

Title: Evaluation of random forest as an approach for covariate selection

Author: Klaus Lindauer, Thomas Frank, Heiner Speth, Ashley Strougo

Institution: Sanofi-Aventis Deutschland GmbH, R&D Operations/Translational Medicine and Early Development (TMED), Pharmacokinetics, Pharmacodynamics and Metabolism Germany

Type: poster

Objectives: Stepwise covariate modeling is a commonly used/accepted procedure for covariate search. This procedure consists of a forward inclusion followed by a backward deletion of covariates from a model. The forward inclusion is a very tedious and lengthy procedure, thereby, substituting it by a faster procedure is expected to markedly fasten the search for covariates. Random forest is a faster classification procedure because it does not require fitting of the whole data but rather focus on the individually estimated η 's where various uncorrelated decision trees are randomly investigated using R or Python [1-4]. The objective of this work is to evaluate the use of random forest as a search algorithm to substitute the forward inclusion step from the stepwise covariate modeling.

Methods: The work was divided in three main steps being the first one the performance of clinical trial simulations of pharmacokinetic profiles using a one-compartment 'hypothetical true model(s)' including body weight and estimated glomerular filtration rate (eGFR) on elimination rate and body weight on volume of distribution (using power functions). In total 10 different scenarios with 100 clinical trials of 300 patients each, were simulated varying the impact of eGFR (exponent varied from 0 to 2) and the unexplained inter-individual variability (IIV) on the elimination rate and volume of distribution (ω defined as 0.1 or 0.3). For the clinical trial simulations, the covariates were sampled from a clinical trial database [5] where the following covariates were recorded at baseline: age, race, body weight, body mass index, low-density lipoprotein, triglyceride, cholesterol, serum gamma glutamyl transferase, serum glutamic-pyruvic transaminase, serum glutamic oxaloacetic transaminase, creatinine, eGFR, and creatinine clearance. In the second step, the simulated pharmacokinetic profiles were fitted using the 'hypothetical base model' (without covariates). Finally, the individually estimated η s for each fitted model was analyzed using random forest. The importance of the hit-list of covariates was evaluated using percentage increase in mean squared errors (%IncMSE) and the total decrease in node impurities. Higher ranked %IncMSE values in combination with outlier detection were used to select covariates. The results of the random forest for each trial was compared with the covariates in the 'hypothetical true model(s)' to define how often 'true' and how often 'false' covariates were selected.

Results Body weight as covariate on volume of distribution was selected in 95 to 100% of all simulated scenarios. Random forest could select body weight on elimination rate constant in 2 to 93% of all simulated scenarios dependent on impact of eGFR and unexplained IIV. Evaluation of eGFR on elimination rate was identified in about 75% of the simulated clinical trials where the investigated scenarios assumed low impact of covariates (exponent equal to 0.25) and in about 100% for the scenarios where mid to high impact was assumed (exponent between 0.5 to 2). In less than 50% of the simulated clinical trials additional false positive covariates were identified. However in these cases only one additional false positive covariate was identified. The scenarios where no covariates were considered (exponent equal to 0), random forest identified correctly no or maximal one false positive covariate in about 98% of simulated clinical trials.

Conclusion: The use of random forest seems to be suitable as a search algorithm to substitute the forward inclusion step from the stepwise covariate modeling. Further investigation comparing random forest with the forward inclusion step from the stepwise covariate modeling is ongoing.

References

- [1] Breiman, L. Random forests. Machine Learn. 2001, 45, 5–32
- [2] Liaw A, Wiener M. Classification and Regression by randomForest. R News 2002;2/3 (12):18–21.
- [3] RandomForestRegressor <https://github.com/7cthunder/RandomForestRegressor> (last accessed Feb 20th 2020)
- [4] Duke University, Durham, NC, USA.
<http://code.env.duke.edu/projects/mget/export/HEAD/MGET/Trunk/PythonPackage/dist/TracOnlineDocumentation/Documentation/ArcGISReference/RandomForestModel.FitToArcGISTable.html>, (last accessed Jan 30th 2020)
- [5] Theroux et al Circulation 2000; 102;3032-3038