A semi-automatic machine learning pipeline for identification of prognostic and predictive factors in cancer trial data

Juliane Manitz, Aslihan Gerhold-Ay, Pascal Kieslich (EMD Serono/Merck Healthcare KGaA Darmstadt Germany)

Summary

We investigate heterogeneity in treatment effect and survival by searching for subpopulations that will particularly benefit from a novel immuno-oncology (IO) treatment in comparison to conventional chemotherapy. We build a machine learning (ML) analysis pipeline identifying prognostic and predictive factors for time-to-event endpoints that can be applied easily to new data from phase 3 clinical trials. Time-to-event outcomes are typically not supported by existing ML pipelines.

Data

Phase 3 clinical trial data from a case-control study comparing a novel immuno-oncology (IO) drug with conventional chemotherapy.

Definition as ML Problem

Dependent variable: Overall survival, the time from treatment initiation to death from any cause; patients alive or lost to follow-up are censored.

Features: Treatment group, patient demographics, disease characteristics, laboratory values, patient-reported outcomes, biomarkers, etc.

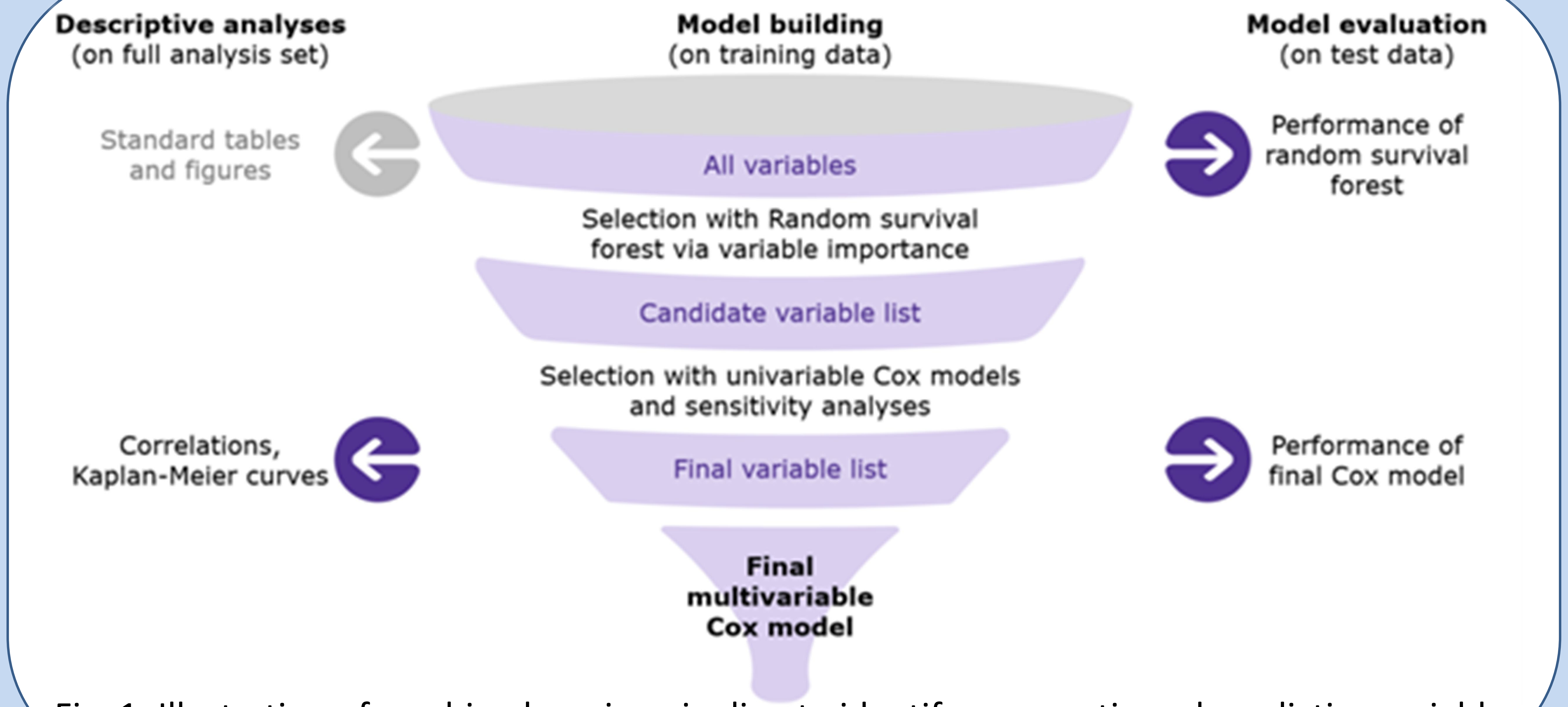
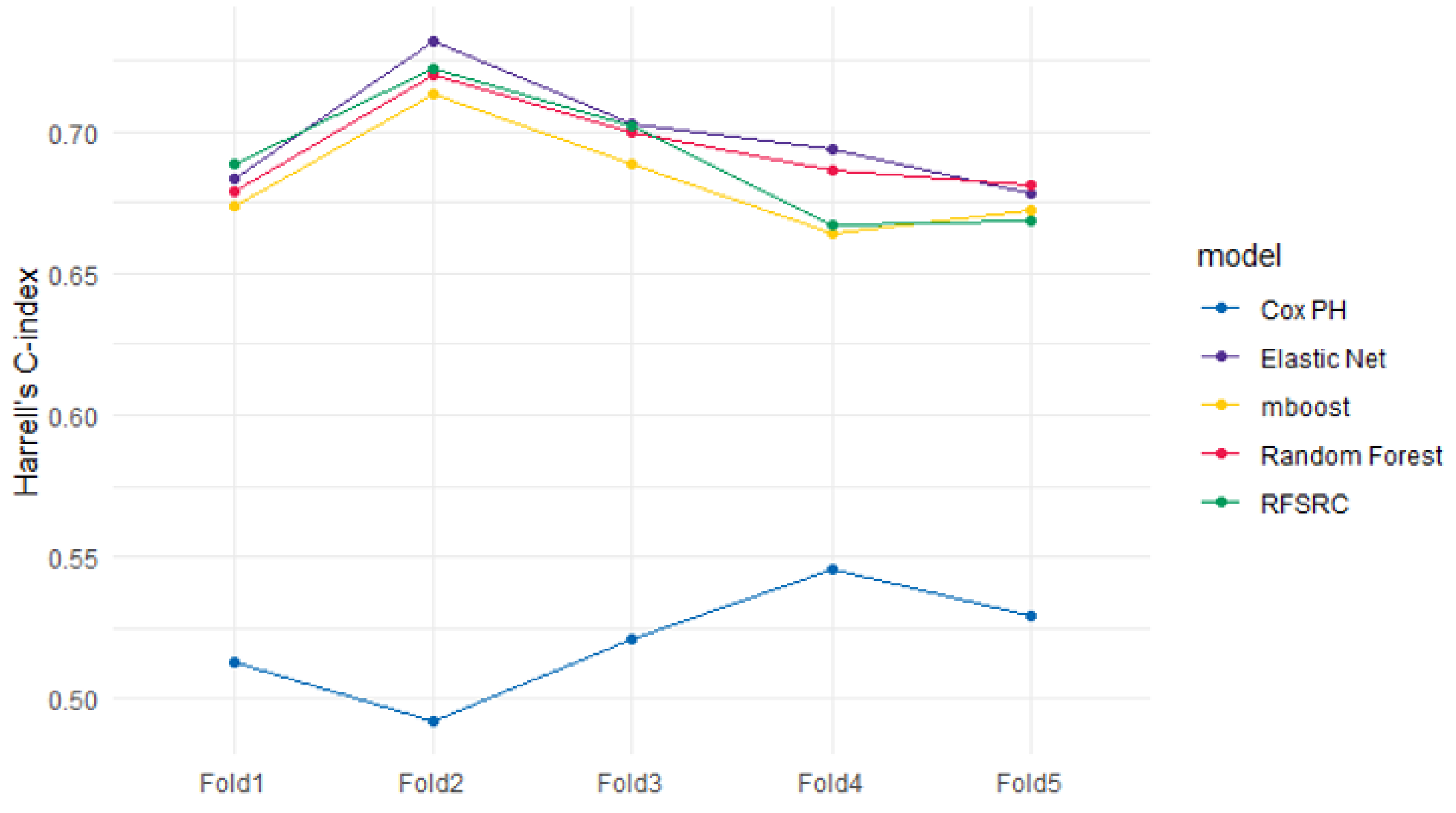

Descriptive analyses
(on full analysis set)
Standard tables and figures
Correlations, Kaplan-Meier curves
Model building
(on training data)
All variables
Selection with Random survival forest via variable importance
Candidate variable list
Selection with univariable Cox models and sensitivity analyses
Final variable list
Final multivariable Cox model
Model evaluation
(on test data)
Performance of random survival forest
Performance of final Cox model

Fig. 1: Illustration of machine learning pipeline to identify prognostic and predictive variables

