| | |
|---|---|
| Name: | Jenna Reps |
| Affiliation: | Johnson & Johnson |
| Email: | jreps@its.jnj.com |
| Presentation type (select one): | Poster |

# Feasibility of using basic heuristics to simplify patient level prediction models

**Jenna Reps, PhD[1], Patrick Ryan, PhD[1]; Peter Rijnbeek, PhD[2]**
**[1]Janssen Research and Development, Titusville, NJ; [2]Erasmus MC, Rotterdam, The Netherlands**

**Abstract** (limit = 200 words; current count = 139 words)

*The PatientLevelPrediction package provides the opportunity to explore the prediction potential of using observational healthcare data across a large number of prediction problems. The package provides a standardized framework implementation that aids the development of complex models containing hundreds or thousands of variables. However, to be clinically used a model generally needs to be simple and contain less than ten variables. In this paper, we investigate three simple heuristic feature selection approaches that can be implemented to reduce the number of variables while determining the effect this has on the discriminative ability of the model. Our results suggest/show that complex models trained via the PatientLevelPrediction package could be simplified slightly by reducing the number of variables without significantly decreasing the discriminative ability. However, more complex feature selection/engineering approaches may be required to reduce the number to less than 10.*

## Introduction

The OHDSI PatientLevelPrediction package[1] enables the development of prediction models that can include thousands of medical concepts recorded into observation healthcare databases as potential predictor variables. The large number of potential predictor variables increases the likelihood of a model overfitting, so the majority of the classifiers contained in the package incorporate some form of regularization. The regularization adds a model complexity cost (e.g., having too many predictor variables) into the loss function that is optimized during training. Lasso logistic regression adds the cost term consisting of the sum of the absolute model coefficient values; this effectively shrinks the coefficients of weakly predictive (or non-predictive) variables to zero. Therefore, lasso logistic regression can be considered to perform variable selection during model training to obtain a parsimonious model with a high predictive power.

In general, even when tens of thousands of predictor variables are initially considered into a prediction model, the lasso regularization causes only a small subset (approximately 100-500 hundred) of variables to be chosen in the final model. A model containing hundreds of variables can be readily implemented using a computer, but cannot be easily implemented by a clinician or patient. Ideally, a model should contain a maximum of around 10 variables to ensure it is simple enough to be used clinically. This prompts the question, can we further simplify the trained models by cutting down the number of variables while not significantly reducing the performance of the model (i.e., the discrimination as measured by the AUC).

## Objective

In this paper we investigate whether three simple heuristics can be implemented to reduce the number of variables in a prediction model without significantly reducing the discriminative ability. The aim is to determine whether model simplification is possible and, if so, gain insight into the best approach to apply. This could help aid the conversion of complex models into clinically useful models.

## Method

In a target population of therapeutically treated depressed adults we predicted the outcomes: stroke and nausea occurring within 1 day of the start of depression treatment and 365 days later. We developed a lasso logistic regression model for which the hyper-parameters were selected using 3-fold cross validation on 75% of the data

used as a train set. The remaining 25% of data was used as the hold out test set for internal validation. We used a database that contained approximately 2 million people in the target population.

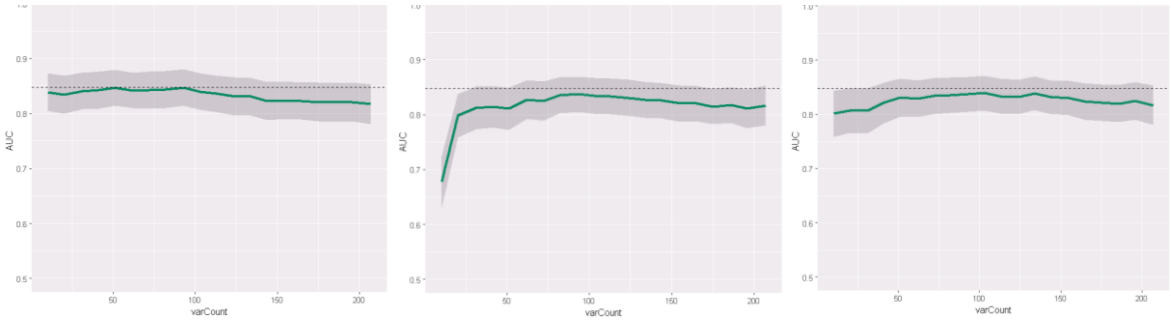
We identified the variables selected by the lasso logistic regression. We then re-ran lasso logistic regression 20 times, each time using the top 100%,95%,…,5% of the lasso selected variables (based on three different criteria) to include in the training set. The three criteria are:

