


RESEARCH

Open Access



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

Cynthia Yang^{1*} , Egill A. Fridgeirsson¹, Jan A. Kors¹, Jenna M. Reips² and Peter R. Rijnbeek¹

*Correspondence:
c.yang@erasmusmc.nl

¹ Department of Medical Informatics, Erasmus University Medical Center, Dr. Molewaterplein 40, 3015 GD Rotterdam, The Netherlands

² Observational Health Data Analytics, Janssen Research and Development, Titusville, NJ, USA

Abstract

Background: There is currently no consensus on the impact of class imbalance methods on the performance of clinical prediction models. We aimed 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 observational health data.

Methods: We developed and externally validated prediction models for various outcomes of interest within a target population of people with pharmaceutically treated depression across four large observational health databases. We used three different classifiers (lasso logistic regression, random forest, XGBoost) and varied the target imbalance ratio. We evaluated the impact on model performance in terms of discrimination and calibration. Discrimination was assessed using the area under the receiver operating characteristic curve (AUROC) and calibration was assessed using calibration plots.

Results: We developed and externally validated a total of 1,566 prediction models. On internal and external validation, random oversampling and random undersampling generally did not result in higher AUROCs. Moreover, we found overestimated risks, although this miscalibration could largely be corrected by recalibrating the models towards the imbalance ratios in the original dataset.

Conclusions: Overall, we found that random oversampling or random undersampling generally does not improve the internal and external validation performance of prediction models developed in large observational health databases. Based on our findings, we do not recommend applying random oversampling or random undersampling when developing prediction models in large observational health databases.

Keywords: Patient-level prediction, Clinical prediction model, Class Imbalance Problem, Machine learning, External validation, Clinical decision support

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 target 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 [1, 2]. 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 [3]. 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 [4, 5]. In contrast, a recent study focusing on logistic regression investigated random oversampling, random undersampling, and Synthetic Minority Oversampling Technique (SMOTE), and found that balancing data using these methods generally did not improve model discrimination [6]. These previous studies focused on low-dimensional datasets with smaller sample sizes; the impact of class imbalance methods on the performance of prediction models developed in large observational health databases is yet unclear. Observational health data typically contain information on thousands of features concerning health conditions and drugs that are routinely recorded in a patient's medical history.

Additionally, to the best of our knowledge, no previous study investigating the impact of class imbalance methods has also assessed external validation performance. External validation refers to evaluating the model performance on data from databases that were not used during model development, while internal validation refers to evaluating the model performance on data from the same database that was used to train the model such as by using a train and test split-sample. Although good internal validation should be an initial requirement for a prediction model, it is often the case that model performance drops on external validation. We are interested in whether class imbalance methods would result in models with better generalizability and robustness.

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 observational health data. We developed and validated models for various outcomes within a target population of people with pharmaceutically treated depression across four large observational health databases. We used three different classifiers (lasso logistic regression, random forest, XGBoost) and varied the target imbalance ratio.

Methods

In this study, we developed and validated prediction models using the Patient-Level Prediction (PLP) framework from the Observational Health Data Sciences and Informatics (OHDSI) initiative [7]. To improve the interoperability of originally heterogeneous data sources, OHDSI uses the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM), which transforms source data into a common format using a set of common terminologies, vocabularies, and coding schemes [8]. The OHDSI PLP framework in turn allows for standardized development and extensive validation of prediction models across observational health databases that are mapped to the OMOP CDM [8, 9].

Data description

We used four observational health databases: three large claims databases from the United States of America (USA) and one large EHR database from Germany with data mapped to the OMOP CDM. The databases are listed in Table 1 and a description of each database is provided in Additional file 1. Each site obtained institutional review board approval for the study or used de-identified data. Therefore, informed consent was not necessary at any site.

For each database, we investigated 21 different outcomes of interest within a target population of people with pharmaceutically treated depression, as described in the OHDSI PLP framework paper [7]. For each of these 21 different outcomes, the prediction problem was defined as follows: “Amongst a target population of patients with pharmaceutically treated depression, which patients will develop <the outcome> during the 1-year time interval following the start of the depression treatment (the index event)?”. The aim of this study was not to obtain the best possible models for these prediction problems but to empirically investigate the impact of random oversampling and random undersampling on model performance. 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 were then applied to obtain the final study population for each outcome of interest within each database [7]: (1) a minimum of 365 days of observation in the database prior to index, and (2) no record of the specific outcome of interest any time prior to index. Additional file 2: Table S1 provides the observed outcome event count and the observed outcome event proportion in the final study population for all prediction outcomes of interest and all databases. For some

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	Date range
IBM MarketScan® Commercial Claims and Encounters Database	CCAЕ	USA	Claims	157 m	2000–2021
IBM MarketScan® Multi-State Medicaid Database	MDCD	USA	Claims	33 m	2006–2021
IBM MarketScan® Medicare Supplemental Database	MDCR	USA	Claims	10 m	2000–2021
IQVIA Disease Analyser Germany EMR	IQVIA Germany	Germany	EHR	31 m	2011–2021

outcomes in IQVIA Germany, no outcome events were observed. Across all remaining study populations, the observed outcome event count ranged from 32 (0.03%) to 7,365 (10.44%).

