A demonstration of the EnsemblePatientLevelPredition package

Jenna M. Reps, Jenna Wong and Ross Williams

Background

Ensemble learning is the process of combining multiple base models[1] with the aim of improving performance and stability[2]. The independent ensemble framework process generally requires three steps:

1) Learning the base models – this can be done across different model designs (different databases, different outcome phenotypes, different target population definitions or different machine learning models).
2) Filtering the base models – models that do not achieve a sufficient performance (any metric included in PatientLevelPrediction) during internal validation (either test, train or cross validation) are excluded.
3) Combining the remaining base models – often either via fusion[3] or stacking[4].

There are many ways to combine base models via ensemble learning, but in EnsemblePatientLevelPredition, we focus on two of the most common methods: 1) fusing the base models by taking a mean or weighted mean of the base models' predicted risks, and 2) stacking the base models by training a new model that uses the base models' predicted risks as inputs and learns how best to combine them. The simplest ensemble is the uniform weighted fusion that simply takes the mean of all base models' predictions. The other fusion ensembles assign weights to each base model's predicted risks based on one of the PatientLevelPrediction performance metrics. The stacking ensembles require previously unseen labelled data to learn how best to combine the base models' predictions. One suitable approach is to learn the stacker model using a small sample of labelled data from the site or dataset in which you wish to make predictions. It is possible to use different types of classifiers as the stacker model, although logistic regression is often used for its simplicity and ease of interpretation.

In this paper, we present the new EnsemblePatientLevelPrediction R package (https://github.com/OHDSI/EnsemblePatientLevelPrediction) and demonstrate how the package can be used to develop advanced ensembles. In a previous study, it was shown that fusion ensembles combining single-database base models developed across different datasets generally outperformed a simple logistic regression stacking ensemble and the single-database models[5]. However, the base models plus the fusion and stacking ensembles all suffered from poor calibration when externally validated[5]. The previous study did not filter any of the base models (although the stacker model could have given a base model a 0 weight), only investigated logistic regression as the stacker model, and did not attempt to recalibrate the base models' predictions to the site or dataset in which the stacker model was being applied. In this study, we will expand the previous experiment to explore ways to improve the ensemble performances. For fusion we investigate the inclusion of base model filtering and recalibration. For stacking we investigate more flexible supervised learning methods (e.g., decision tree) for combining base model predictions, as well as the inclusion of additional predictors in the stacker model, such as age in years and Charlson comorbidity index, to facilitate recalibration of and allow for interactions with base model predictions.

Methods

Prediction Task: We focus on the task of predicting 1-year risk of hypotension within patients with a first episode of pharmaceutically-treated depression (index is date of depression diagnosis). Inclusion criteria: >=365 days observation in the database prior to index and no prior hypotension.

Data Sources: We use five databases, see Table 1.

Table 1- Summary of databases used in study

Database Name	IBM MarketScan Medicare Supplemental Beneficiaries	IBM MarketScan Medicaid	Optum® De-Identified Clinformatics® Data Mart Database	Optum® de-identified Electronic Health Record Dataset	IMB MarketScan Commercial Claims and Encounters
Abbreviated Name	MDCR	MDCD	OptumClaims	Optum EHR	CCAE
Type of data	US insurance claims	US insurance claims	US insurance claims	US electronic healthcare database	US insurance claims
Observation period	Jan 1, 2000 and Dec 31, 2021	Jan 1, 2006 and April 30, 2021	May 1, 2000 and Dec 31, 2021	Jan 1 2007 and June 30, 2021	Jan 1, 2000 and Dec 31, 2021
Number of covered lives	10.39M	33.36M	91.68M	101.04M	162.22M

The use of IBM Health MarketScan® and Optum databases were reviewed by the New England Institutional Review Board and were determined to be exempt from broad Institutional Review Board approval.

Base Models: In four databases (CCAE, MDCR, MDCD, OptumClaims) we trained four regularized logistic regression models (LASSO) using a 75% train set and 25% test set split and 3-fold cross validation in the train set to select the regularization hyper-parameter (maximizing AUROC). The covariates offered to the base models were age in 5-year bins, gender, race, ethnicity, binary variables indicating the presence of a record of conditions/drugs/measurements/procedures/devices/observations – each in the prior 30 days and prior 365 days, and number of visits in the prior 30 days and prior 365 days.

Ensemble Designs: In this study, we compare five ensemble designs for combining the base models, see Table 2.

