MACHINE LEARNING-BASED COVARIATE SELECTION IN POPULATION PHARMACOKINETICS: A SYSTEMATIC EVALUATION OF BORUTA ALGORITHM WITH TREE-BASED METHODS AGAINST SCM+

Ibtissem Rebai (1), Stephen Duffull (1), Ayman Akil (1), Anna Largajolli* (1), Floris Fauchet* (1)
(1) Certara, Princeton, NJ, Radnor; * Contributed equally.

Boruta-LightGBM showed the most reliable covariate selection performance across both linear and TMDD models.

Adding Lasso pre-processing reduced false positive covariates and improved robustness in complex correlated scenarios compared with SCM+.

Background and Objective

Machine learning (ML) approaches have emerged as promising alternatives for covariate selection when applied to empirical Bayes estimates (EBEs) [1].

This study evaluates the Boruta feature selection algorithm [2] combined with four different tree-based ML models (i.e., Random Forest, XGBoost, LightGBM, and CatBoost), with or without Lasso pre-processing [3], and compares their performance to SCM+ [4]. across linear and target-mediated drug disposition (TMDD) models.

Methods

Simulation Design

- 2 popPK models:
  linear and TMDD 2CPT model (Qss approximation, IV)
- 6 covariate scenarios :
  None, Single, Multiple covariates
  × 100 replicated datasets × 180 subjects each
- 7 covariates from NHANES:
  Age, Sex, Race, Weight, Albumin, Creatinine, Hemoglobin
- Covariate effect:
  β = 0.75 (power & linear for continuous & categorical)
- EBEs obtained from NONMEM base model (FOCEI) used as ML inputs

==========

ML Methods Evaluated

- Boruta algorithm with 4 tree-based algorithm: Random Forest, XGBoost, LightGBM, CatBoost
- Lasso as a pre-screening step (λ.min and λ.1se via 5-fold cross-validation)

==========

Performance Metrics

- Type I error: false positive rate (Scenario 0 only)
- Score 1 (S1): proportion of datasets identifying the true covariate (with or without extras)
- Score 2 (S2): proportion of datasets with exclusive correct covariate identification

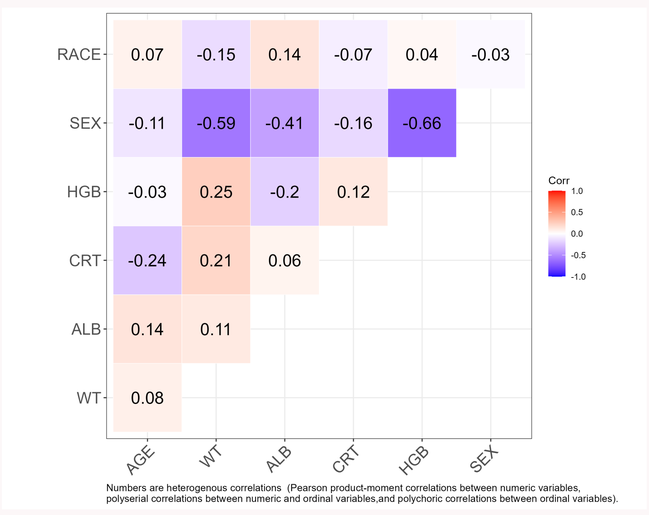
Results – Evaluation Study Design

- All parameters well estimated
- Low shrinkage in all scenarios except for the CL parameter in the TMDD model (shrinkage~ 20%) presenting moderate shrinkage

Correlation Covariate (Figure 2)

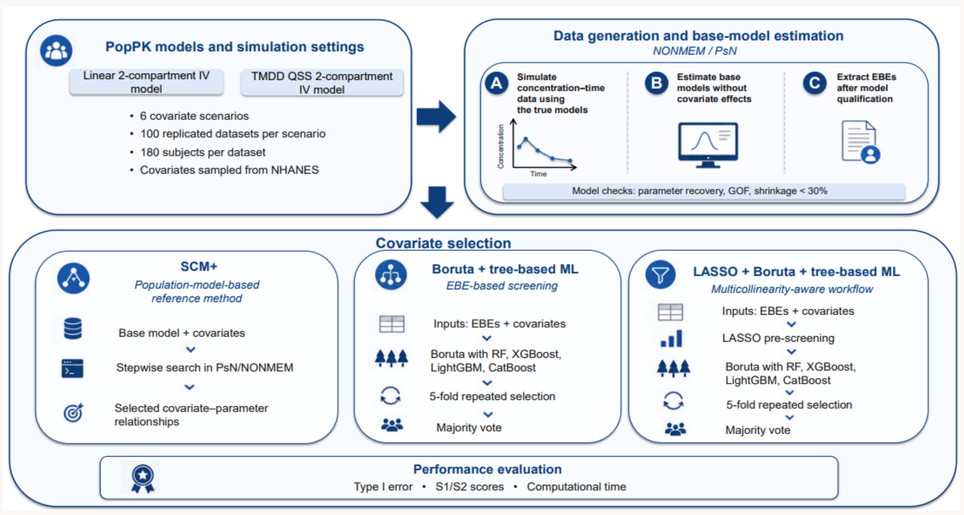
- High correlation between:
  Gender & Weight and Gender & Hemoglobin

