Unlocking the Code: Harnessing Machine Learning to Predict Treatment Resistance in Lung Cancer Patients

Johnson&Johnson

Fabian Kreimendahl

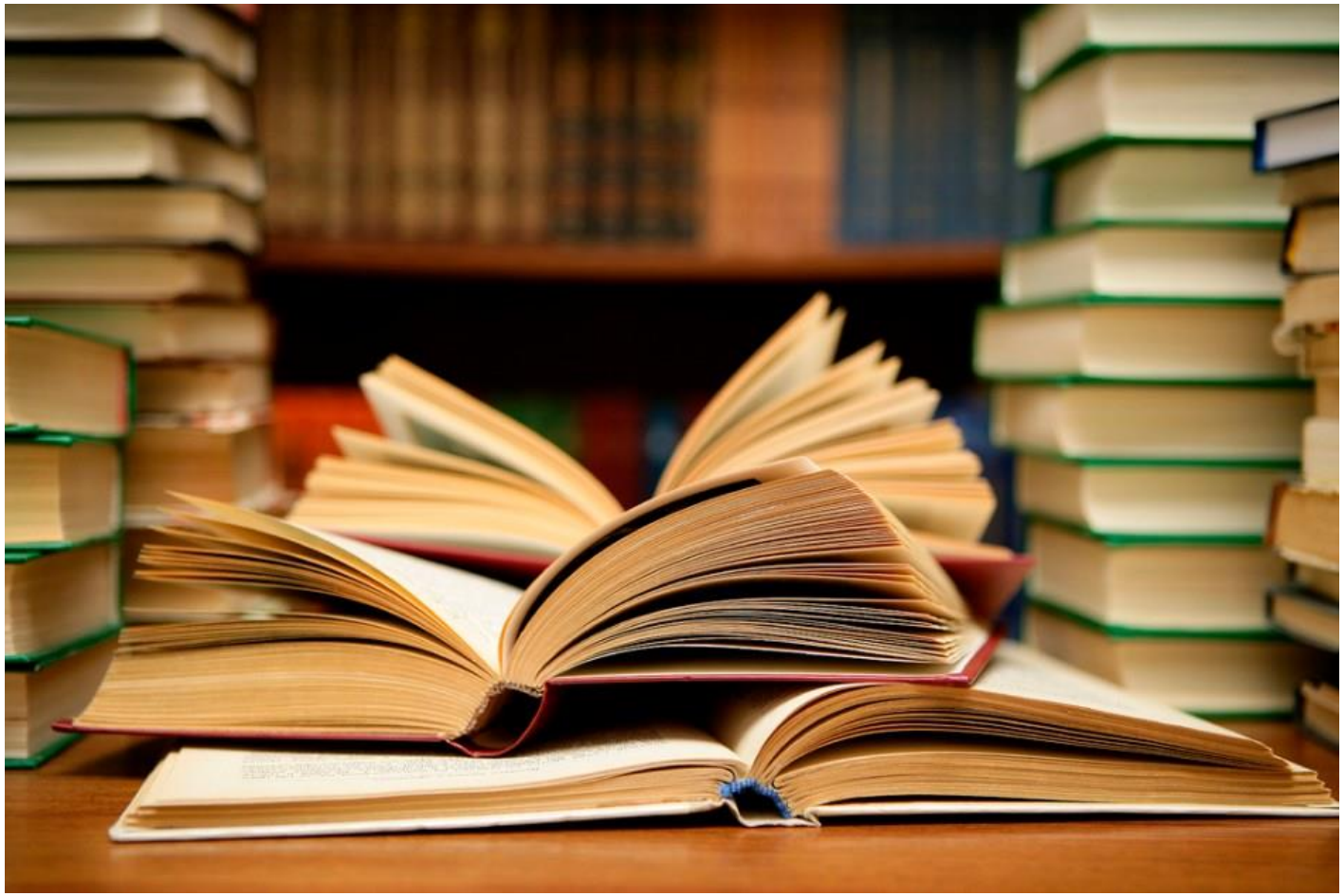
AI in Clinical Research and Drug Development
BBS Basel
Sep 25, 2024

Methodology

Johnson&Johnson
Innovative Medicine

Johnson&Johnson
Innovative Medicine

Literature Search & Review

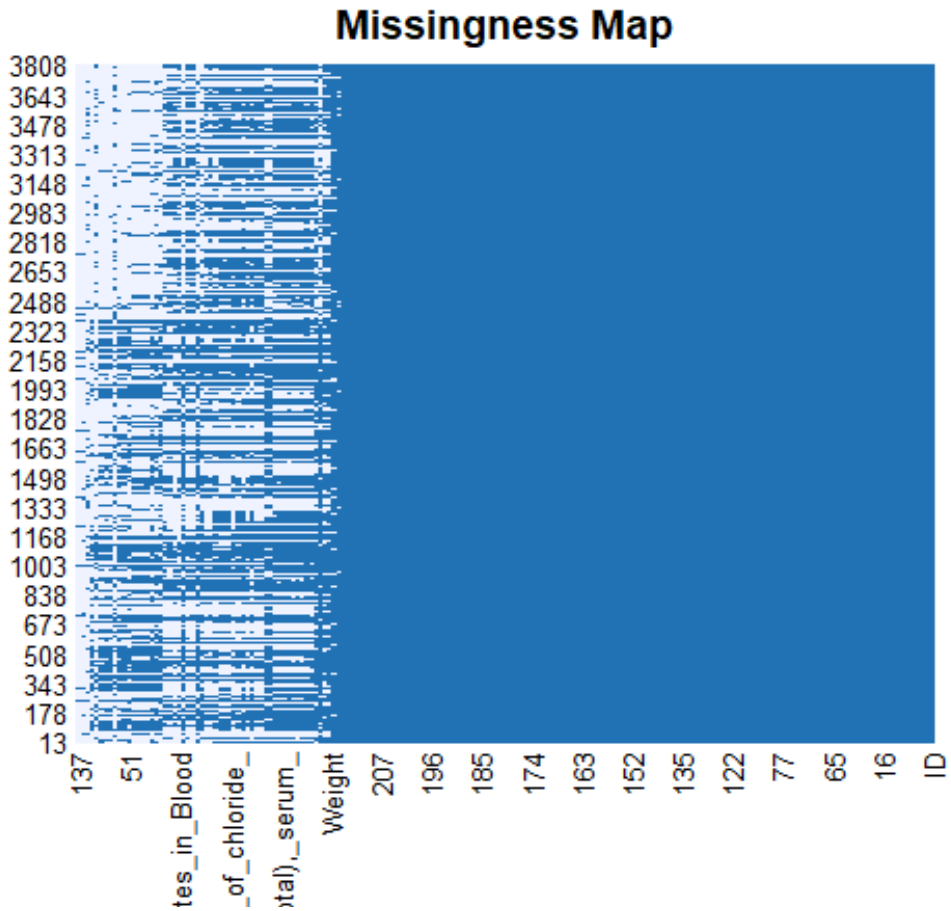


“Molecular Mechanisms”

Johnson&Johnson
Innovative Medicine

Literature Search & Review

Data Imputation?



