Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data

Cynthia Yang[1], Egill A. Fridgeirsson[1], Jan A. Kors[1], Jenna M. Reps[2], Peter R. Rijnbeek[1]
[1]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
[2]Observational Health Data Analytics, Janssen Research and Development, Titusville, NJ, USA

Background

Many datasets used for clinical prediction modeling exhibit an unequal distribution between their outcome classes and are hence imbalanced; typically, only a small proportion of patients in a study population experiences a certain outcome of interest. In the machine learning literature, the term class imbalance problem has been used to describe a situation in which a classifier may not be suitable for imbalanced data. It has been suggested that a prediction model developed using imbalanced data may become biased towards the larger class (also referred to as the majority class) and may be more likely to misclassify the smaller class (also referred to as the minority class) (1). As a result, various methods have been proposed to improve prediction performance when developing prediction models using imbalanced data. Such methods are also referred to as class imbalance methods.

In our previous systematic review on clinical prediction modeling using electronic health record (EHR) data, we found that class imbalance methods were increasingly applied in the period 2009-2019 (2). However, there is currently no consensus on the impact of class imbalance methods on the performance of clinical prediction models. Several previous studies suggest that class imbalance methods may indeed improve performance of clinical prediction models (3, 4). In contrast, a recent study focusing on logistic regression investigated the impact of random oversampling, random undersampling, and Synthetic Minority Oversampling Technique (SMOTE), and suggests that these methods result in miscalibration without improving model discrimination (5). These previous studies focused on datasets with small sample size; the impact of class imbalance methods on the performance of prediction models developed using large observational health data is yet unclear.

The aim of this study is to empirically investigate the impact of random oversampling and random undersampling, two commonly used class imbalance methods, on the internal and external validation performance of prediction models developed using large observational health data.

Methods

In this study, we developed and validated prediction models using the OHDSI Patient-Level Prediction (PLP) framework (6). We used three claims databases from the United States of America (USA) and one EHR database from Germany (listed in Table 1) with data 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 (6). For consistency across the experiments and to reduce computational efforts, we sampled an initial study population of 100,000 patients from each database. Further inclusion criteria (minimum observation time of 365 days prior to index, no prior outcome) were then applied to obtain the final study populations.

The imbalance ratio (IR) is defined as the number of patients who do not experience the outcome (the negative class) divided by the number of patients who do experience the outcome (the positive class).

The original IRs ($IR_{original}$) in the final study populations ranged from 8.6 to 245.3 with a median of 84.0. We randomly sampled towards a target IR: $IR_{target} = \min(IR_{original}, x)$ with $x \in \{20, 10, 2, 1\}$.

We considered three different classifiers: lasso logistic regression implemented using the glmnet R package (7), random forest implemented using the Scikit-learn Python package (8), and XGBoost implemented using the xgboost R package (9). A stratified random subset of 75% of the patients in the final study population was used as a training set and the remaining subset of 25% of the patients was used as a test set. First, 3-fold cross-validation (CV) was performed on the training set for hyperparameter tuning. The sampling strategy was only applied to the training folds; it was not applied to the validation fold to allow for a proper evaluation of the model during CV (10).

Next, the test set was used for internal validation. We evaluated each prediction task using the area under the receiver operating characteristic curve (AUROC). The impact of random sampling was then assessed using the difference from the AUROC of the original data model, calculated as internal AUROC difference = $AUROC_{sampled,\ internal} - AUROC_{original,\ internal}$, with $AUROC_{original,\ internal}$ the AUROC of the original data model for which no sampling strategy was applied on internal validation. The impact of the sampling strategy on model calibration was assessed using plots of the predicted risks against the observed risks (11, 12). When random oversampling or random undersampling is applied, the outcome proportion in the data used to train the classifier is modified, resulting in a mismatch between the predicted and observed risks and thus miscalibration is expected. We investigated whether this miscalibration could be corrected by recalibrating the models towards the original IRs, and we assessed the calibration plots both before and after recalibration (13).

Finally, we externally validated each developed model across the other databases. We evaluated the impact of the sampling strategy on model discrimination using the external AUROC difference = $AUROC_{sampled,\ external} - AUROC_{original,\ external}$, with $AUROC_{original,\ external}$ the AUROC of the original data model for which no sampling strategy was applied on external validation.