Candidate predictors

Candidate predictors were extracted from data routinely recorded in the databases. These included binary indicators of 5-year age groups (0–4, 5–9, etc.) and sex, as well as a large set of binary indicators of recorded OMOP CDM concepts for health conditions and drug groups [10]. For health conditions and drug groups, we considered data from the 365 days prior to index. No feature selection methods were used for selecting candidate predictors prior to model training. The initial study population contained 13,207 candidate predictors in CCAE, 14,237 in MDCD, 13,499 in MDCR, and 6,494 in IQVIA Germany. A list of all the candidate predictors per database is available in Additional file 3.

Handling of missing data

Observational health data rarely reflect whether a feature is not observed or missing. In the observational health data used in this study, if a candidate predictor was not recorded in a patient's history, the candidate predictor defaulted to a value of 0 (corresponding to not observed) for this patient. Age group and sex are required by the OMOP CDM and were always recorded.

Statistical analysis methods

For our experiments, we varied the prediction task, the sampling strategy, and the classifier, resulting in a total of 1,566 prediction models = 58 (prediction tasks) \times 9 (8 sampling strategies + 1 control) \times 3 (classifiers). The details of what was varied are described in the rest of this section.

We refer to a combination of one of the 21 prediction problems and one of the four databases as a prediction task. For each prediction task, a random stratified 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. To increase statistical power for our analysis, prediction tasks for which the test set contained less than 100 outcome events were excluded from further analysis [11]. This resulted in a total of 58 prediction tasks across the four databases; two outcomes ('acute liver injury inpatient' and 'decreased libido') were omitted from further analysis. The imbalance ratio (IR) is defined as the number of patients who do not experience the outcome (the minority class) divided by the number of patients who do experience the outcome (the majority class). An IR = 1 hence represents balanced data, while data with an IR > 100 are typically considered severely imbalanced [12]. The original IRs (IR_{original}) in the final study populations ranged from 8.6 to 245.3 with a median of 84.0 (Table 2).

First, we developed an original data model (without sampling strategy) for each of the 58 prediction tasks. We then investigated random oversampling and random undersampling: for random oversampling, data from the minority class were randomly replicated (with replacement) and added to the original dataset; for random undersampling, data from the majority class were randomly selected and removed from the original dataset.

Table 2 Original imbalance ratios

Outcome of interest	CCAE	MDCD	MDCR	IQVIA Germany
Acute myocardial infarction		221.4	60.5	
Alopecia	191.8	205.8	203.6	
Constipation	43.6	18.6	16.3	85.3
Delirium		245.3	84.5	
Diarrhea	28.9	16.8	16.7	
Fracture	174.3	104.7	34.1	185.8
Gastrointestinal hemorrhage	209.5	83.1	44.6	
Hyponatremia	157.7	85.8	32.4	
Hypotension	131.6	47.6	21.5	178.3
Hypothyroidism	73.1	76.6	31.9	158.9
Insomnia	19.5	12.7	17.9	57.2
Ischemic stroke inpatient			101.1	
Nausea	20.6	8.6	15.7	60.5
Open-angle glaucoma			194.0	
Seizure	180.5	71.4	90.8	
Suicide and ideation	49.7	19.2	164.6	
Tinnitus	158.0	174.8	83.4	152.3
Ventricular arrhythmia and sudden cardiac death inpatient		220.7	90.9	
Vertigo	167.5	202.4	71.9	204.5

Each column represents a database

We randomly sampled towards a target IR: $IR_{\text{target}} = \min(IR_{\text{original}}, x)$ with $x \in \{20, 10, 2, 1\}$; this resulted in a total of eight different sampling strategies.

Three different classifiers were considered: L1-regularized logistic regression (also known as “lasso logistic regression” or “lasso”), random forest, and XGBoost. The algorithms were all implemented within the OHDSI PLP framework [7], with lasso logistic regression implemented using the glmnet R package [13], random forest using the Scikit-learn Python package [14], and XGBoost using the xgboost R package [15]. The model development and internal validation procedure is illustrated in Fig. 1. First, we performed hyperparameter tuning using threefold cross-validation (CV) on the training set [16]. The sampling strategy was only applied to the training folds within CV; it was not applied to the validation fold to allow for a realistic evaluation of the model during CV [17]. Next, the model was refit on the full training set using the tuned hyperparameters, and the final model was internally validated on the test set (i.e., the held out 25% of patients from the development database).