899 of 3808 with complete data on relevant Vars

Missing (16%)
Observed (84%)

We decided to
- not impute missing data
- use **only complete observations**

Sensitivity analysis with imputation:
Mean / Mod / Median / Expectation Maximization Bootstrapping

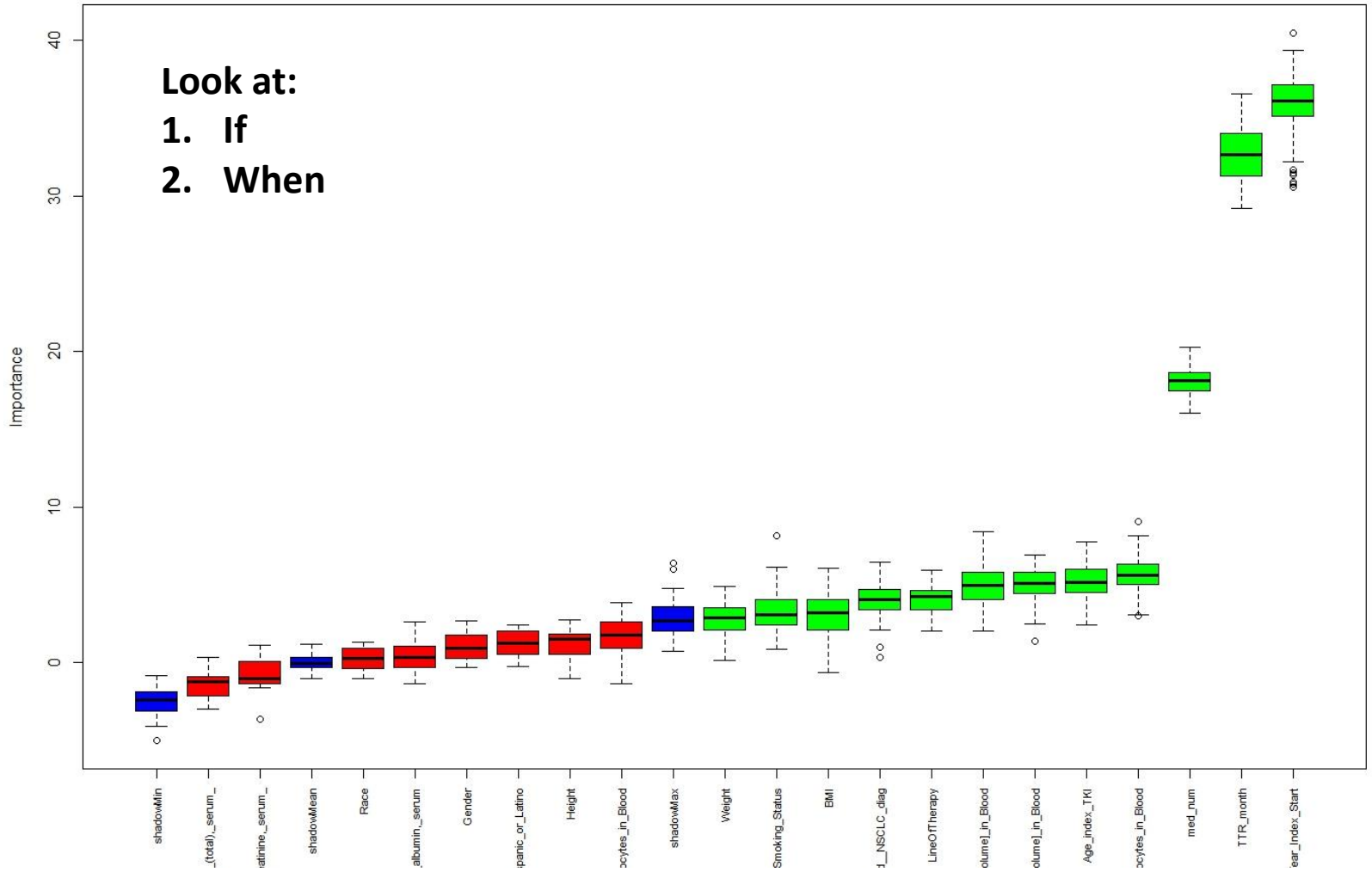
-> No gain in AUC / Brier

Johnson&Johnson
Innovative Medicine

Literature Search & Review

Data Imputation?

Boruta Feature Selection



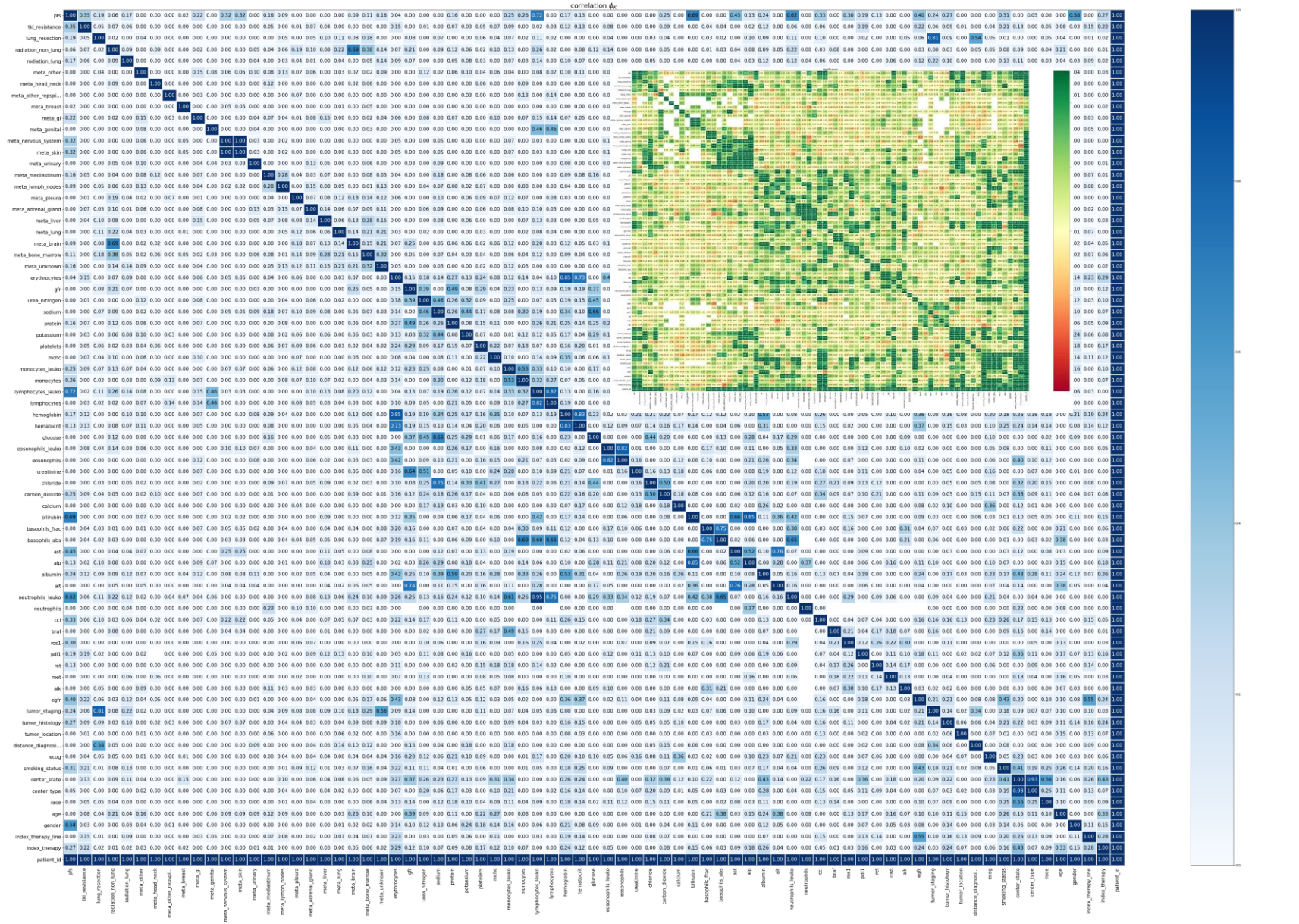
Johnson&Johnson
Innovative Medicine

Literature Search & Review

Data Imputation?