Table 2 - The five different ensemble strategies

Design	**Base Models**	**Filter Rule**	**Combination**
Design 1	Models developed across four databases (CCAE, MDCR, MDCD, OptumClaims) with the same model design.	None	Fusion with uniform weights
Design 2	Models developed across four databases (CCAE, MDCR, MDCD, OptumClaims) with the same model design	Remove worse performing model based on test AUROC	Fusion with uniform weights
Design 3	Models developed across four databases (CCAE, MDCR, MDCD, OptumClaims) with the same model design	None (* however the stacker's final model may not use a base model)	Stacker with logistic regression model
Design 4	Models developed across four databases (CCAE, MDCR, MDCD, OptumClaims) with the same model design	None	Stacker with decision tree model

Design 5	Models developed across four databases (CCAE, MDCR, MDCD, OptumClaims) with the same model design	None	Stacker with decision tree model plus age in years and Charlson comorbidity index predictors as additional stacker inputs, in addition to the base model predictions.

Evaluation: We evaluate the external validation performance of the ensembles on the Optum EHR data (a database not used by the base models). The stacker models require extra data to learn how to combine the base models. It makes sense to use a small amount of labelled data from the database or site in which the model will be applied. Therefore, in this study, we use 25% of the Optum EHR data to fit the stacker models (designs 3-5). To make the comparison with the other ensemble designs fair, we recalibrate the fusion models (designs 1-2) using the same 25% of Optum EHR database. We used the weakRecalibration() function in the PatientLevelPrediction package that learns a new logistic regression slope and intercept. In the 75% of the unseen Optum EHR data, we evaluate the discriminative performance using the AUROC metric and the calibration using both the calibration-in-the-large and the E-statistics[6].

Results

Table 3 shows the calibration (mean predicted risk/observed risk and the E-statistics) and discrimination (AUROC with 95% confidence intervals) for the different ensemble designs when applied to the unseen 75% of Optum EHR data. The AUROC was similar and ranged between 0.8205-0.8258 across the five ensemble designs investigated. The tree-based stacker models achieved the lowest E-statistic values and their mean predicted risks were closer to the true observed risk.

Table 3 - discrimination and calibration performance for the ensemble designs when applied to 75% of the Optum EHR data

Design	**AUROC (95% CI)**	**Mean observed risk**	**Mean predicted risk**	**E-statistic**	**E90**
1: fusion uniform weights all four models - recalibrated	0.8251 (0.8195-0.8306)	0.01970187	0.01934277 (before recalibration: 0.05493634)	0.00344895	0.008653025
2: fusion uniform weights top 3 models - recalibrated	0.8205 (0.8149-0.8261)	0.01970187	0.01935836 (before recalibration: 0.0537191)	0.002701462	0.007206211
3. Stacker with all base models combined using logistic regression	**0.8258 (0.8203-0.8313)**	0.01970187	0.01932958	0.008290899	0.02145309
4. Stacker with all base models combined using decision tree algorithm	0.8217 (0.8161-0.8274)	0.01970187	**0.01942972**	0.001044224	**0.001145373**
5. Stacker with all base models plus age and Charlson comorbidity combined using decision tree algorithm	0.8217 (0.8161-0.8274)	0.01970187	0.01937698	**0.0009886911**	**0.001145373**

Conclusion

In this study, we presented the new EnsemblePatientLevelPrediction R package and demonstrated its ability to develop advanced ensemble models. Our experiment compared the discrimination and calibration performances for ensembles designs when applied to make predictions for patients in a new database. The results show that using 25% of the Optum EHR dataset (~100,000 data points) to either recalibrate the fusion models or fit the stacker models resulted in similar discrimination performance. The calibration was good for all five ensembles, as the mean predicted risk was similar to the observed risk and the E-statistics were low. Recalibrating the fusion models fix the calibration issue. However, the tree-based ensembles appear to produce better calibrated models. These results show that if there is a small amount of labelled data available where the model is intended to be applied, then recalibrating or fitting a stacker can improve calibration. In future work we plan to expand this experiment to more prediction tasks and explore the minimum recalibration/stacker fitting data size required to obtain good calibration in new data.

References/Citations

1. Tsoumakas G, Partalas I, Vlahavas I. A taxonomy and short review of ensemble selection. In Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications 2008;1–6.
2. Rokach L. Ensemble-based classifiers. Artif Intell Rev. 2010;33(1–2):1–39.

3. Fumera G, Roli F. Performance analysis and comparison of linear combiners for classifier fusion. In Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR) Springer, Berlin, Heidelberg. 2002:424–432.
4. Wolpert DH. Stacked generalization. Neural Netw. 1992;5(2):241–59.
5. 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 Dec;22(1):1-4.
6. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in medicine. 2019 Sep 20;38(21):4051-65.