Detailed definitions of the inclusion criteria and outcome definitions, including code lists, as well as the analytical source code that were used for the analysis, including example code, are available at https://github.com/mi-erasmusmc/RandomSamplingPrediction.

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

Database full name	Database short name	Country	Data type	Population size	Data range
IBM MarketScan® Commercial Database	CCAE	USA	Claims	157m	2000-2021
IBM MarketScan® Medicare Supplemental Database	MDCR	USA	Claims	10m	2000-2021
IBM MarketScan® Multi-State Medicaid Database	MDCD	USA	Claims	33m	2006-2021
IQVIA Disease Analyser Germany EMR	IQVIA Germany	Germany	EHR	31m	2011-2021

Results

First, we investigated the impact on model discrimination in terms of internal AUROC difference (Figure 1). We can see that although there were some cases with a positive AUROC difference, on average random oversampling and random undersampling did not improve the AUROC. For lasso logistic regression and XGBoost, the impact of random sampling on model discrimination was

relatively small, with a maximum absolute AUROC difference below 0.06. However, for random oversampling with random forest, we observed a larger impact on model discrimination; the AUROC differences had a wider range and performance tended to deteriorate for smaller IRs.

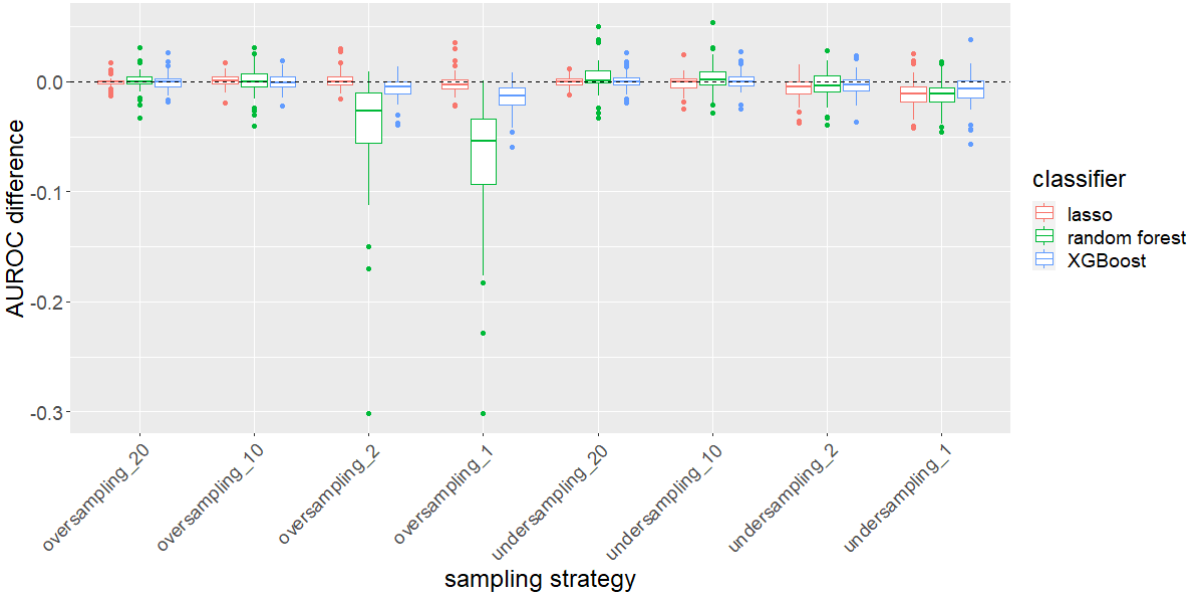

Figure 1. *Internal AUROC differences across all prediction problems and databases for each sampling strategy and classifier. A negative difference means that the original data model had a higher AUROC than the data model based on random sampling.*

Figure 2 shows that model calibration on internal validation clearly deteriorated for all sampling strategies, for all three classifiers. More specifically, the calibration plots indicate increased overestimation for random oversampling or random undersampling towards smaller target IRs, compared to the original data model. This is in line with expectations, since the models with smaller target IRs were trained on datasets with increased outcome proportions. Figure 3 shows that after recalibration, the calibration plots resembled those of the original data models, although for random oversampling with random forest the models appeared to slightly underestimate risks.