Boruta Feature Selection

PhiK Correlation Analysis



Johnson&Johnson
Innovative Medicine

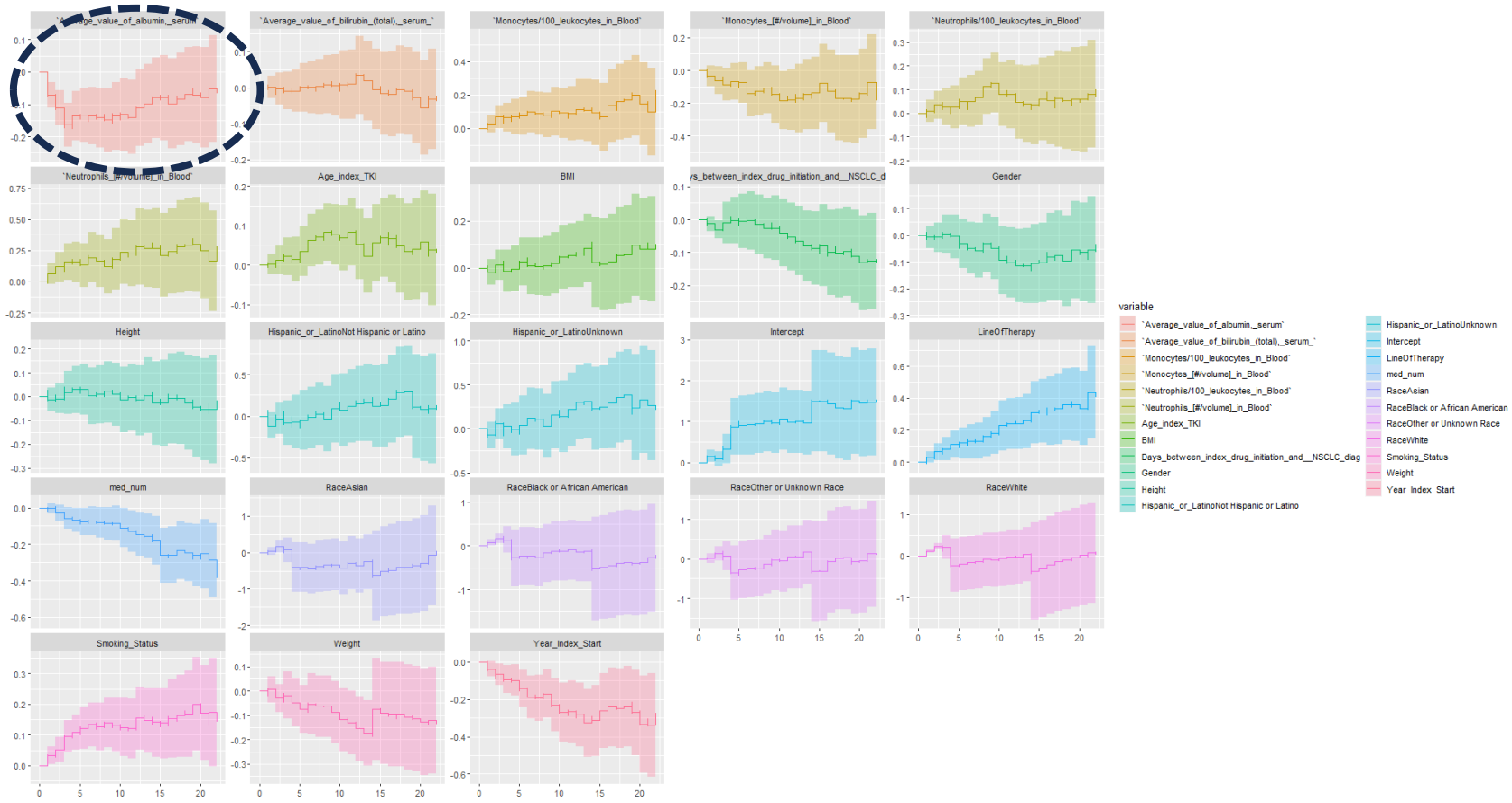
Literature Search & Review

Data Imputation?

Boruta Feature Selection

PhiK Correlation Analysis

Aalen's Additive Regression Model



Johnson&Johnson
Innovative Medicine

Age
Gender
Height
Weight
Race
Year Index Start

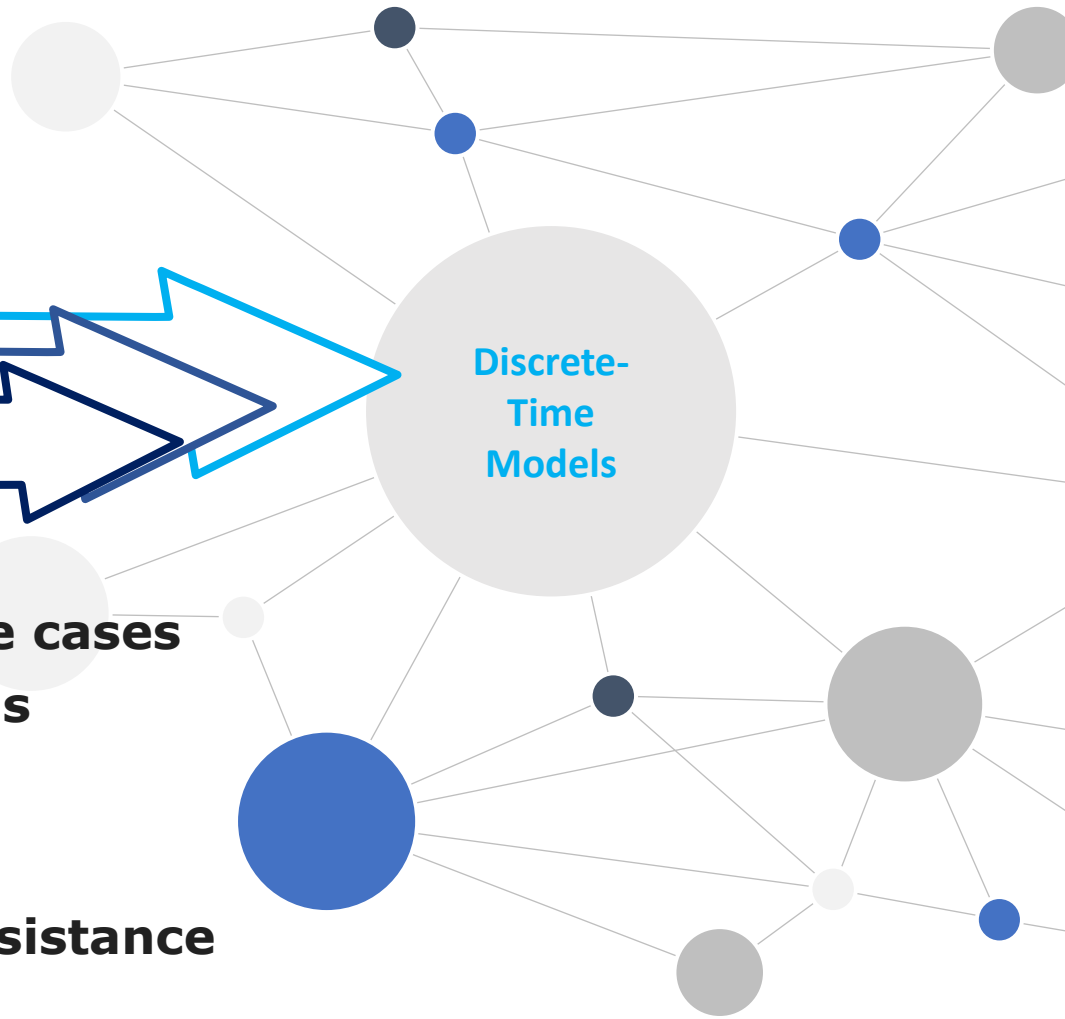
Demo graphic

Smoking Status
Line of Therapy
Medication (Afatinib, Erlotinib, Osimertinib)
Days bewteen Initiation & NSCLC Diagnosis