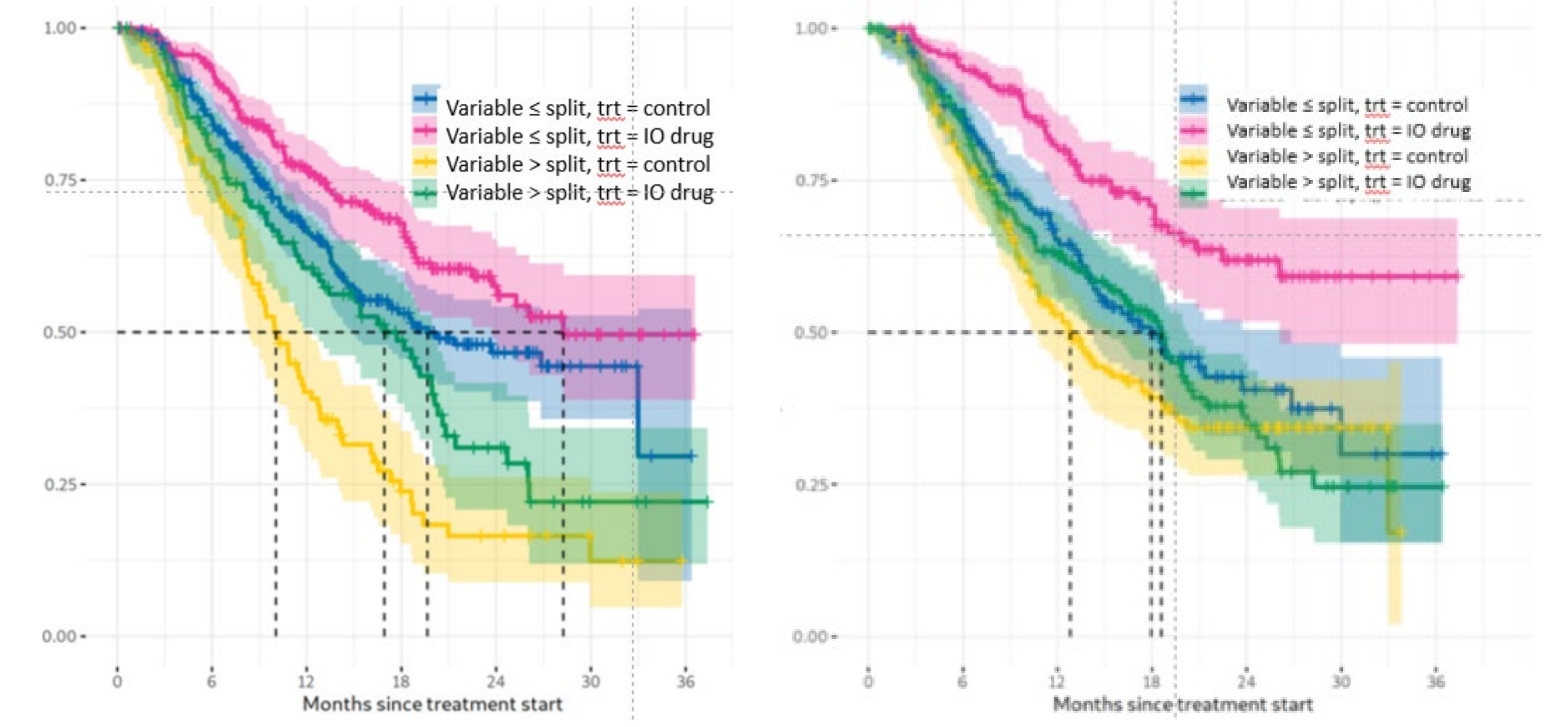
Methods & Results

The machine learning pipeline consists of the following steps (see Fig. 1):

1. Data preparation, split into training and test data, and imputation of missing values
2. Benchmark random forest with other ML algorithms incl. elastic net, boosting and simple Cox model with known predictors



3. Build the predictive model using a random forest which allows covariate selection of potentially high-dimensional interactions
   a. pooled data across treatment arms
   b. each treatment arm independently
4. Develop candidate variable list with variables selected in at least two models using permutation-based variable importance measures
5. Univariate modelling for each variable with treatment interaction to differentiate between prognostic and predictive effects



6. Fit a final Cox proportional hazard model to provide clinical insights and interpretability.

Following data preparation and specification of covariate candidates, all analysis steps are automated. Results are provided in interactive analysis reports.

Conclusion & Discussion

We exemplify the machine learning pipeline using phase III clinical trial data and successfully identify possible prognostic and predictive factors. However, this is an exploratory data analysis, which is performed with the purpose of hypotheses generation and gaining internal insights. Results of subpopulation analysis might be subject to increased probability of type I and II errors and further (clinical) validation is pending.