mlcov: R package for Covariate Selection Using Machine Learning

Ibtissem Rebai (1), Vincent Duval (1), Ayman Akil (1), James Craig (1), Mike Talley (1), Anna Largajolli* (1), Floris Fauchet*(1).
(1) Certara, Princeton, NJ, USA; * Contributed equally.

Package support : james.craig@certara.com; mike.talley@certara.com; ibtissem.rebai@certara.com

Background & Objective

- Previous work [1] evaluating the performance of Boruta algorithm (BOAL) [2] implemented in R [3] using XGboost in combination with Lasso regularized regression method [4] led us to establish a new framework for covariate selection.
- ***mlcov*** **R package (https://github.com/certara/mlcov) is now available to the pharmacometrics community.**
- **This work compares the *mlcov* R package and the traditional Stepwise Covariate Modeling (SCM) methodology [5] on a real-world data. Results of both approaches were compared with respect to covariates identified as clinically relevant.**

Methods

Machine Learning algorithms

➢ **Boruta**

1. Creates a **set of shadow covariates** generated by randomly permuting the values of the original covariates and compares their **importance scores** to the original covariates provided by training an XGBoost model (Fig. 1).
2. **Identify covariates** by repeating the process and evaluating the number of hits (=importance greater than the maximum importance of all shadow covariates) in binomial distribution to provide decision of covariate selection (Fig. 2).

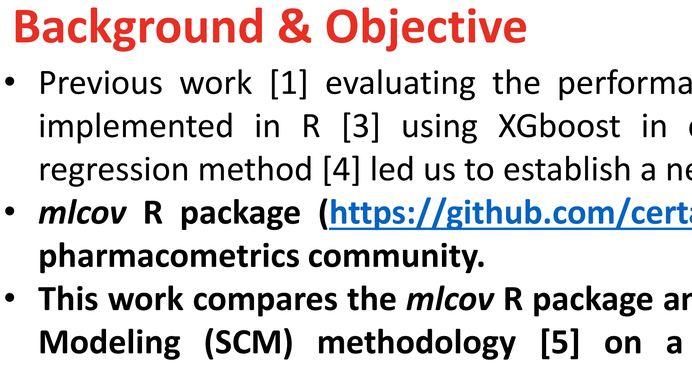

Fig. 1

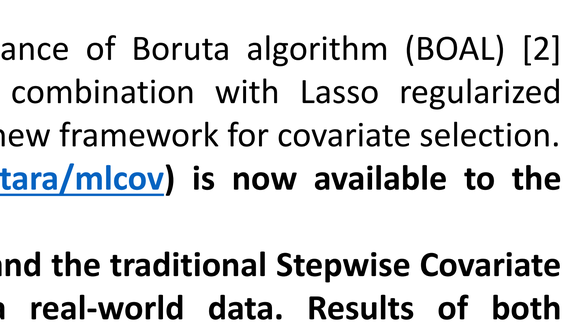

Fig. 2

➢ **XGBoost**

Gradient boosting technique that employs a series of **Decision Trees** to make predictions. It assigns an importance score to covariates, with more crucial one receiving higher scores.

➢ **Lasso**

Regularization techniques addressing collinearity in statistical modeling that is applied with the **glmnet package in R** before the BOAL to reduce **correlation** between covariates.

➢ **Majority Voting Ensemble (MVE)**

Covariate selection framework implemented to repeat the entire process on five random subsets of the dataset using **a voting mechanism** to obtain the final covariate selection.

mlcov package

Data splitting

- The dataset including the empirical Bayesian estimates of the individual parameters (EBEs) and the sets of the covariates is randomly split into five equal subsets (or folds).

Covariate selection

- 4/5 subsets (80%) are used to apply Lasso algorithm as a pre-processing step, followed by BOAL to select the relevant covariates. This process is repeated 5 times, with different folds used each time.

Voting mechanism

- The number of times each covariate is selected in the five folds is calculated. The covariates with the highest selection count (more than 2 times) are considered as the final selected covariates.

ml_cov_search()

Residuals Plot

- Residual plots are used: 1) to assess how chosen covariates capture data trends 2) to reveals potential overlooked trends with the unselected covariates.

generate_residuals_plot()

Real-world data

- PopPK model developed on Phase 2/ Phase 3 data including N=1957 patients.
- 14 covariates relationships tested for both SCM and *mlcov* (Tab. 1).