Medical

Labora-tory

Neutrophils #/volume in blood
Neutrophils /100 leukocytes in blood
Ø Albumin
Ø Bilirubin
Ø Creatine
Monocytes #/volume in blood
Monocytes /100 leukocytes in blood

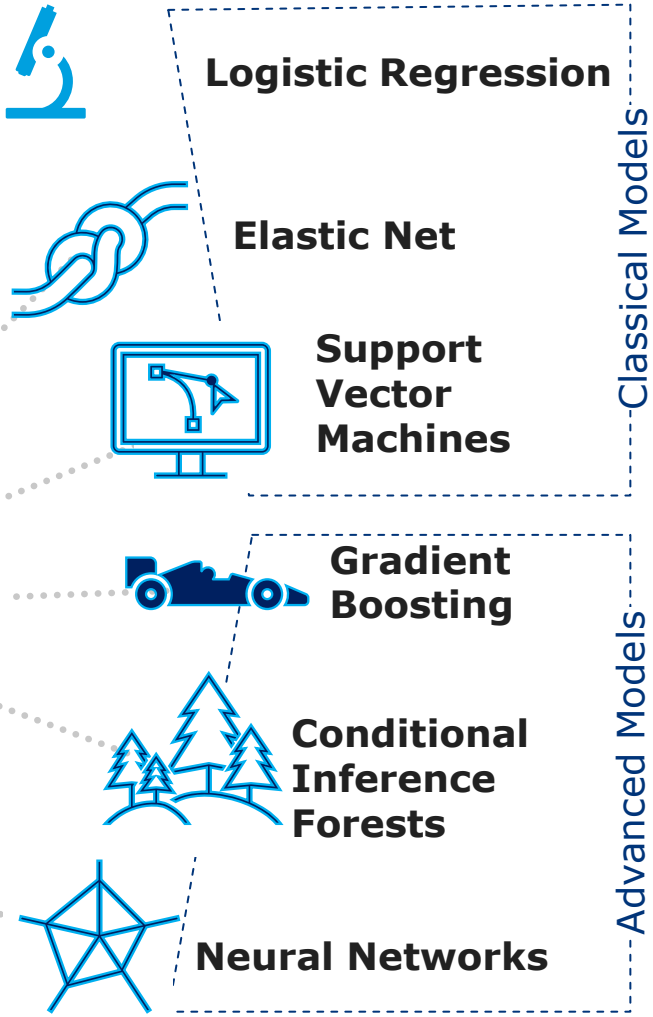


N=899 complete cases for analysis

Predicting TKI resistance

Johnson&Johnson
Innovative Medicine

Discrete-Time Model options*



- **80 / 20 Training / Testing Split**
- **10 fold Cross Validation**
- **Max predicted Intervals set to 25 (Longest Obs: 78m)**
- **Time frame not capped**
- **Hyperparameter tuning optimizes predictive accuracy (Brier Score)**
- **No. of Bayesian Optimization process iterations not limited**

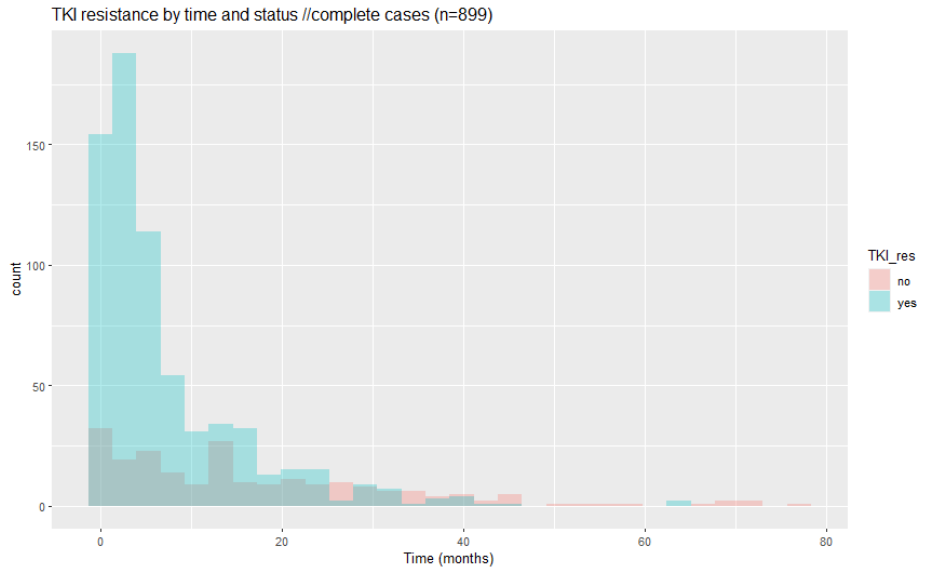
The explicit goal was to
Maximize AUC
Minimize Brier

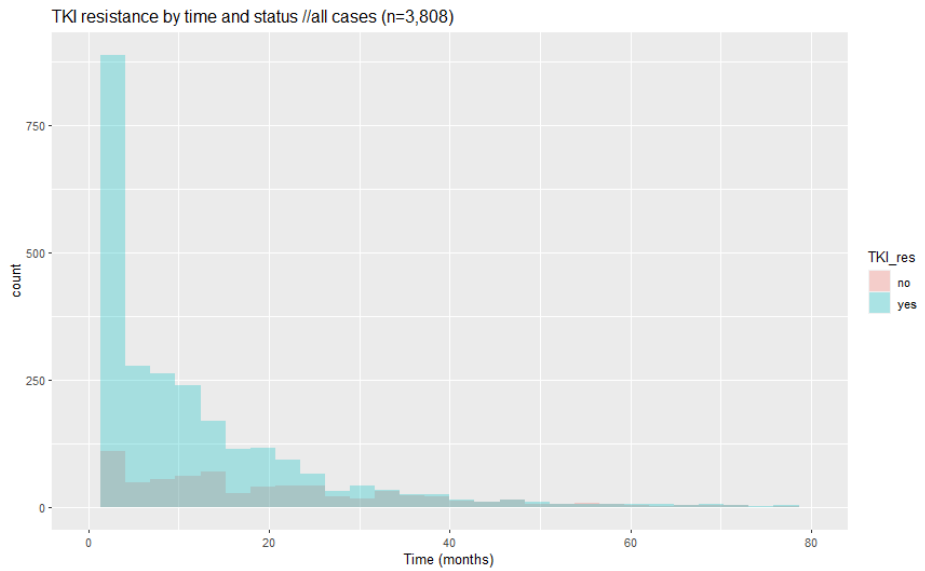
*Cox & Random Survival Forest are time continuous -> included in analysis only as sensitivity analysis

