An initial investigation into more complex stacking methods to improve transportability of prediction models developed across multiple databases

Cynthia Yang[1], Egill A. Fridgeirsson[1], Jan A. Kors[1], Jenna M. Reps[1,2], Peter R. Rijnbeek[1], Ross D. Williams[1], Jenna Wong[3]

[1]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands, [2]Observational Health Data Analytics, Janssen Research and Development, Titusville, NJ, USA, [3]Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA, USA

Background

Ensemble learning combines multiple models to improve performance compared to a single model. One method of ensemble learning is "stacking", where the predicted probabilities from a set of models (base learners) are used as features to train a meta-learner for the outcome. A previous study investigated the use of stacking to combine models developed across multiple databases to improve model transportability (i.e., to see if a stacking ensemble could perform better in new data than its base learners trained on only a single database). This study showed that stacking ensembles combining L1-regularized (lasso) logistic regression models each trained on a different observational health database (the base learners) could result in better external validation performance than its base learners (1). To train the meta-learner to combine the base learner predictions in a new (external validation) dataset, the stacking ensemble required using some labeled data from the external validation database.

In the previous study, logistic regression was used for the meta-learner, which is most commonly used but imposes some limitations for model transportability: 1) it applies a fixed weight to each base learner across its full range of its predictions, which may be suboptimal if a base learner has poorly calibrated regions (e.g., at the extremes), and 2) it applies the same set of fixed weights for the base learners across all individuals in the new dataset, essentially a "one-size-fits-all" approach which may be suboptimal if different sets of weights are better for different groups of individuals (e.g., by age or sex). In this work, we investigate two incremental enhancements to improve the transportability of stacking ensembles. First, we investigate using random forest as a meta-learner to combine the predictions from the base learners in a manner that imposes neither fixed nor linear weights. Second, we additionally include age and sex (in addition to the base learner predictions) as features in the random forest meta-learner to allow for the potential of different combinations of base learners by baseline patient features (interactions).

An important consideration when using stacking ensembles (especially more complex stacking methods) is the amount of labeled data required to train a meta-learner in the external validation database, and whether this meta-learner performs better than fitting a new model on the labeled data. In this work, we investigate these questions by fitting stacking ensembles using varying amounts of labeled data from the external validation database (for the hypothetical situation in which that would be the amount of data available from a new database) and compare its performance to a custom model developed using the same amount of labeled data from the new database (as the benchmark). We wanted to see at what amount of labeled data stacking offers added value over developing a model using new data alone.

Methods

We developed and validated prediction models using the OHDSI Patient-Level Prediction (PLP) framework (2). We used four large claims databases and one large electronic health record (EHR) database from the United States of America (USA) (Table 1) with all databases mapped to the OMOP CDM. For each database, we investigated 21 different outcomes within a target population of people with pharmaceutically treated depression, as described in the PLP framework paper (2). To reduce computational efforts, we sampled 500,000 patients from the target population cohort for each database that contained more than 500,000 patients in the cohort (as in the previous study (1)). Inclusion criteria (minimum observation time of 365 days prior to index, no prior outcome) were applied to obtain the final study populations.

Table 1. Databases included in the study with data mapped to the OMOP CDM

Full name	Short name	Country	Data type	Population size	Date range
IBM MarketScan® Commercial Claims and Encounters	CCAE	USA	Claims	157m	2000-2021
IBM MarketScan® Multi-State Medicaid	MDCD	USA	Claims	33m	2006-2021
IBM MarketScan® Medicare Supplemental	MDCR	USA	Claims	10m	2000-2021
Optum® De-Identified Clinformatics® Data Mart	Optum Claims	USA	Claims	91m	2000-2021
Optum® De-Identified Electronic Health Record	Optum EHR	USA	EHR	101m	2007-2021

A lasso logistic regression model was developed as a base learner for each database, where we performed 3-fold cross-validation during model development to tune the regularization parameter. We iteratively combined the base learners from four of the databases, which we stacked using a sample of *x* labeled observations from the remaining fifth database to train the meta-learner. To do the stacking, we used the predictions from the base learners in the *x* observations as features to train the meta-learner. We fit the following meta-learners: 1) traditional logistic regression using the base learner predictions as the only features, 2) random forest using the base learner predictions as the only features, and 3) random forest using the base learner predictions plus age and sex of the *x* individuals in the labeled training data as features. Finally, as the benchmark, we used the same sample of *x* observations to fit new custom models for the fifth database in the same way as the base learners were developed in each of the other four databases.