Parameters	Covariates tested
CL/F	weight, albumin, creatinine clearance (CRCL), sex, race, ethnicity
V/F	weight (WGT), albumin (ALB), sex, race, ethnicity (ETHN)
Ka	age, formulation (FORM), device

Tab. 1

Implementation *mlcov*

ID	CL	V	KA	WGT	ALB	SEX	...
1	0.41	5.9	0.30	64.2	4.7	0	
2	0.52	10.0	0.58	53.8	4.5	0	
3	0.39	5.6	2.70	58.1	4.4	0	
4	0.75	14.0	0.19	66.7	4.3	0	
5	0.34	5.7	1.10	47.7	4.3	0	
6	0.59	11.0	1.50	68.1	4.2	1	

```
devtools::install_github("certara/mlcov")
library(mlcov)

result_CL <- ml_cov_search(data = read.csv('Base_model_Outputs.csv',header = T), #NONMEM output (EBEs+cov)
                           pop_param =  c('CL'),
                           cov_continuous = c("WGT","CRCL","ALB"),
                           cov_factors = c("SEX","ETHNIC","RACE"))

residuals_CL <- generate_residuals_plot(data = read.csv('Base_model_Outputs.csv',header = T),
                                        result = result_CL,
                                        i = c('CL'))
```

Results

Tab. 2	SCM	*mlcov*
Number of covariate effect selected	9	6
Covariate rejected by user	1	0
Execution time	13h	5min

The parameter estimates are similar regarding set of covariates identified by the two methodologies (Tab. 3).

Tab. 3	ALB CL/F	CRCL CL/F	ETHN CL/F	Race CL/F	Sex CL/F	WGT CL/F	ALB Vc/F	WGT Vc/F
SCM selection	-1.35	0.17	0.23	0.19	0.1	0.66	-0.54	0.8
***mlcov* selection**	-1.13	0.18	-	0.2	-	0.73	-0.54	0.81
Importance Score	0.14	0.35	0.03	0.04	0.03	0.38	0.72	0.27

Sex and Ethnicity not selected by *mlcov,* likely due to their correlations with bodyweight and race (Fig. 4).

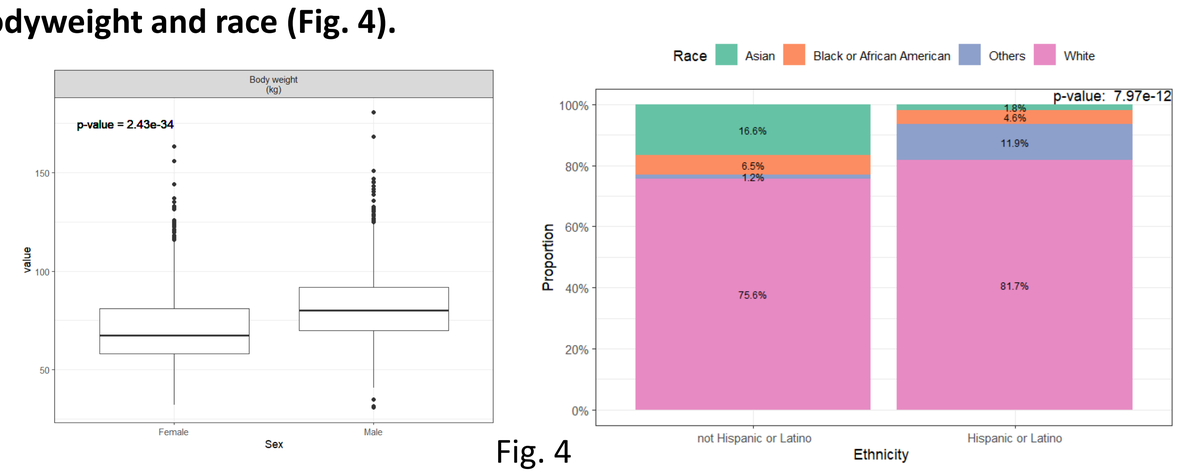

Fig. 4

Multivariate forest plots

- EBEs associated with individual set of covariates were used for this assessment. The parameter uncertainty and the residual variability were not considered.
- Covariates unselected by *mlcov* (Sex and Ethnicity) showed no clinical relevance (included in gray area covering the 0.8 to 1.25-fold change in exposure metric).
- **Similar trends are observed between both approaches resulting in same conclusions on the clinical relevance of the covariates.**