Results

Johnson&Johnson
Innovative Medicine

Johnson&Johnson
Innovative Medicine



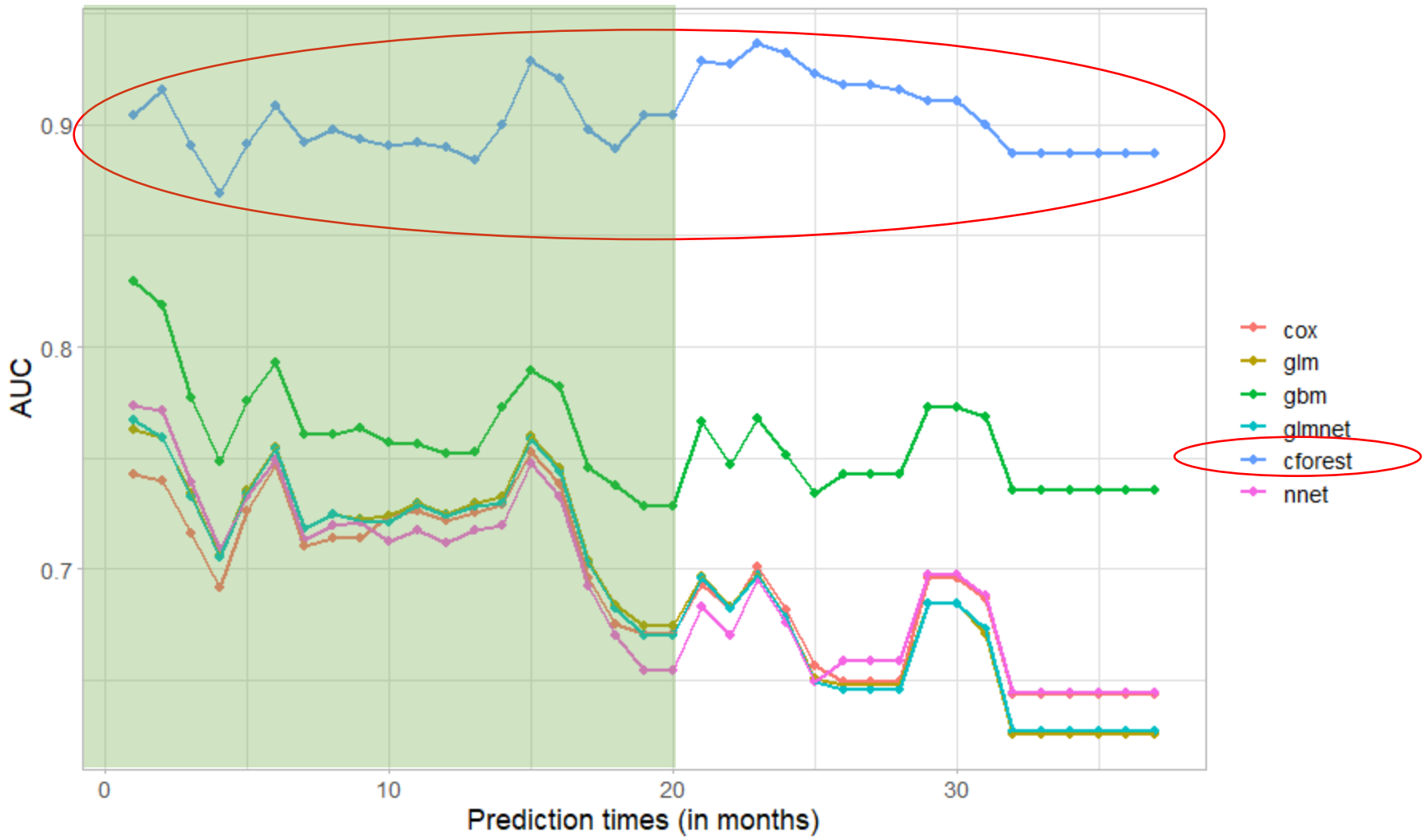


N=899 complete cases for analysis

	Complete Cases (n=899)	All cases (n=3,808)
Time To Resistance (months)	M = 9.5 SD = 11,7 med = 5	M = 13.1 SD = 18.7 med = 6
TKI Resistance	75.6%	79.0%
Age	67.9 y (±10.1)	67.2 y (±10.5)
Gender (female)	552 (61.4%)	2,419 (63.5%)
Weight	74.8 kg (±19.4)	74.7 kg (±22.9)
Height	165.5 cm (±10.1)	164.6 cm (±10.3)