Model evaluation

We evaluated model discrimination for each developed model using the area under the receiver operating characteristic curve (AUROC) with 95% confidence intervals [18]. The impact of the sampling strategy on model discrimination was then assessed using the difference from the original data model AUROC, calculated using internal AUROC difference = $AUROC_{\text{sampled, internal}} - AUROC_{\text{original, internal}}$, with $AUROC_{\text{original, internal}}$ the AUROC of the original data model for which no sampling strategy was applied on internal validation. A positive AUROC difference therefore means that the sampling

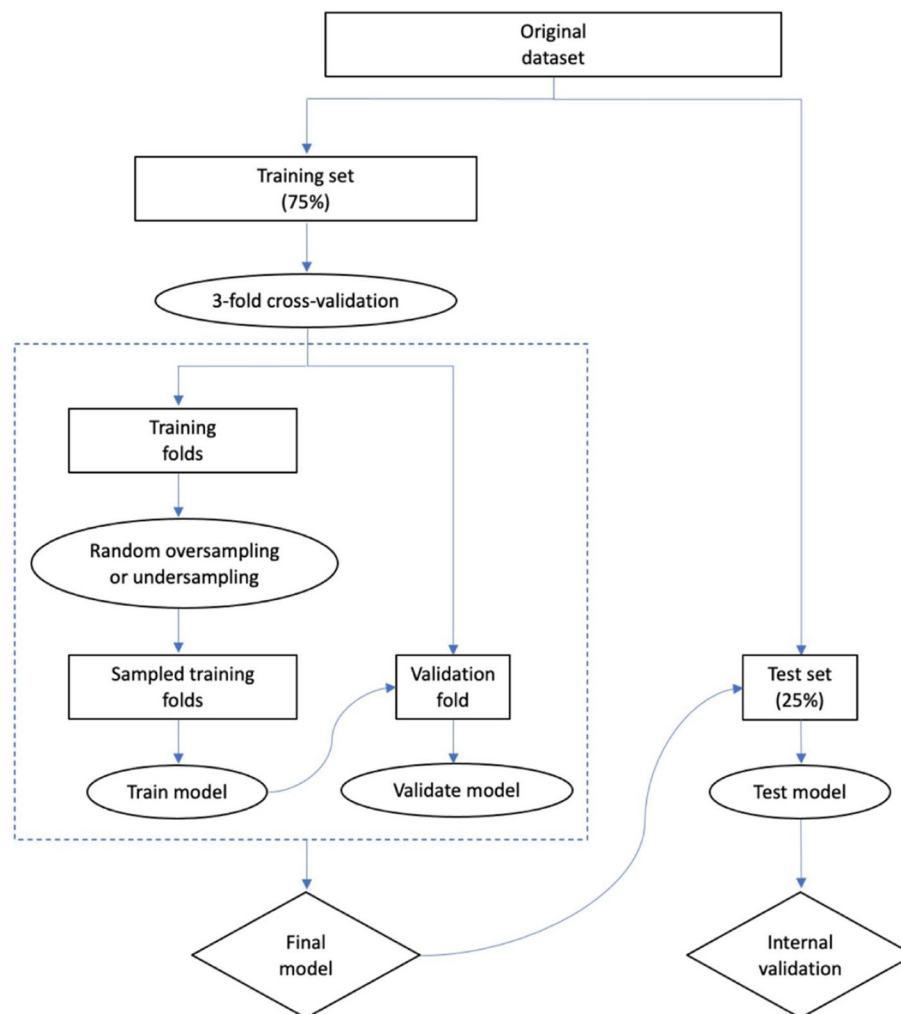


Fig. 1 Flow chart of the model development and internal validation procedure

strategy resulted in an increased AUROC compared to when no sampling strategy was applied, while a negative AUROC difference means that the sampling strategy resulted in a decreased AUROC. We also evaluated discrimination using the maximum F1-score across all prediction thresholds for each model.

The impact of the sampling strategy on model calibration (in the moderate sense) was assessed using plots of the mean predicted risks against the observed outcome event proportions, categorized using percentiles of the predicted risks by each model [19, 20]. Without sampling, the mean predicted risks and the observed outcome event proportions are expected to be equal on internal validation. However, 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 risks and the observed outcome event proportions, 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 [21]. Recalibration towards the original IR was done by adding a correction

to the logit function of the estimated probabilities to obtain the logit function of the recalibrated probabilities as follows:

$$\ln(\text{odds}(\text{probability}_{i,\text{recalibrated}})) = \ln(\text{odds}(\text{probability}_{i,\text{estimated}})) + C \text{ for individual } i$$

$$\text{with } C = \ln\left(\frac{\text{odds}(\pi_{\text{original}})}{\text{odds}(\pi_{\text{target}})}\right) \text{ where } \pi_{\text{original}} = \frac{1}{IR_{\text{original}}+1} \text{ and } \pi_{\text{target}} = \frac{1}{IR_{\text{target}}+1}.$$

We also investigated whether the best sampling strategy in terms of AUROC could be identified prior to evaluating the performance on the test set by selecting the sampling strategy with the highest AUROC during CV. We consider the eight different sampling strategies as well as the option of no sampling strategy, i.e., the original data model, and assessed the internal AUROC difference. We tested whether the median AUROC differences between the original data model and the selected model were significantly different from 0 using the Wilcoxon signed-rank test ($p < 0.05$).