The remaining labeled data in the fifth database were used to evaluate performance of the meta-learners and the custom models, where we evaluated discrimination using the area under the receiver operator characteristic curve (AUC).

Preliminary results

Figures 1 and 2 show the AUC of the meta-learners and the custom models for *x* = 2,000 and *x* = 20,000, respectively. We can see that with a small *x*, a custom model often could not be developed because of too few outcome events. For most external validation tasks, logistic regression as the meta-learner achieved the best performance for both *x* = 2,000 and *x* = 20,000. For *x* = 2,000, the stacking ensemble using either logistic regression or random forest for the meta-learner generally outperformed the custom model. However, for *x* = 20,000, the custom model often outperformed the stacking ensemble using random forest as meta-learner. Adding age and sex of the individuals in the *x* training observations as additional baseline features in the random forest meta-learner showed similar performance as not including them in the meta-learner. For some external validation tasks, *x* = 2,000 and *x* = 20,000 yielded models with similar performance; however, in most cases, using a larger amount of labeled data from the external validation database to train the meta-learner resulted in a higher AUC.

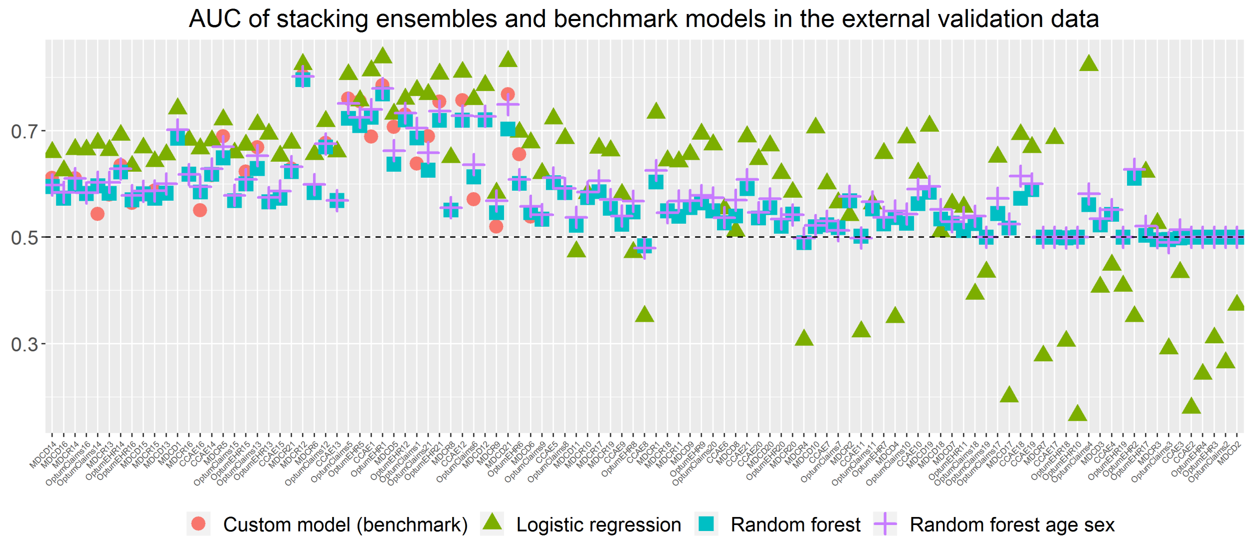

Figure 1. Results when using *x* = 2,000 observations from the external validation database for training, across all outcome and external validation database combinations in order of decreasing observed outcome risk (from left to right).