Johnson&Johnson
Innovative Medicine

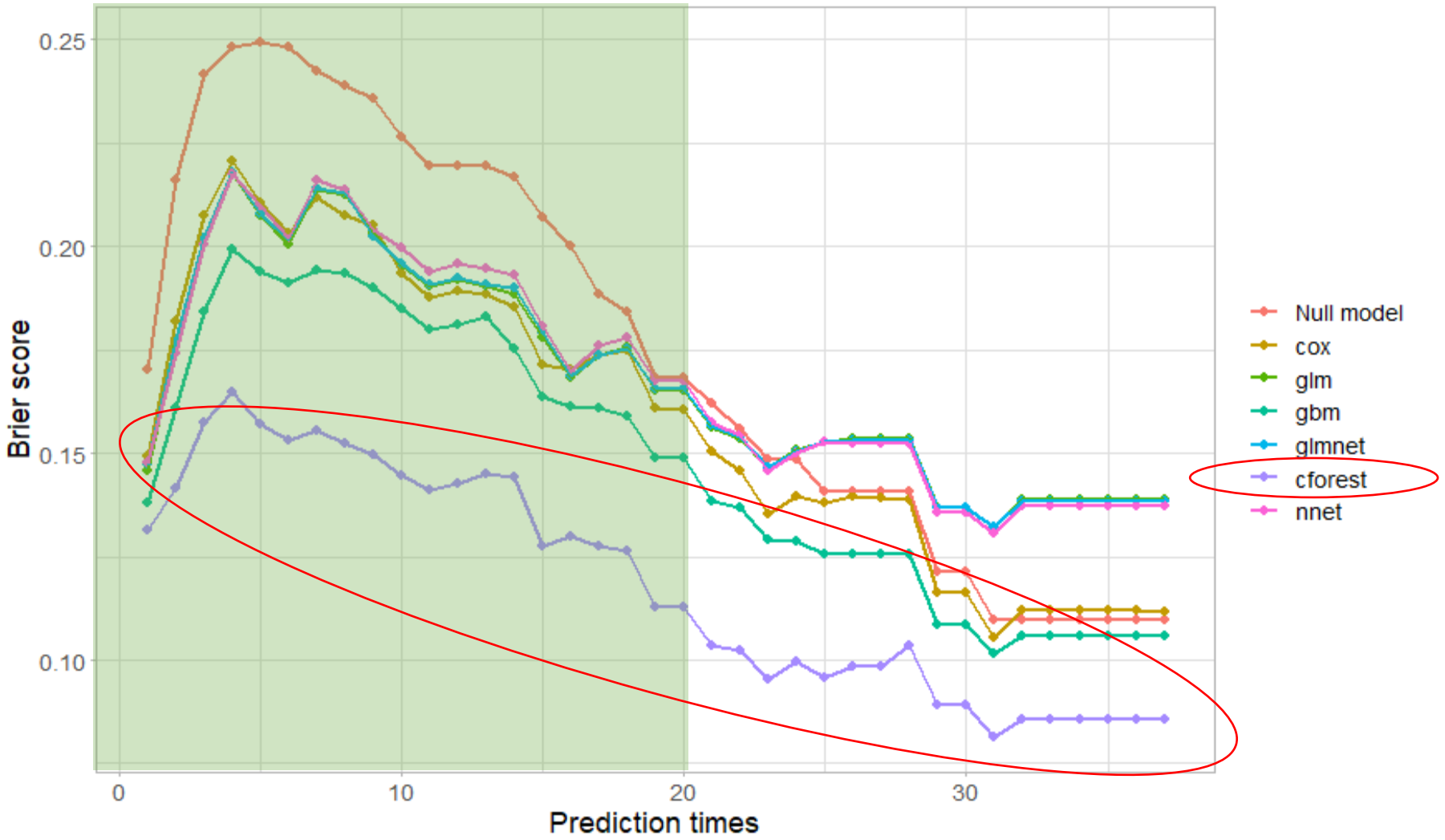
AUC Analysis



Johnson&Johnson
Innovative Medicine

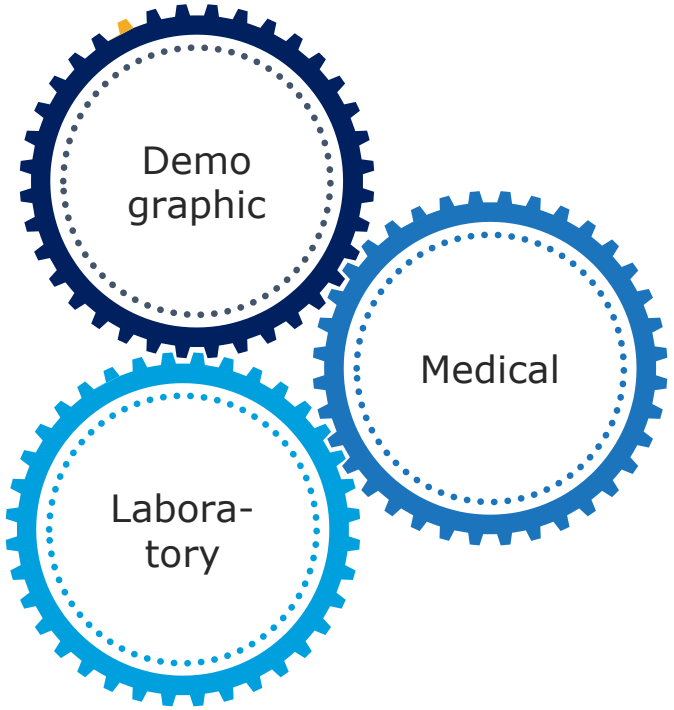
AUC Analysis

Brier Score Analysis

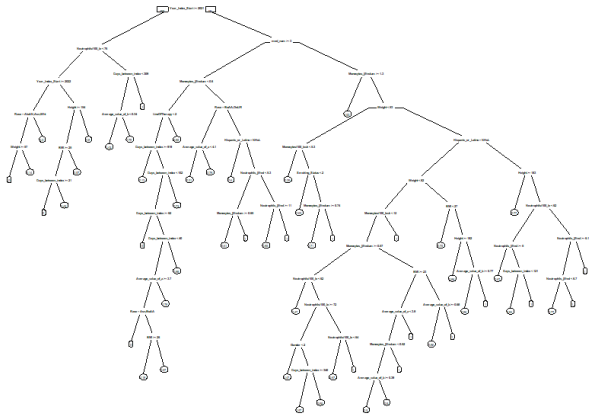


Slide to be viewed in presentation mode

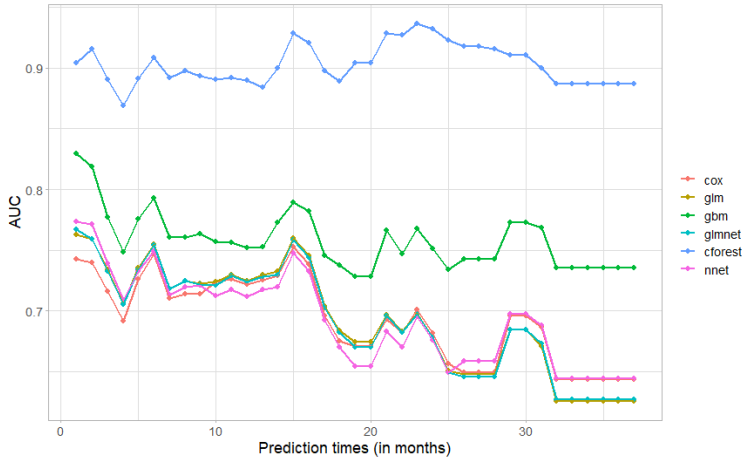
Summary



Conditional Inference Forests

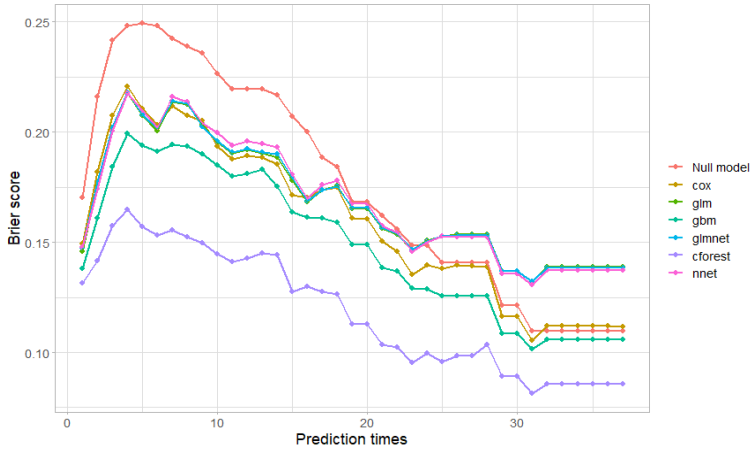


~90% AUC



(higher is better)

~0.14 Brier Score



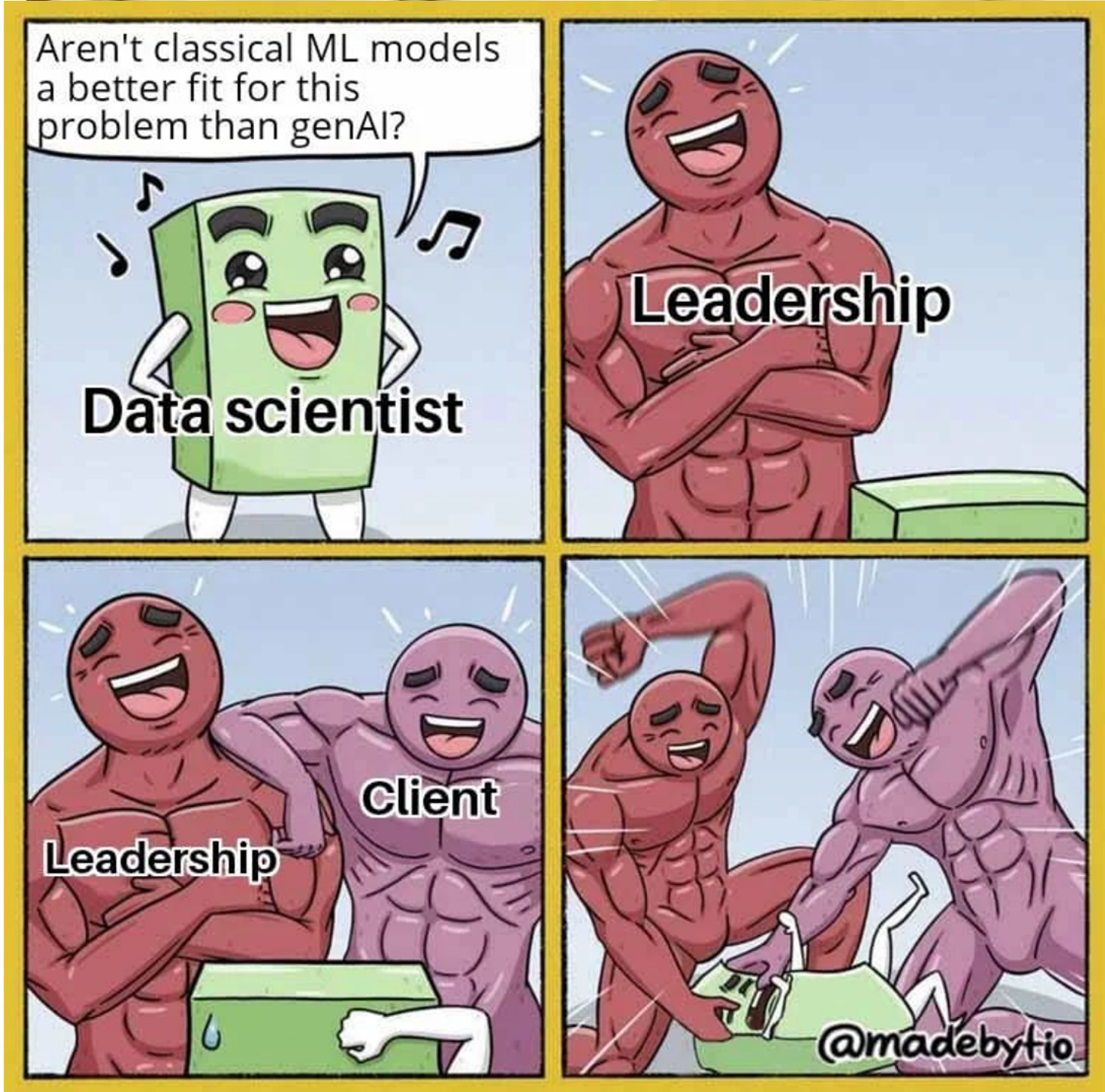
(lower is better)