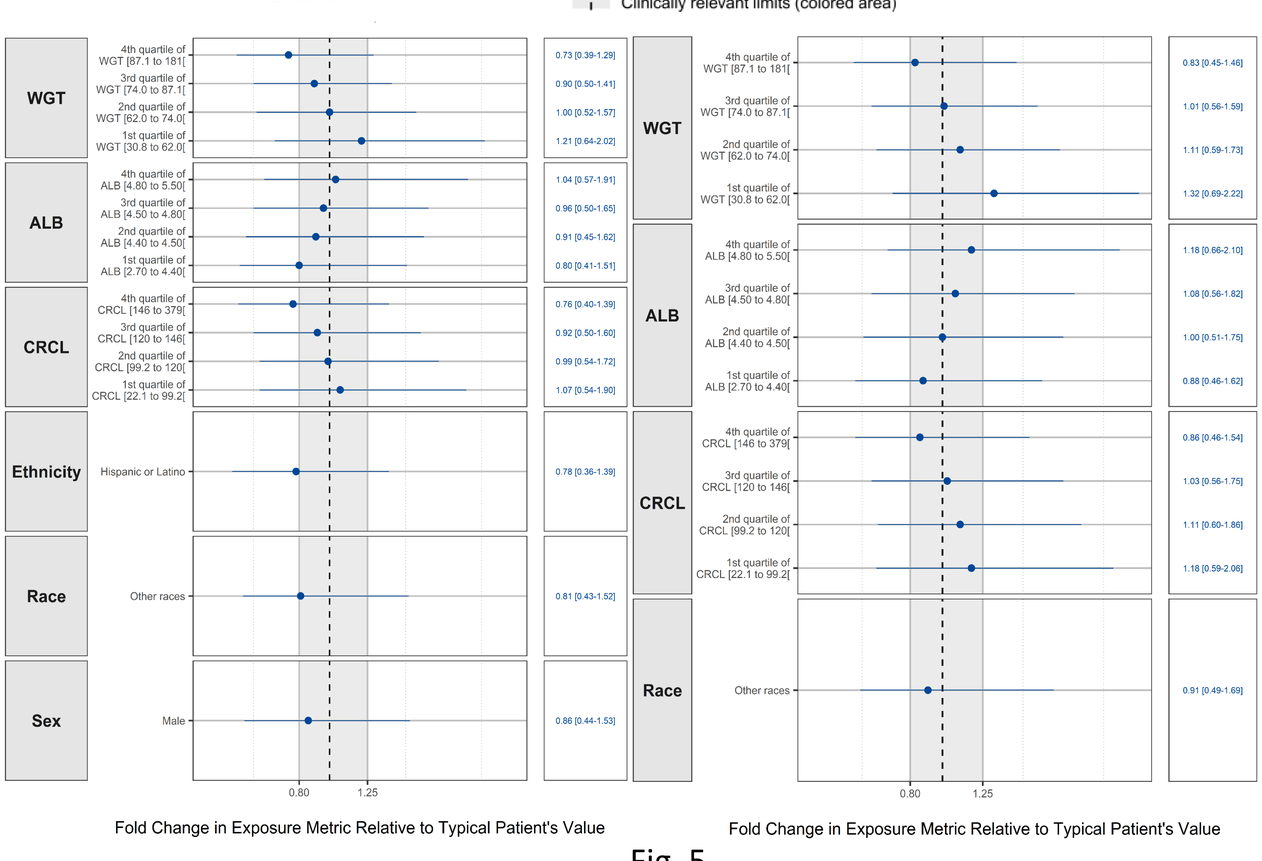

Fig. 5

Conclusion

- Regarding the two approaches, similar conclusions are reached about clinical implications based on covariate.
- The covariate selection process can become efficient and user friendly by using Machine Learning framework algorithms as implemented in the *mlcov* package.

[1] Rebai I., Duval, V., Akil, A., Teusher, N., Largajolli, A. and Fauchet, F. Evaluation of the Boruta Machine Learning Algorithm for Covariate Selection. 31st PAGE meeting 2023, A Coruna, Spain, June 2023
[2] KURSA, Miron B.; JANKOWSKI, Aleksander; RUDNICKI, Witold R. Boruta–a system for feature selection. Fundamenta Informaticae, 2010, 101.4: 271-285
[3] Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. Journal of Statistical Software, 36(11), 1–13.
[4] Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1-22
[5] Jonsson E, Karlsson M (1998) Automated covariate model building with NONMEM. Pharm Res 15(9):1463–1468