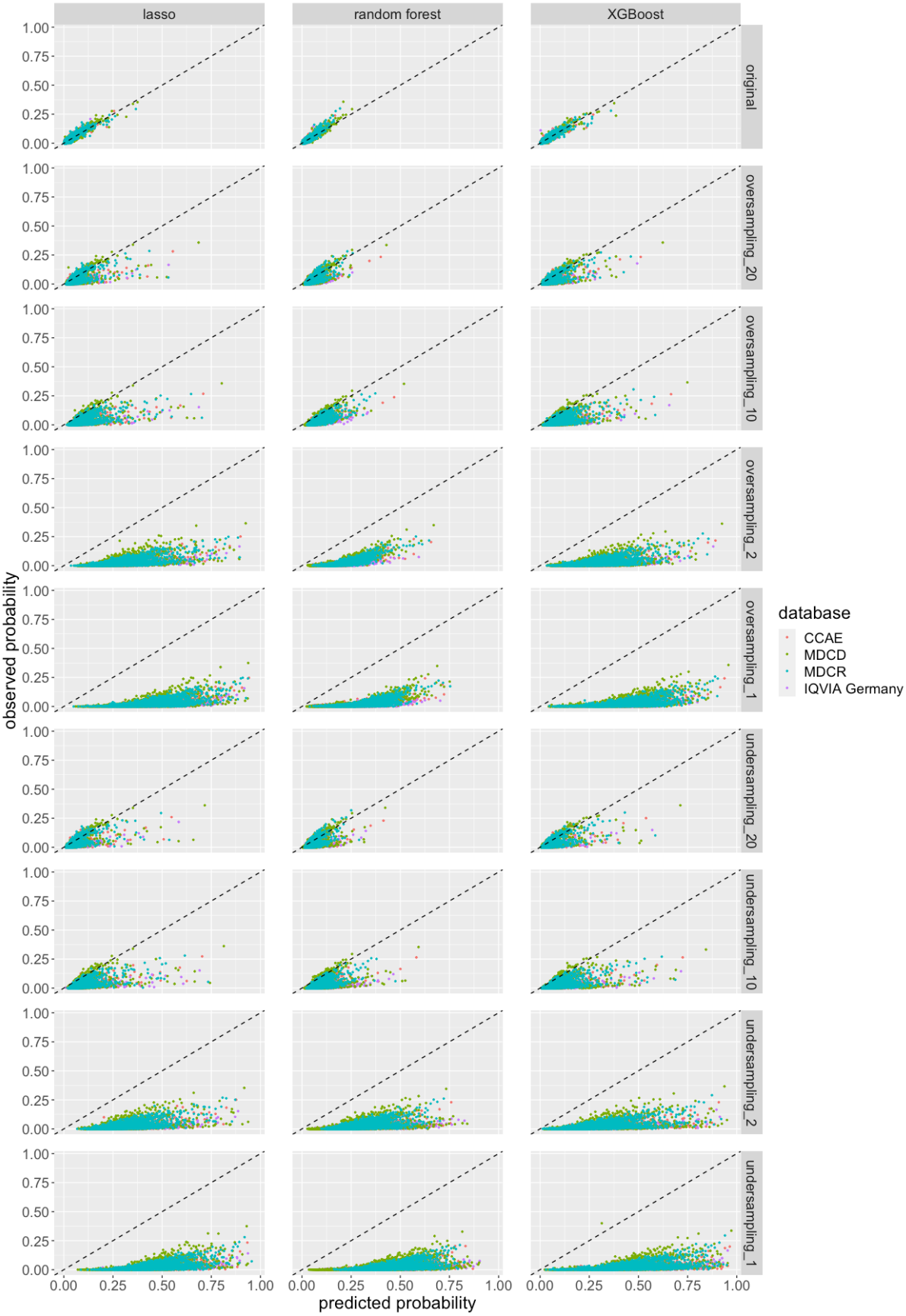

Figure 2. *Calibration plots across all prediction problems and databases for each sampling strategy and classifier on internal validation prior to recalibration towards the original imbalance ratios.*