Finally, we investigated whether the sampling strategy resulted in better generalizability and robustness by externally validating each developed model across the other databases [22]. To increase statistical power for our analysis, external validation tasks for which the external validation dataset contained less than 100 outcome events were excluded from further analysis [11]. We evaluated the impact of the sampling strategy on model discrimination using the external AUROC difference = $\text{AUROC}_{\text{sampled, external}} - \text{AUROC}_{\text{original, external}}$ with $\text{AUROC}_{\text{original, external}}$ the AUROC of the original data model for which no sampling strategy was applied on external validation. The impact of the sampling strategy on model calibration was assessed in the same way as on internal 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>.

Results

We developed an original data model for which no sampling strategy was applied and eight different models for which a sampling strategy was applied across a total of 58 prediction tasks and three different classifiers. We hence developed and externally validated a total of 1,566 prediction models. The original AUROCs ranged from 0.58 to 0.87 (Table 3).

First, we investigated the impact on model discrimination in terms of AUROC difference for each sampling strategy and classifier on internal validation (Fig. 2). We can see that although there were some cases with a positive AUROC difference, indicating that the sampling strategy resulted in a higher AUROC compared to when no sampling strategy was applied, random oversampling and random undersampling generally 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 difference in AUROC 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, with the largest difference around -0.3 . Moreover, we investigated the AUROC differences for each sampling strategy and classifier by number of outcome events on

Table 3 Original data model AUROCs (with 95% confidence intervals)

Outcome of interest	Classifier	CCAE	MDCD	MDCR	IQVIA Germany
Acute myocardial infarction	Lasso		0.86 (0.82–0.89)	0.71 (0.69–0.73)	
	Random forest		0.87 (0.84–0.90)	0.69 (0.66–0.71)	
	XGBoost		0.87 (0.85–0.90)	0.71 (0.69–0.73)	
Alopecia	Lasso	0.61 (0.57–0.66)	0.69 (0.65–0.73)	0.69 (0.65–0.73)	
	Random forest	0.58 (0.53–0.63)	0.65 (0.61–0.70)	0.68 (0.64–0.72)	
	XGBoost	0.64 (0.59–0.68)	0.68 (0.64–0.73)	0.68 (0.64–0.72)	
Constipation	Lasso	0.67 (0.64–0.69)	0.65 (0.63–0.66)	0.66 (0.65–0.68)	0.80 (0.78–0.83)
	Random forest	0.66 (0.64–0.69)	0.64 (0.62–0.66)	0.64 (0.63–0.66)	0.81 (0.79–0.83)
	XGBoost	0.67 (0.65–0.69)	0.65 (0.63–0.66)	0.66 (0.65–0.68)	0.80 (0.77–0.83)
Delirium	Lasso		0.79 (0.75–0.84)	0.75 (0.72–0.78)	
	Random forest		0.80 (0.76–0.84)	0.73 (0.70–0.76)	
	XGBoost		0.80 (0.75–0.84)	0.74 (0.71–0.77)	
Diarrhea	Lasso	0.65 (0.63–0.67)	0.67 (0.66–0.69)	0.64 (0.62–0.65)	
	Random forest	0.64 (0.62–0.66)	0.67 (0.65–0.69)	0.62 (0.61–0.64)	
	XGBoost	0.63 (0.61–0.66)	0.67 (0.66–0.69)	0.63 (0.61–0.65)	
Fracture	Lasso	0.61 (0.56–0.66)	0.70 (0.67–0.74)	0.67 (0.65–0.70)	0.82 (0.78–0.86)
	Random forest	0.61 (0.56–0.65)	0.66 (0.63–0.70)	0.65 (0.63–0.67)	0.80 (0.77–0.84)
	XGBoost	0.62 (0.57–0.67)	0.69 (0.65–0.72)	0.67 (0.65–0.69)	0.82 (0.79–0.86)
Gastrointestinal hemorrhage	Lasso	0.73 (0.67–0.78)	0.74 (0.71–0.77)	0.73 (0.71–0.76)	
	Random forest	0.72 (0.67–0.77)	0.75 (0.72–0.78)	0.72 (0.70–0.74)	
	XGBoost	0.70 (0.65–0.75)	0.74 (0.71–0.77)	0.72 (0.70–0.75)	
Hyponatremia	Lasso	0.74 (0.69–0.78)	0.84 (0.81–0.86)	0.66 (0.64–0.68)	
	Random forest	0.73 (0.68–0.77)	0.83 (0.80–0.85)	0.64 (0.62–0.66)	
	XGBoost	0.74 (0.70–0.78)	0.84 (0.81–0.86)	0.66 (0.64–0.68)	
Hypotension	Lasso	0.74 (0.70–0.78)	0.75 (0.73–0.77)	0.72 (0.71–0.74)	0.71 (0.66–0.75)
	Random forest	0.74 (0.70–0.78)	0.74 (0.72–0.77)	0.71 (0.70–0.73)	0.71 (0.67–0.75)
	XGBoost	0.74 (0.71–0.78)	0.75 (0.73–0.78)	0.72 (0.70–0.74)	0.71 (0.67–0.75)
Hypothyroidism	Lasso	0.80 (0.78–0.83)	0.76 (0.72–0.79)	0.83 (0.81–0.85)	0.86 (0.82–0.89)
	Random forest	0.79 (0.76–0.82)	0.74 (0.71–0.78)	0.82 (0.80–0.84)	0.87 (0.84–0.90)
	XGBoost	0.80 (0.77–0.83)	0.75 (0.72–0.78)	0.83 (0.81–0.85)	0.86 (0.82–0.89)
Insomnia	Lasso	0.64 (0.62–0.66)	0.61 (0.60–0.63)	0.67 (0.65–0.69)	0.60 (0.57–0.63)
	Random forest	0.62 (0.61–0.64)	0.60 (0.58–0.61)	0.66 (0.64–0.67)	0.58 (0.55–0.60)
	XGBoost	0.64 (0.62–0.66)	0.61 (0.60–0.63)	0.67 (0.65–0.69)	0.59 (0.56–0.62)
Ischemic stroke inpatient	Lasso			0.79 (0.76–0.82)	
	Random forest			0.76 (0.73–0.79)	
	XGBoost			0.78 (0.75–0.81)	
Nausea	Lasso	0.67 (0.66–0.69)	0.66 (0.65–0.68)	0.66 (0.64–0.68)	0.75 (0.73–0.77)
	Random forest	0.65 (0.64–0.67)	0.65 (0.64–0.66)	0.64 (0.63–0.66)	0.75 (0.72–0.77)
	XGBoost	0.66 (0.65–0.68)	0.66 (0.65–0.67)	0.66 (0.64–0.68)	0.75 (0.73–0.77)
Open-angle glaucoma	Lasso			0.76 (0.71–0.82)	
	Random forest			0.77 (0.72–0.82)	
	XGBoost			0.79 (0.75–0.84)	
Seizure	Lasso	0.75 (0.70–0.79)	0.74 (0.71–0.77)	0.74 (0.70–0.77)	
	Random forest	0.73 (0.69–0.78)	0.71 (0.68–0.74)	0.73 (0.70–0.77)	
	XGBoost	0.72 (0.67–0.76)	0.73 (0.70–0.76)	0.73 (0.69–0.76)	
Suicide and ideation	Lasso	0.79 (0.77–0.81)	0.76 (0.74–0.77)	0.73 (0.69–0.77)	
	Random forest	0.75 (0.73–0.77)	0.72 (0.71–0.74)	0.64 (0.59–0.68)	
	XGBoost	0.79 (0.77–0.81)	0.75 (0.74–0.77)	0.71 (0.67–0.75)	