Figure 2. Correlation matrix Between Covariates



Framework

Figure 1. Framework

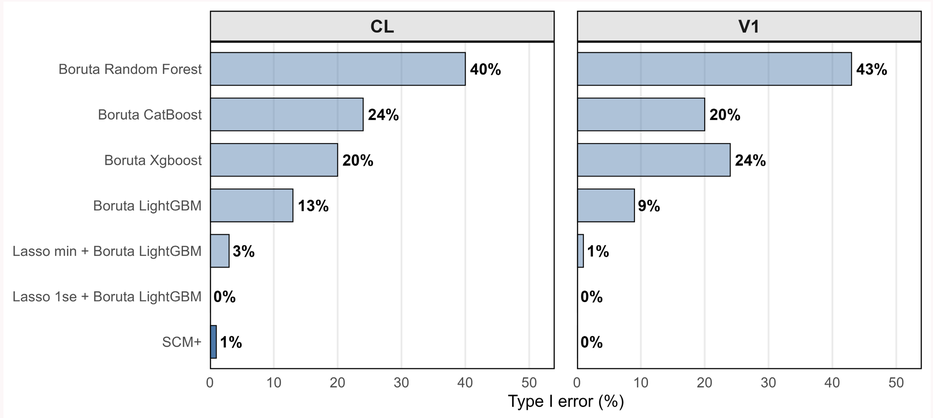


Results – Linear Model – ML vs SCM+

Type I Error Control (Scenario 0 — No true covariate — Figure 3)

SCM+: best control (<2% false positive rate)
Boruta LightGBM: lowest Type I error among ML methods (9–13%)
Boruta Random Forest: poor control (≥40%) => excluded
Lasso + Boruta LightGBM: Type I error reduced to <5%, comparable to SCM+

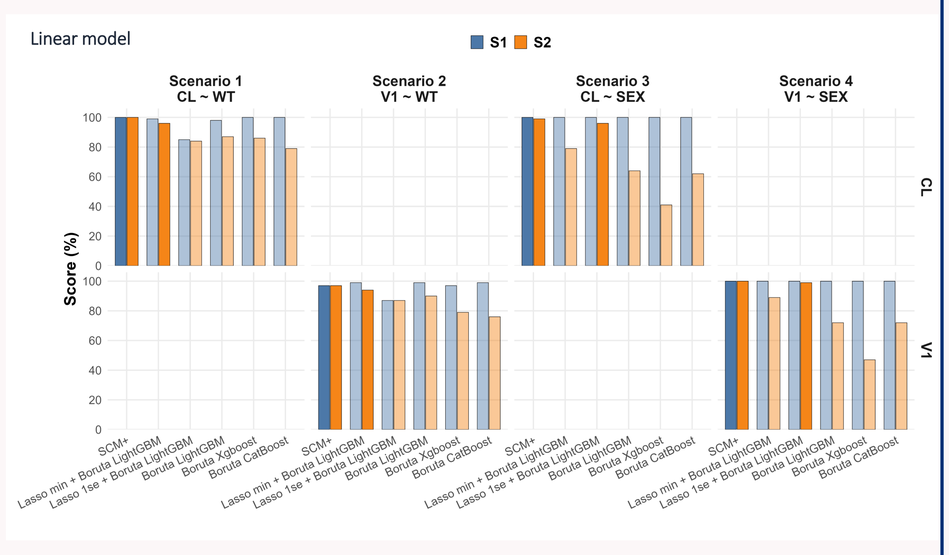
Figure 3. Type I error



Single Covariate Scenarios (Figure 4)

SCM+: near-perfect S1 and S2 scores (~100%) for all scenarios
Boruta LightGBM: S1 ~100%, S2 65–90% ; high true covariate detection, occasional extra covariates selected
Lasso (λ.min) + Boruta LightGBM: S1 and S2 approaching 100% for continuous covariates
Lasso (λ.1se) + Boruta LightGBM: highest S2 scores (96–99%) for categorical covariates

