Creating parsimonious patient-level prediction models using feature selection

Aniek F. Markus[1], Egill A. Fridgeirsson[1], Ross D. Williams[1]
[1] Department of Medical Informatics, Erasmus University Medical Center,
Rotterdam, The Netherlands

Background

Patient-level prediction models can help identify a patient's personalized risk of some future medical event. These models can aid medical decision making. Machine learning models are frequently published, but few have any clinical impact. Their implementation in daily practice is often limited. Often these models are too complex, making it difficult to manually compute predictions or to integrate the model into an electronic health record (EHR) system[1]. Simple and parsimonious patient level prediction (PLP) models (i.e. models with a reduced set of variables) may reduce the complexity of clinically implementing a model. Furthermore, deploying a simple and thus interpretable model can contribute to implementation of trustworthy AI in health care[2]. Interpretability is influenced by amongst others model class and the number of variables. In this study we investigate different data-driven feature selection methods to generate parsimonious models that still perform well.

Methods

We develop patient-level prediction models using different feature selection methods. Feature selection methods can work based on a pre-processing step assessing the importance of features (filter methods), the subset evaluation technique (wrapper methods), or be included in the model algorithm (embedded). In this work, we investigated Normalized Joint Mutual Information Maximization (NJMIM), Iterative Hard Thresholding (IHT) and Least Absolute Shrinkage and Selection Operator (LASSO). NJMIM is a filter method that greedily adds features with maximum information value, we combined this with a logistic regression using an L2 penalty after the filtering to produce a model. IHT is an embedded method that uses a combination of gradient descent and a hard thresholding operator (input parameter specifying the maximum number of covariates) to fit a sparse model in an iterative fashion. LASSO is an embedded method that uses L1 regularization. We use LASSO as baseline because it is commonly used to develop patient-level prediction models.

All prediction models in this study are developed using the PatientLevelPrediction framework[3]. The feature selection methods are evaluated using model performance as measured by the area under the receiver operating characteristic curve (AUROC) in the test set and model complexity defined as the number of covariates in the model. Model performance will be plotted as a function of model complexity to investigate the relationship between model parsimony and prediction performance.

We conducted the study using data from the Integrated Primary Care Information (IPCI) database which consists of Dutch general practice patient data[4]. We aimed to predict dementia in a general population of older adults. The target cohort consisted of patients between 55-84 years of age with a recorded visit between 1 January 2014 – 31 December 2014. We used the earliest recorded visit to a healthcare provider as the index event. The outcome cohort is incident dementia with a time-at-risk of 5 years.

Results

We identified 301,226 patients of which 4,768 were newly diagnosed with dementia within the time-at-risk (outcome rate 1.6%). The results are displayed in Figure 1. All developed models achieved good performance. LASSO achieved the best AUROC performance of 0.868. The IHT method performed with a slight decrease in AUROC (0.005 points), however, it selected only 5 covariates which compares favorably to the LASSO method with 319 covariates. For all models age was the largest coefficient (in absolute value) by far.

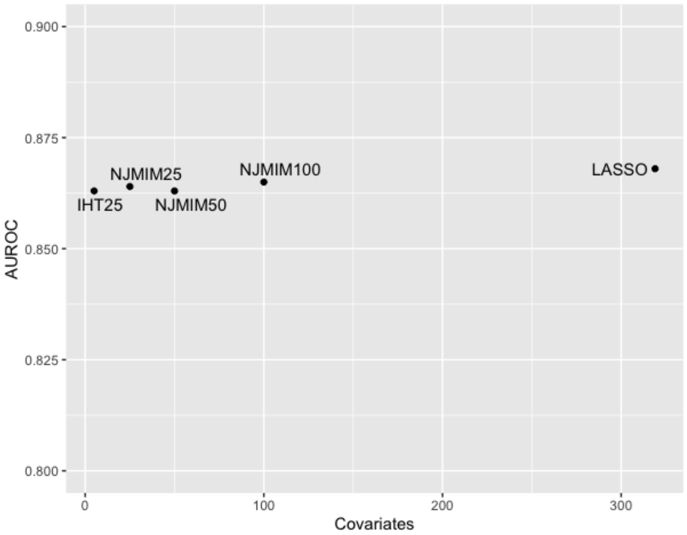

Figure 1. Results for the different feature selection methods and the LASSO model as baseline. IHT# = Iterative Hard Thresholding with threshold of maximum # covariates; NJMIM# = Normalized Joint Mutual Information Maximization with threshold of maximum # covariates.

Conclusion

For the current prediction task, the different data-driven feature selection methods all give similar performance. The most important covariate included in all models was age, which explains why a large reduction in covariates was possible without a drop in predictive performance. This work suggests both IHT and NJMIM can be considered as good methods to create parsimonious clinical prediction models for some prediction tasks. More follow-up research is needed to validate these findings in more prediction tasks and other databases. In future work we also plan to extend this study to more feature selection methods (e.g. component wise gradient boosting, Boruta random forest, maximum relevance minimum redundancy) and perform external validation to see how the selection methods affect the generalizability of the models.

Funding

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union’s Horizon 2020 research and innovation programme and EFPIA.

References

1. Lee, T. C., Shah, N. U., Haack, A., & Baxter, S. L. (2020). Clinical implementation of predictive models embedded within electronic health record systems: a systematic review. In Informatics (Vol. 7, No. 3, p. 25). Multidisciplinary Digital Publishing Institute.
2. Markus AF, Kors JA, Rijnbeek PR. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform. 2021;113:103655.
3. Reps, J. M., Schuemie, M. J., Suchard, M. A., Ryan, P. B., & Rijnbeek, P. R. (2018). Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. Journal of the American Medical Informatics Association, 25(8), 969-975.
4. de Ridder, M. A. J., de Wilde, M., de Ben, C., Leyba, A. R., Mosseveld, B. M. T., Verhamme, K. M. C., . . . , Rijnbeek, P. R. (2022). Data Resource Profile: The Integrated Primary Care Information (IPCI) database, The Netherlands. International Journal of Epidemiology.