Take aways

- Focus construction of ML Algorithms on **Feature Selection**
- Data Imputation is not a must – consider **trade-off quality vs. quantity**
- Approach model selection open-minded
  exclude only those that **violate assumptions**
- Fine Tune, Fine Tune, Fine Tune…

Johnson&Johnson
Innovative Medicine

Q&A



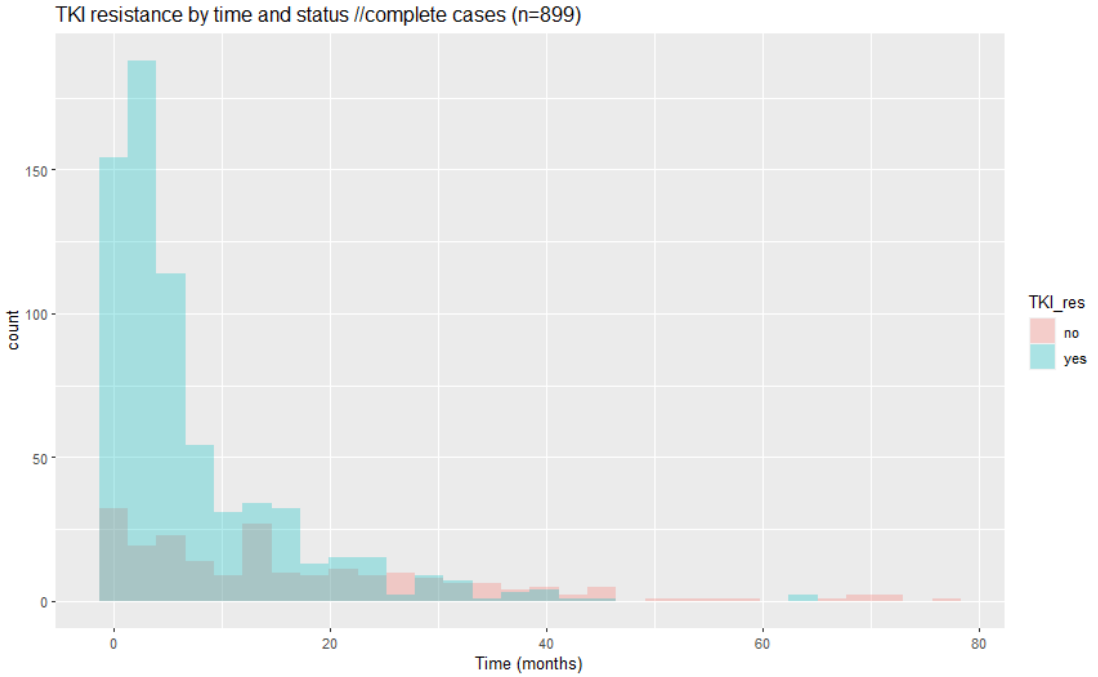
Johnson&Johnson
Innovative Medicine

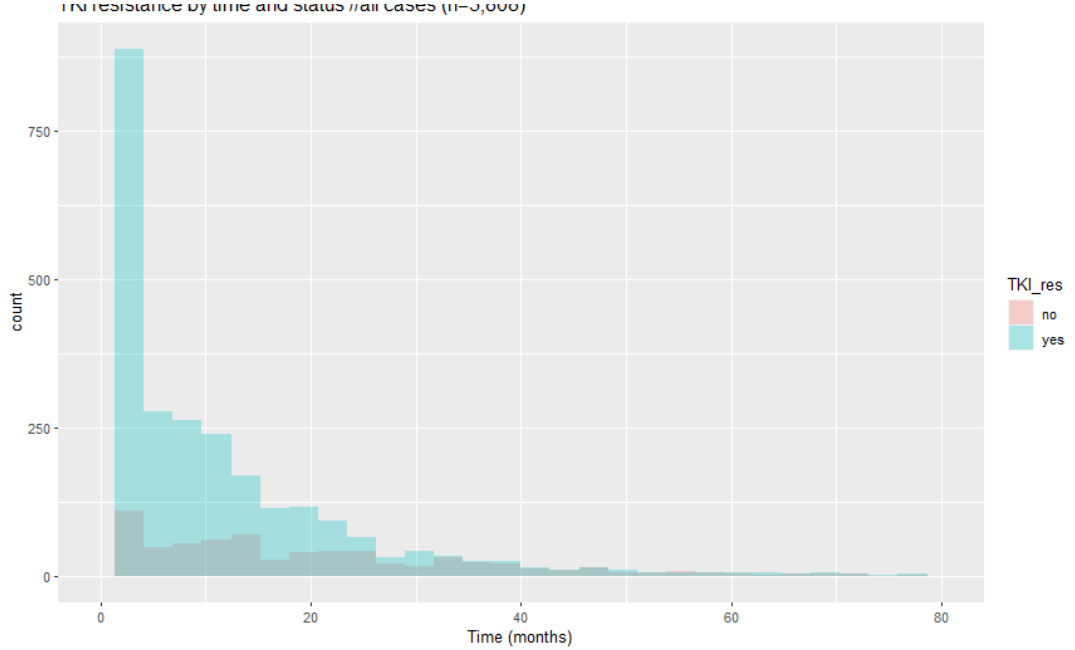
Back-Up

Johnson&Johnson
Innovative Medicine

Q: How different is the complete cases population from the overall population?

	Complete Cases (n=899)	All cases (n=3,808)
Time To Resistance (months)	M = 9.5 SD = 11,7 med = 5	M = 13.1 SD = 18.7 med = 6
TKI Resistance	75.6%	79.0%
Age	67.9 y (±10.1)	67.2 y (±10.5)
Gender (female)	552 (61.4%)	2,419 (63.5%)
Weight	74.8 kg (±19.4)	74.7 kg (±22.9)
Height	165.5 cm (±10.1)	164.6 cm (±10.3)





What‘s the logic behind conditional forests?

•An implementation of the random forest and bagging ensemble algorithms utilizing conditional inference trees as base learners.

•The main idea behind CForest is that many trees are built in parallel between the same start and goal states. The key concepts of **Conditional Inference Forests** are:

•Every time a tree finds a better solution, it is shared with all other trees so that all trees have the best solution found so far.

•Trees are expanded into regions that are known to be beneficial. Samples that cannot lead to a better solution are immediately discarded.

•Trees are pruned every time a better solution is found. Those states in the tree that do not help to find a better solution are removed from the tree.