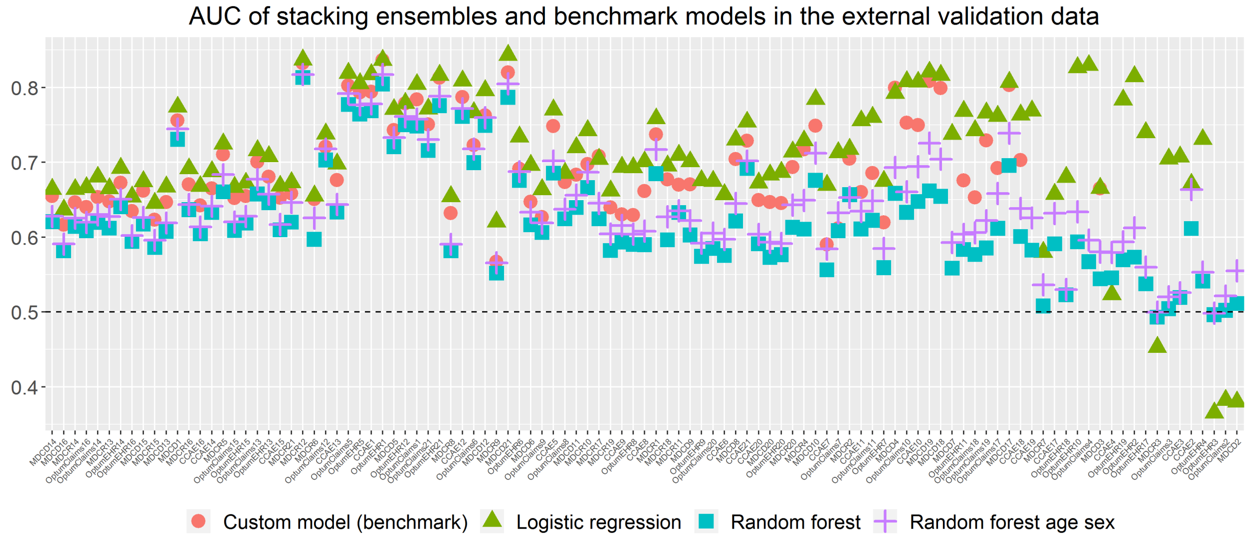

Figure 2. Results when using *x* = 20,000 observations from the external validation database for training, across all outcomes and external validation database combinations in order of decreasing observed outcome risk (from left to right).

Conclusion

In this work, we investigated several approaches for potentially improving the performance of stacking ensembles combining base learners trained on different databases. Our preliminary results do not show that using a random forest meta-learner improves the performance of a stacking ensemble compared to using logistic regression. However, our results do provide insights into the impact of training set size on stacking performance. We find that a small amount of training data can be sufficient to develop a stacking ensemble (that includes base learners with large numbers of features) in a new database, while a small amount of data is not sufficient to develop a new model with similarly large number of features as the base learners in the stacking ensemble. Evaluating more incremental amounts of training data from the external validation database than investigated in this study will allow for a better understanding of the amount of training data up to which developing a stacking ensemble may have added value over developing a custom model for the new data. Further investigation is needed to better understand the reasons for the poorer performance of the random forest meta-learners compared to the traditional logistic regression meta-learners. We are also interested in investigating other flexible but more parametric approaches for developing a meta-learner that better targets the desired associations and interactions we wanted to allow in this study. Finally, future work should evaluate the performance of these stacking approaches in terms of calibration, in addition to discrimination.

Funding

This work has received support from the European Health Data & Evidence Network (EHDEN) project. EHDEN 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. Reps JM, Williams RD, Schuemie MJ, Ryan PB, Rijnbeek PR. Learning patient-level prediction models across multiple healthcare databases: evaluation of ensembles for increasing model transportability. BMC Medical Informatics and Decision Making. 2022;22(1):142.
2. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. 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. 2018;25(8):969-75.