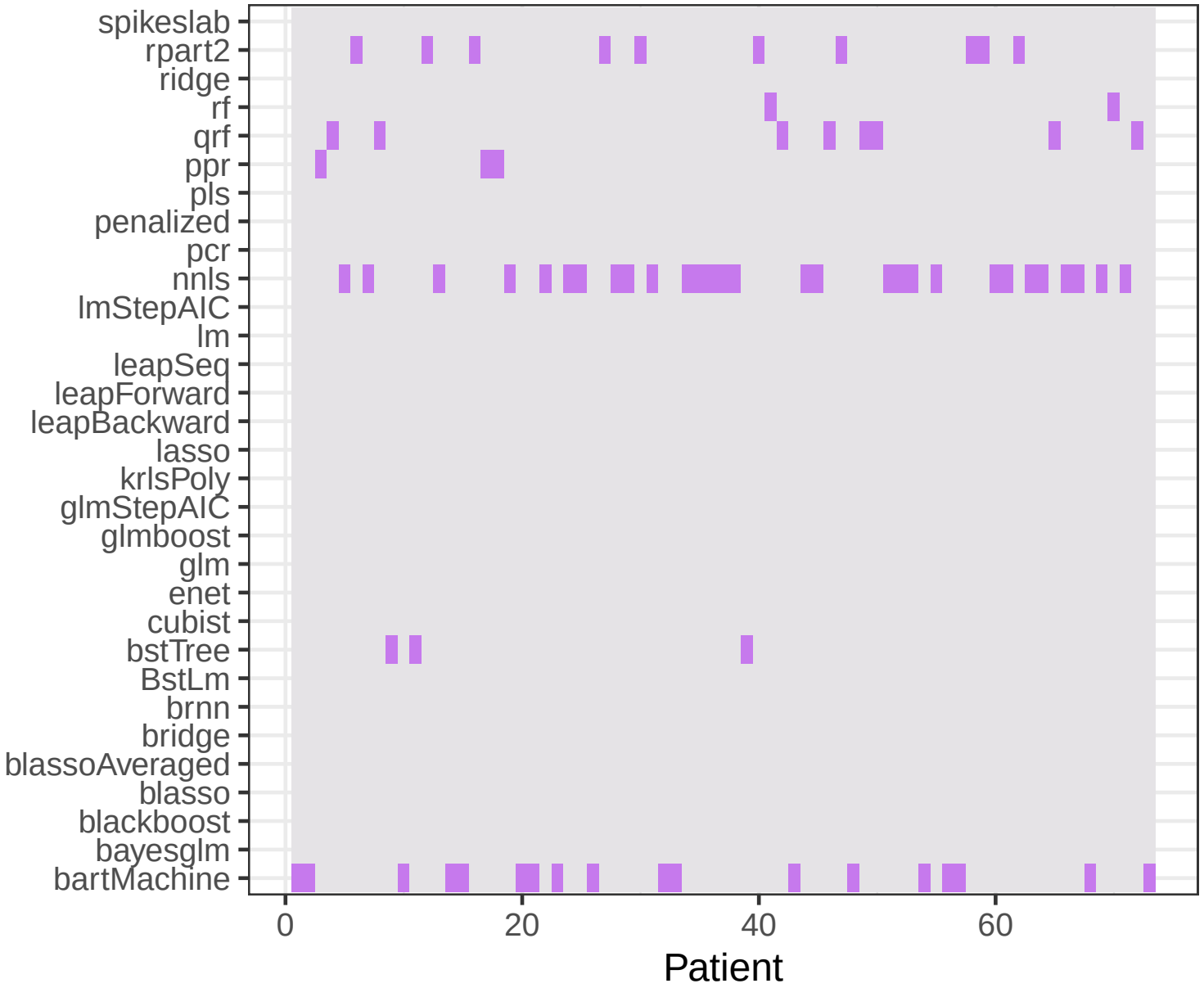
Difficult cases Non-linear Complex $n = 500$

only **one model** captures the PITE direction



Difficult cases Non-linear Complex $n = 1000$

only **one model** captures the PITE direction



Conclusions

- This work evidences potential approaches for different contexts.

Conclusions

- This work evidences potential approaches for different contexts.
- An analyst should know what characteristics their dataset presents.

Conclusions

- This work evidences potential approaches for different contexts.
- An analyst should know what characteristics their dataset presents.
- Risk is more variable.

Conclusions

- This work evidences potential approaches for different contexts.
- An analyst should know what characteristics their dataset presents.
- Risk is more variable.
- RF approaches perform poorly in terms of risk across non-trivial scenarios.

Conclusions

- This work evidences potential approaches for different contexts.
- An analyst should know what characteristics their dataset presents.
- Risk is more variable.
- RF approaches perform poorly in terms of risk across non-trivial scenarios.
- Effectiveness of Regularization in high-correlation scenarios.

Conclusions

- This work evidences potential approaches for different contexts.
- An analyst should know what characteristics their dataset presents.
- Risk is more variable.
- RF approaches perform poorly in terms of risk across non-trivial scenarios.
- Effectiveness of Regularization in high-correlation scenarios.
- Methods based on trees (blackboost, bartMachine, rpart2) present typically low risk in non-linear cases, while linear approaches struggle with risk in simple non-linear data.

Conclusions

- This work evidences potential approaches for different contexts.
- An analyst should know what characteristics their dataset presents.
- Risk is more variable.
- RF approaches perform poorly in terms of risk across non-trivial scenarios.
- Effectiveness of Regularization in high-correlation scenarios.
- Methods based on trees (blackboost, bartMachine, rpart2) present typically low risk in non-linear cases, while linear approaches struggle with risk in simple non-linear data.
- Leave complex methods for large sample sizes.

References I

Thomas Jaki, Chi Chang, Alena Kuhlemeier, and M. Lee Van Horn. Predicting individual treatment effects: Challenges and opportunities for machine learning and artificial intelligence. *KI - Künstliche Intelligenz*, January 2024. ISSN 1610-1987. doi: 10.1007/s13218-023-00827-4. URL http://dx.doi.org/10.1007/s13218-023-00827-4.

Andrea Lamont, Michael D Lyons, Thomas Jaki, Elizabeth Stuart, Daniel J Feaster, Kukatharmini Tharmaratnam, Daniel Oberski, Hemant Ishwaran, Dawn K Wilson, and M Lee Van Horn. Identification of predicted individual treatment effects in randomized clinical trials. *Statistical Methods in Medical Research*, 27(1):142–157, March 2016. doi: 10.1177/0962280215623981. URL https://doi.org/10.1177/0962280215623981.

Comparing Modern Machine Learning Methods for Predicting Individual Treatment Effect

Pamela Solano[1] and Thomas Jaki[1,2]

[1]Regensburg University, [2]Cambridge University

July 24, 2024