Table 3 (continued)

Outcome of interest	Classifier	CCAЕ	MDCD	MDCR	IQVIA Germany
Tinnitus	Lasso	0.66 (0.62–0.70)	0.69 (0.64–0.74)	0.60 (0.56–0.63)	0.60 (0.56–0.65)
	Random forest	0.64 (0.60–0.68)	0.71 (0.67–0.76)	0.58 (0.55–0.62)	0.62 (0.58–0.66)
	XGBoost	0.66 (0.62–0.70)	0.69 (0.65–0.74)	0.59 (0.55–0.62)	0.60 (0.55–0.65)
Ventricular arrhythmia and sudden cardiac death inpatient	Lasso		0.83 (0.79–0.87)	0.77 (0.74–0.79)	
	Random forest		0.84 (0.81–0.87)	0.76 (0.73–0.79)	
	XGBoost		0.83 (0.79–0.87)	0.77 (0.74–0.80)	
Vertigo	Lasso	0.65 (0.61–0.70)	0.72 (0.67–0.76)	0.62 (0.59–0.65)	0.63 (0.57–0.68)
	Random forest	0.63 (0.58–0.68)	0.70 (0.66–0.74)	0.59 (0.55–0.62)	0.65 (0.60–0.70)
	XGBoost	0.63 (0.58–0.67)	0.71 (0.66–0.75)	0.60 (0.57–0.64)	0.63 (0.59–0.68)