What are the parameter results of the winning model?

```
[1] "Optimizing DiscreteTime-cforest"
elapsed = 68.39 Round = 1   intervals = 22.0000   mtry = 4.0000   Value = -0.225178
elapsed = 27.24 Round = 2   intervals = 8.0000    mtry = 8.0000   Value = -0.2180285
elapsed = 34.30 Round = 3   intervals = 12.0000   mtry = 9.0000   Value = -0.2169
elapsed = 55.68 Round = 4   intervals = 19.0000   mtry = 13.0000  Value = -0.2181103
elapsed = 65.51 Round = 5   intervals = 18.0000   mtry = 17.0000  Value = -0.2207953
elapsed = 30.75 Round = 6   intervals = 8.0000    mtry = 13.0000  Value = -0.2223858
elapsed = 25.33 Round = 7   intervals = 7.0000    mtry = 17.0000  Value = -0.2245203
elapsed = 37.94 Round = 8   intervals = 15.0000   mtry = 7.0000   Value = -0.2166563
elapsed = 9.84  Round = 9   intervals = 5.0000    mtry = 4.0000   Value = -0.2254332
elapsed = 64.54 Round = 10  intervals = 24.0000   mtry = 14.0000  Value = -0.2211411
elapsed = 47.33 Round = 11  intervals = 19.0000   mtry = 9.0000   Value = -0.2180478
elapsed = 30.25 Round = 12  intervals = 12.0000   mtry = 6.0000   Value = -0.2185194
elapsed = 74.73 Round = 13  intervals = 25.0000   mtry = 19.0000  Value = -0.2192173
elapsed = 31.78 Round = 14  intervals = 13.0000   mtry = 8.0000   Value = -0.2169926
elapsed = 41.70 Round = 15  intervals = 16.0000   mtry = 11.0000  Value = -0.2186583
elapsed = 59.67 Round = 16  intervals = 25.0000   mtry = 12.0000  Value = -0.2207182
elapsed = 34.31 Round = 17  intervals = 15.0000   mtry = 2.0000   Value = -0.2453982
elapsed = 56.11 Round = 18  intervals = 25.0000   mtry = 6.0000   Value = -0.2221802
elapsed = 43.88 Round = 19  intervals = 14.0000   mtry = 19.0000  Value = -0.2216234
elapsed = 13.25 Round = 20  intervals = 5.0000    mtry = 10.0000  Value = -0.2194843
elapsed = 69.28 Round = 21  intervals = 25.0000   mtry = 18.0000  Value = -0.2195931
elapsed = 46.32 Round = 22  intervals = 15.0000   mtry = 15.0000  Value = -0.2209005
elapsed = 16.22 Round = 23  intervals = 5.0000    mtry = 19.0000  Value = -0.2189276
elapsed = 12.02 Round = 24  intervals = 5.0000    mtry = 6.0000   Value = -0.2188034
elapsed = 60.34 Round = 25  intervals = 25.0000   mtry = 9.0000   Value = -0.2181861
elapsed = 31.41 Round = 26  intervals = 10.0000   mtry = 11.0000  Value = -0.2200767
elapsed = 48.13 Round = 27  intervals = 18.0000   mtry = 6.0000   Value = -0.2168099
elapsed = 79.76 Round = 28  intervals = 19.0000   mtry = 7.0000   Value = -0.2195975
elapsed = 49.95 Round = 29  intervals = 23.0000   mtry = 10.0000  Value = -0.2162741
elapsed = 52.02 Round = 30  intervals = 20.0000   mtry = 19.0000  Value = -0.2234684

 Best Parameters Found:
Round = 29      intervals = 23.0000             mtry = 10.0000 Value = -0.2162741
```

Q: Is the sample size sufficiently large to calculate efficient prediction models?

- We chose quality over quantity when analyzing TKI resistance:
  - Data imputation has not show improvements in AUC / Brier
  - The initial dataset with n=3,808 was not tremendously large either way
  - Analysis of baseline variables shows homogeneity of “complete cases” & “total” population
- Conditional Random forest are especially useful when dealing with “small n, large p” situations,

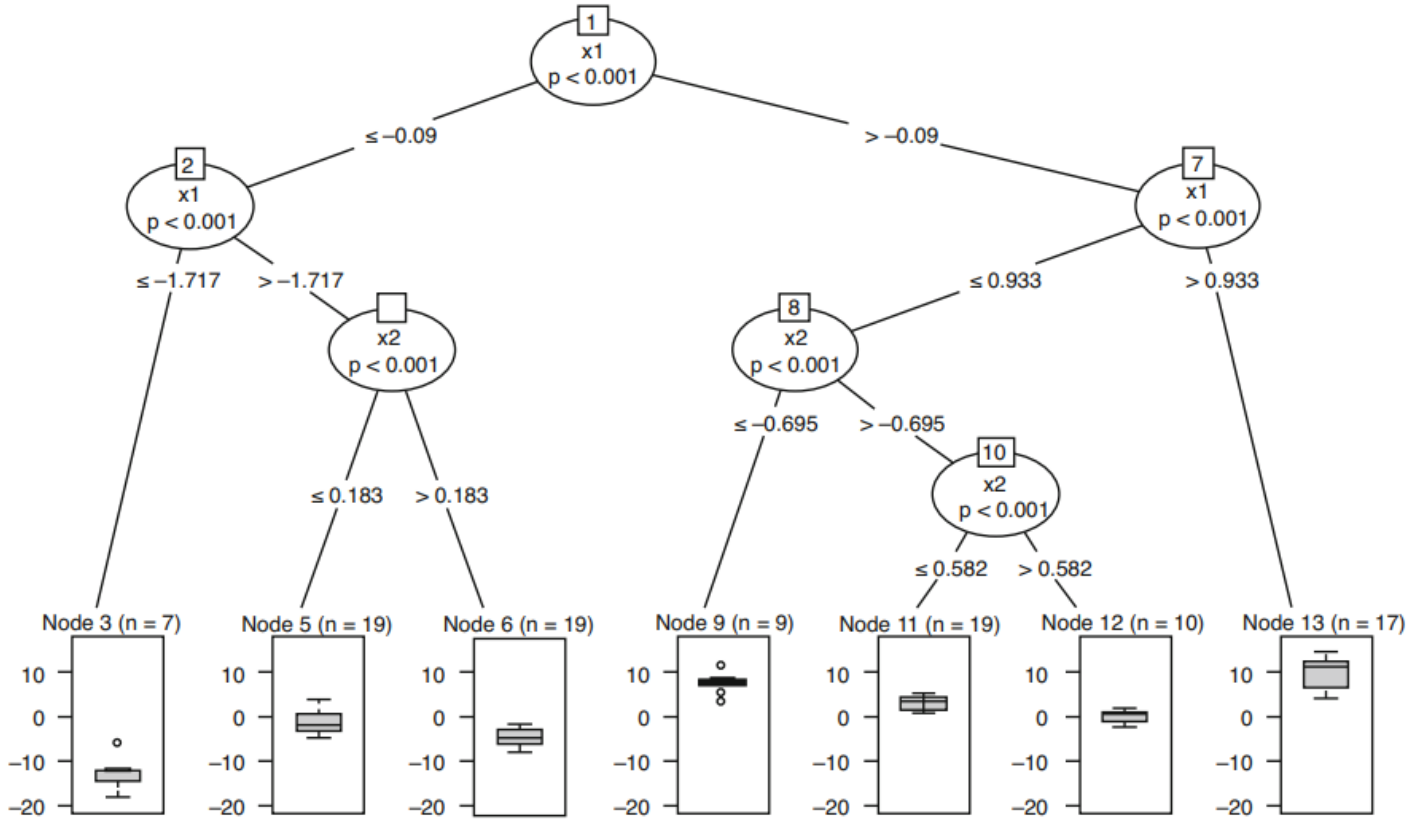
i.e. when parametric models are problematic.

- Potential issue of overfitting: Debatable!

Q: What does a conditional forest look like? What are the difference to random forests?

- Every time a tree finds a better solution, it is shared with all other trees so that all trees have the best solution found so far.

- Trees are expanded into regions that are known to be beneficial. Samples that cannot lead to a better solution are immediately discarded.



Q: What are „mtry“ and „intervals“ at cforest?

- An “Interval” is just the number of the optimal number of classification buckets from the model
- “Mtry” is the **number of input variables randomly sampled as candidates** at each node for random forest like algorithms.