1. Absolute coefficient value from the lasso logistic regression
2. Overall prevalence of predictor in the target population
3. Absolute mean difference between the frequency the variable occurred in those with the outcome minus those without the outcome

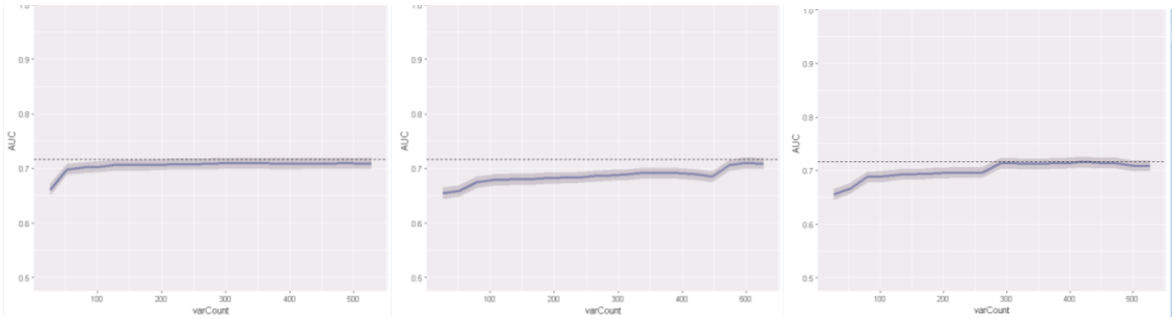
The simplified models were then evaluated on the independent test set using the area under the receiver operating characteristic curve (AUC).

**Results**

For the outcome stroke, out of 41623 variables the complete trained model contained only 207 variables. The complete model obtained an AUC of 0.85 (95% confidence interval: 0.81-0.88) and was well calibrated.
**Figure 1.** The AUC of the stroke model as function of the number of variables of the full model for the three simple heuristics. The left: absolute coefficient value, middle: prevalence in target population and right: mean difference. The dashed line is the AUC on the complete original lasso regression model for stroke.




For the outcome nausea**,** out of 41623 variables the complete trained model contained only 527 variables. The complete model obtained an AUC of 0.72 (95% confidence interval: 0.71-0.73) and was well calibrated.
**Figure 2.** The AUC of the nausea model as function of the number of variables of the full model for the three simple heuristics. The left: absolute coefficient value, middle: prevalence in target population and right: mean difference. The dashed line is the AUC on the complete original lasso regression model for nausea.




**Discussion**

The results show that the simple post training feature selection heuristics based on coefficient value or mean difference can be implemented to simplify a PatientLevelPrediction model to some degree without reducing the discriminative ability significantly but none of the three approaches was able to generate simple models with less than ten variables while maintaining the same (or similar) discriminative ability as the complete model.

Interestingly, the regularized model trained with just the selected variables did worse than when trained on thousands of variables. This is probably due to the hyper-parameter for regularization being selected to have more weighting on the cost of model complexity when there were thousands of variables and therefore reducing overfitting. Interestingly, it seems stroke can be discriminated from non-stroke with fewer variables than discriminating nausea from non-nausea, this may be due to nausea have numerous causes.

## Conclusion

In this paper we investigated three simple heuristics that aim to simplify a model. The results suggest it may be possible to simplify PatientLevelPrediction models but more target populations and outcomes need to be investigated, and additional feature selection heuristics should be explored.

## References

1. Jenna Reps, Martijn J. Schuemie, Marc A. Suchard, Patrick B. Ryan and Peter R. Rijnbeek (2017). PatientLevelPrediction: Package for patient level prediction using data in the OMOP Common Data Model. R package version 1.2.1.