Each column represents a database

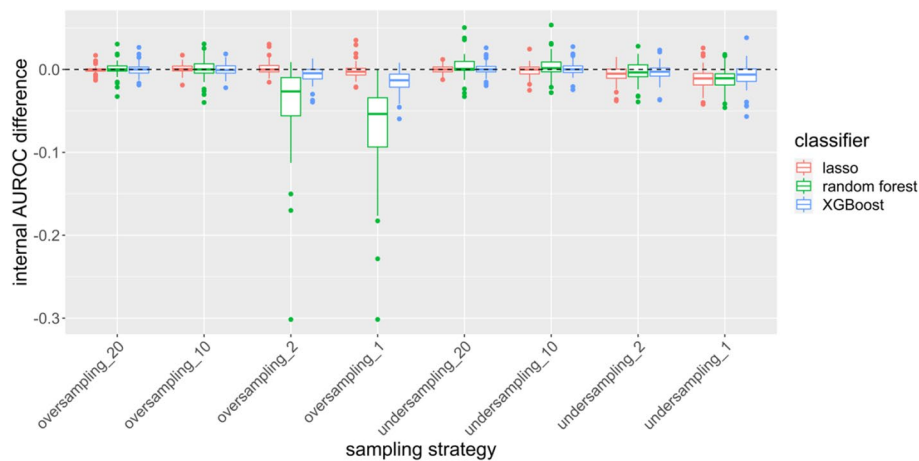


Fig. 2 Internal AUROC differences across all prediction problems and databases for each sampling strategy and classifier. A positive difference means original data model had a lower AUROC, and a negative difference means original data model had a higher AUROC

internal validation (Fig. 3). It appears that overall, and in particular for random oversampling with random forest, the impact of random sampling on the AUROC shows more variation when the number of outcome events is lower. We also investigated the impact on model discrimination in terms of difference in maximum F1-score for each sampling strategy and classifier on internal validation (Additional file 4). We found that random oversampling and random undersampling generally did not improve the maximum F1-score.

Figure 4 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 using increased outcome proportions. To investigate whether this miscalibration could be corrected, we recalibrated the models towards the original IRs. Figure 5 shows that after recalibration, the calibration plots resembled those of the original data models,

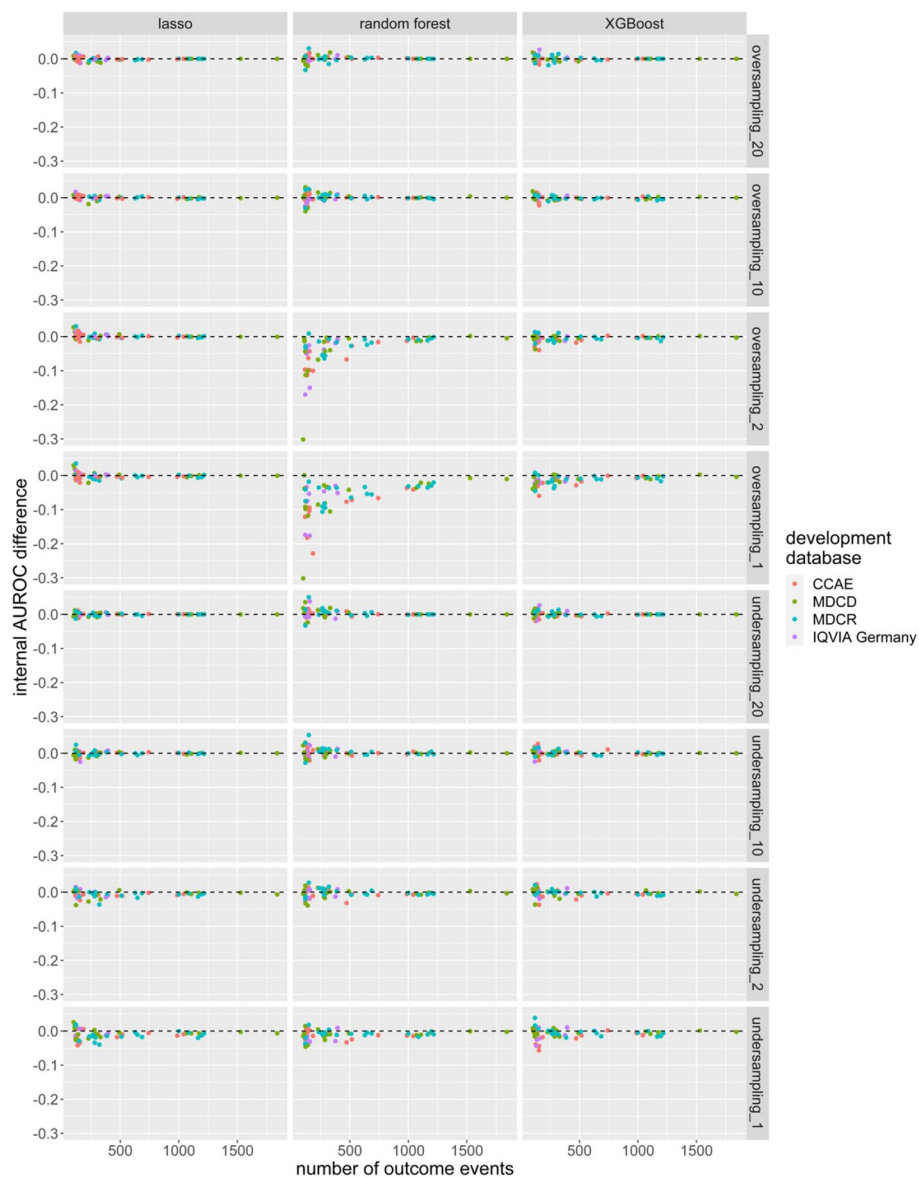


Fig. 3 Internal AUROC differences across all prediction problems and databases for each sampling strategy and classifier by number of outcome events. A positive difference means original data model had a lower AUROC, and a negative difference means original data model had a higher AUROC

although for random oversampling with random forest the models appeared to underestimate risks instead. The same calibrations plots per outcome of interest are available in Additional file 5.