Figure 4. Covariate Selection Performance by Scenario (S1 and S2)

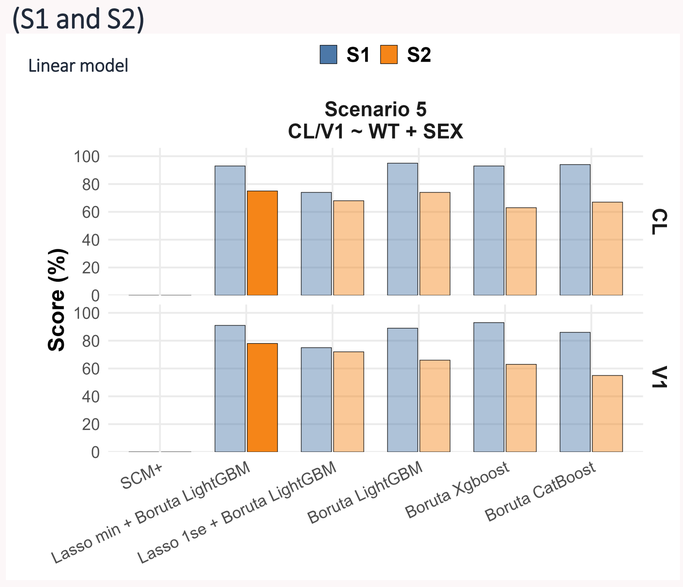


Darked shades represent SCM+ (all bars) and best Boruta method per panel (based on S2)

Multiple Covariate Scenario (Figure 5)

SCM+: S1 & S2 = 0% in 100% of datasets (missed WT effect systematically)
Boruta LightGBM: S1 ~95%, S2 ~75%
Lasso (λ.min) + Boruta LightGBM: S1 ~95%, S2 ~75%
--> best overall ML performance

Figure 5. Covariate Selection Performance for Scenario 5 (S1 and S2)

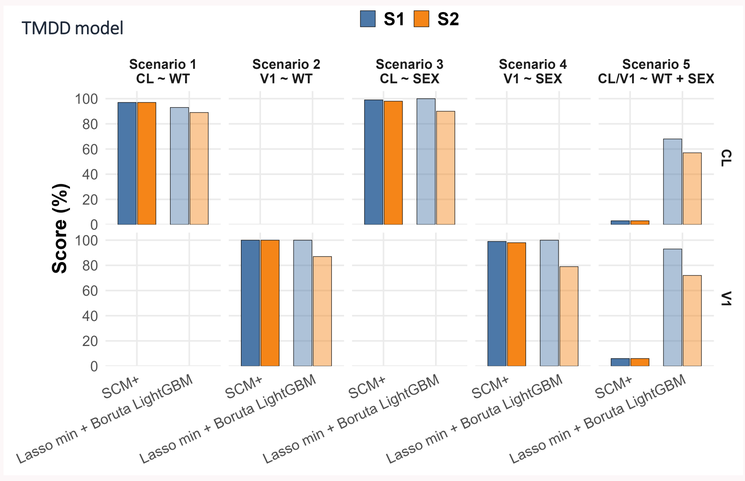


Darked shades represent best Boruta method per panel (based on S2)

Results – TMDD model (Figure 6) – ML vs SCM+

Lasso + Boruta LightGBM maintained strong performance on the complex TMDD model
Type I error remained <5%, comparable to SCM+
S2 (exact selection) slightly reduced vs. linear model (~7–10%), but true covariate detection maintained

Figure 6. Covariate Selection Performance by Scenario (S1 and S2) with Lasso Pre-processing



Darked shades represent SCM+

Discussion

SCM+ is highly effective for simple, single-covariate scenarios, with near-perfect detection rates and excellent Type I error control. SCM+ performs poorly when multiple correlated covariates are present in the model.

Lasso pre-screening dramatically improved ML false positive control by filtering correlated covariates before Boruta evaluation.

ML-based methods offer a substantial computational advantage over SCM+, reducing covariate selection runtime from hours to minutes.

Lasso + Boruta LightGBM is recommended as the ML approach for covariate selection.

[1] Sibieude et al. *J Pharmacokinet Pharmacodyn.* 2021;48(4):597-609.
[2] Kursa et al. *Fundamenta Informaticae.* 2010;101(4):271-285.
[3] Tibshirani. *J R Stat Soc Series B.* 1996;58(1):267-288.
[4] Ahamadi et al. *J Pharmacokinet Pharmacodyn.* 2019;46(3):273-285.