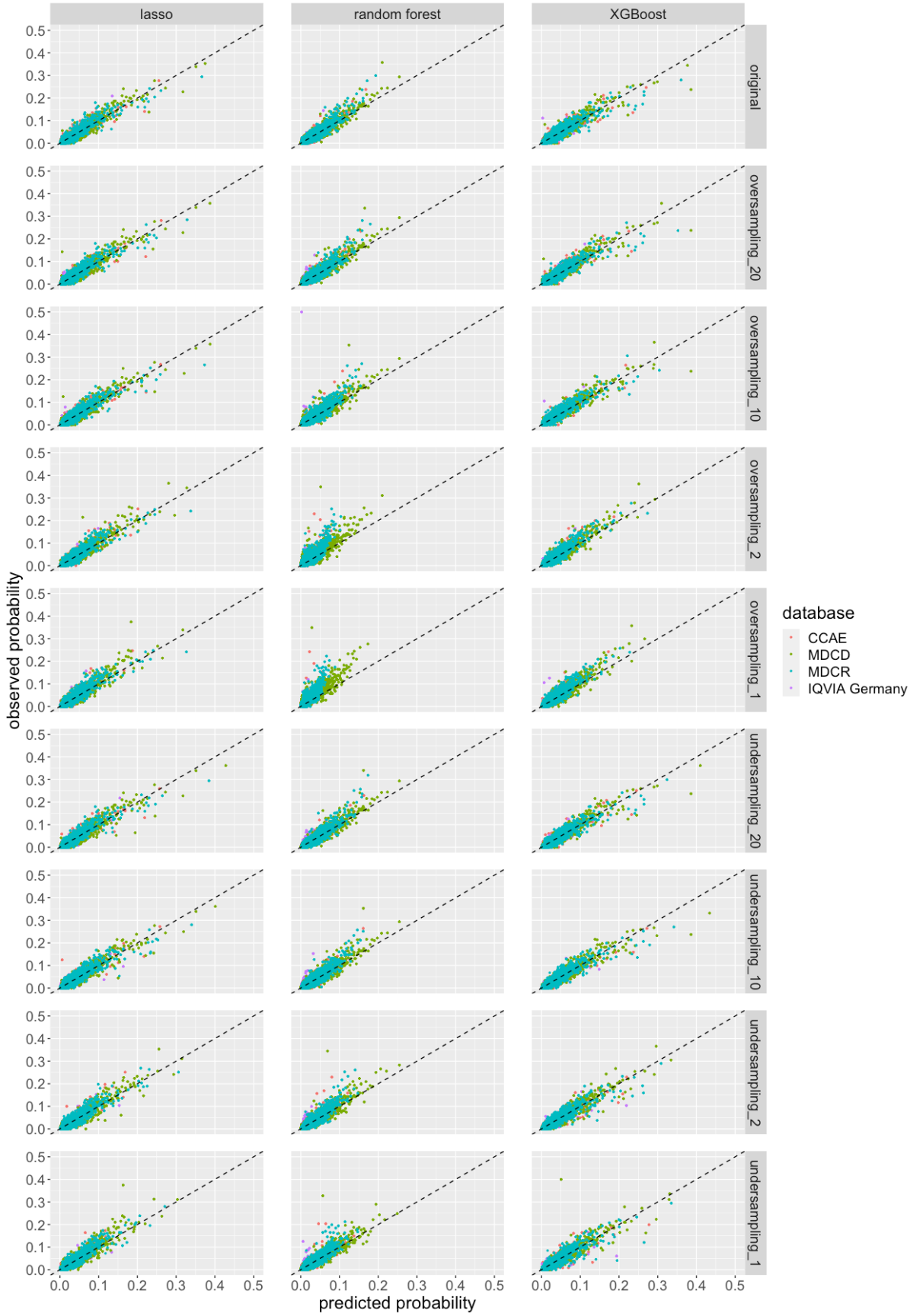

Figure 3. *Calibration plots across all prediction problems and databases for each sampling strategy and classifier on internal validation after recalibration towards the original imbalance ratios.*

Finally, we investigated the impact of random sampling on external validation performance by assessing the external AUROC differences across all prediction tasks for each sampling strategy and classifier (Figure 4). The results were consistent with internal validation; on average, random oversampling and random undersampling did not improve the AUROC on external validation compared to when no sampling strategy was applied. For random oversampling with random forest, the external validation AUROC shows more variation and relatively large drops.