Next, we were interested in whether the best sampling strategy in terms of AUROC could be identified prior to evaluating the performance on the test set by selecting the sampling strategy with the highest AUROC during CV. The option of no sampling strategy, i.e., the original data model, was also considered. Table 4 shows the resulting median AUROC differences across all prediction problems for each database and classifier on internal validation. Only for random forest we found positive median AUROC differences, but these were not significantly different from zero. Hence, selecting the

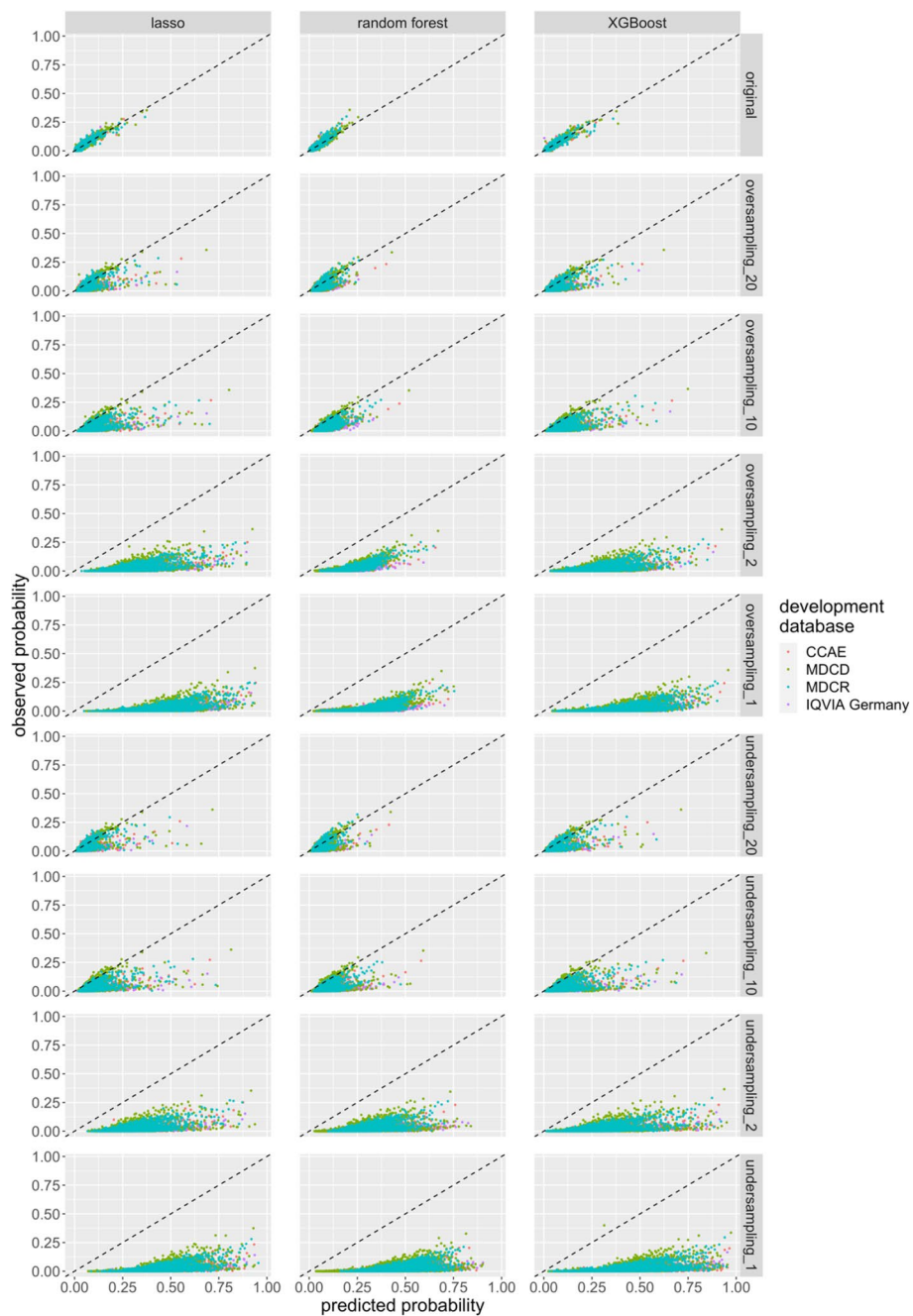


Fig. 4 Calibration plots across all prediction problems and databases for each sampling strategy and classifier on internal validation prior to recalibration towards the original imbalance ratios

sampling strategy based on the highest AUROC during CV generally did not improve the test AUROC.