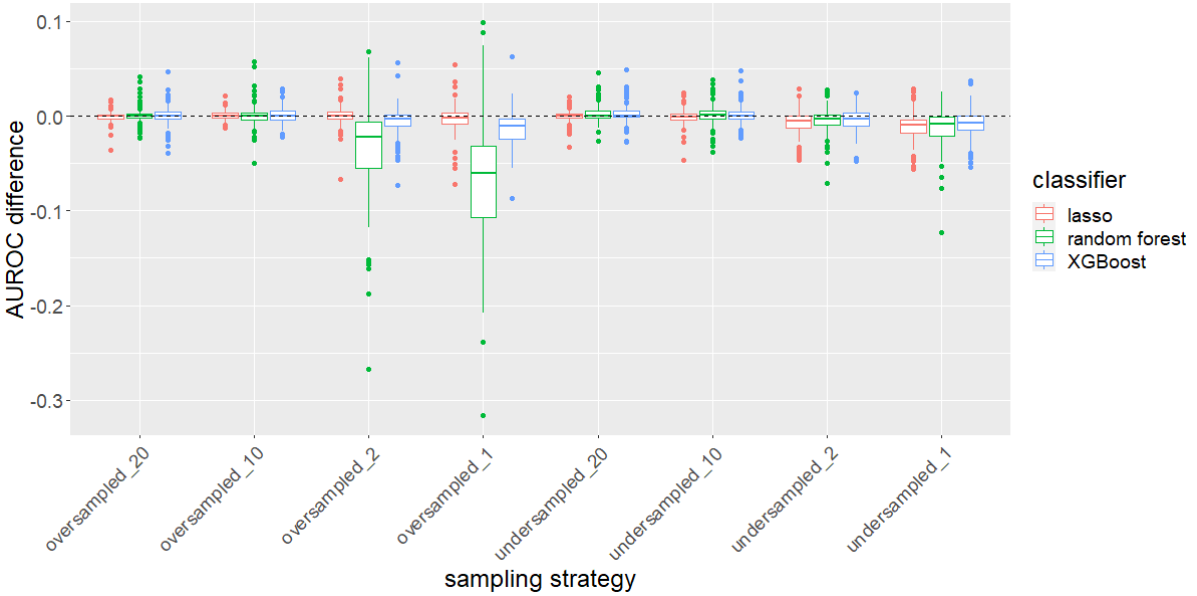

Figure 4. *External AUROC differences across all prediction problems and databases for each sampling strategy. A negative difference means that the original data model had a higher AUROC than the data model based on random sampling.*

Conclusion

In this study, we empirically investigated the impact of random oversampling and random undersampling across various outcomes of interest within a target population of people with pharmaceutically treated depression. Overall, our results suggest that random oversampling and random undersampling on average do not improve the internal and external validation performance. Based on our findings, we do not recommend applying random oversampling or random undersampling when developing prediction models using large observational health data.

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. He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering. 2009;21(9):1263-84.
2. Yang C, Kors JA, Ioannou S, John LH, Markus AF, Rekkas A, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J Am Med Inform Assoc. 2022.
3. Liu J, Wong ZSY, So HY, Tsui KL. Evaluating resampling methods and structured features to improve fall incident report identification by the severity level. J Am Med Inform Assoc. 2021;28(8):1756-64.
4. Fotouhi S, Asadi S, Kattan MW. A comprehensive data level analysis for cancer diagnosis on imbalanced data. J Biomed Inform. 2019;90:103089.
5. Goorbergh Rvd, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. arXiv preprint arXiv:220209101. 2022.
6. 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.
7. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33(1):1 - 22.
8. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.
9. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA: Association for Computing Machinery; 2016. p. 785–94.
10. Blagus R, Lusa L. Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models. BMC Bioinformatics. 2015;16:363.
11. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. BMC Medicine. 2019;17(1):230.
12. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74:167-76.
13. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating: Springer New York; 2008.