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 (Fig. 6). The results were consistent with internal validation; generally, random oversampling and random undersampling did not improve the

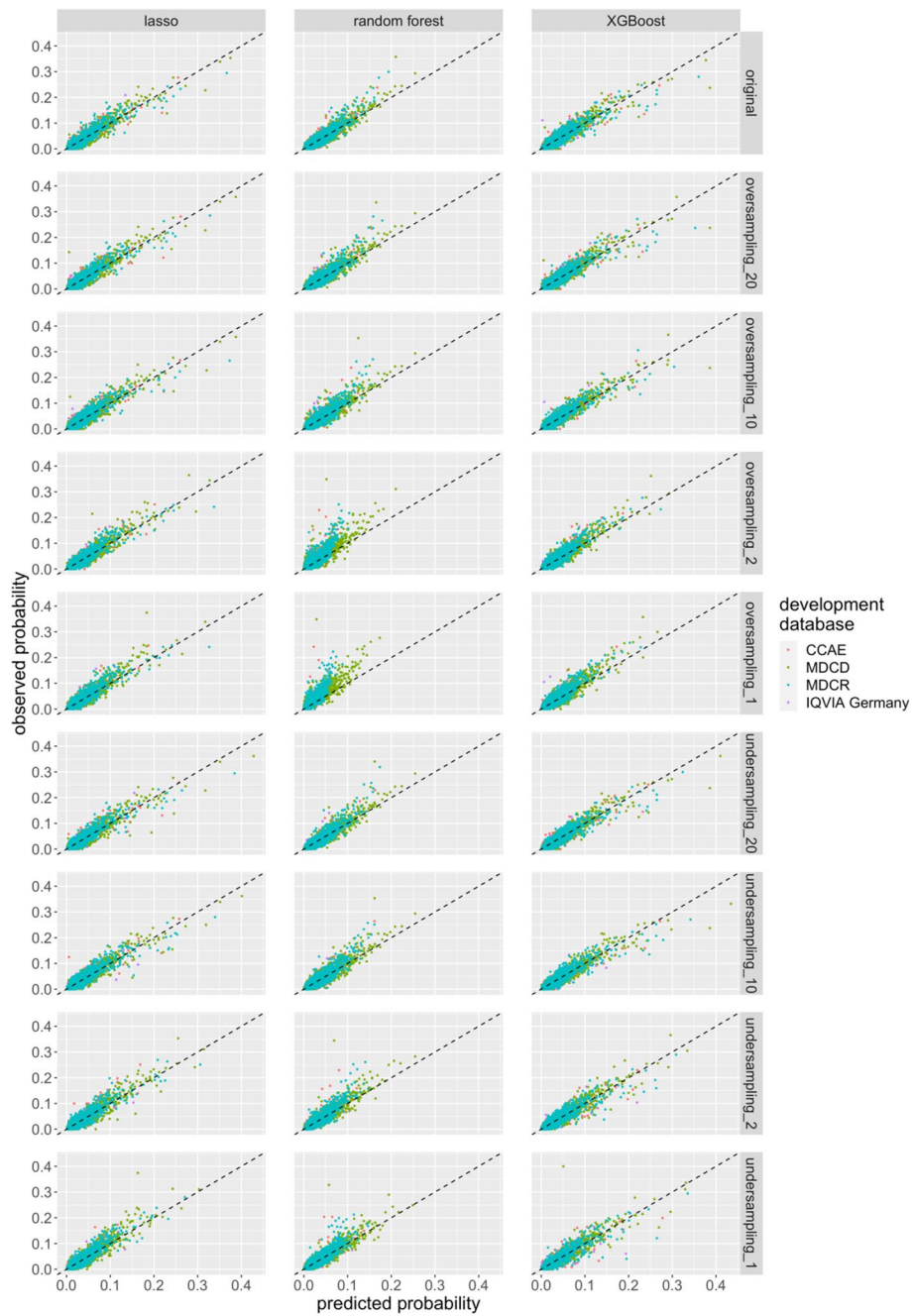


Fig. 5 Calibration plots across all prediction problems and databases for each sampling strategy and classifier on internal validation after recalibration towards the original imbalance ratios

AUROC on external validation compared to when no sampling strategy was applied. For random oversampling with random forest, we found more variation and larger drops in external validation AUROC. We also found that random oversampling and random undersampling generally did not improve the maximum F1-score on external validation compared to when no sampling strategy was applied (Additional file 4). The calibration plots on external validation before and after recalibration are available in Additional

Table 4 Median internal AUROC differences (with interquartile range) across all prediction problems for each database and classifier when choosing the sampling strategy with the highest AUROC during CV

Database	Number of prediction tasks	Lasso	Random forest	XGBoost	All classifiers
CCAE	14	-0.0025 (0.0073)	0.0001 (0.0106)	0 (0.0067)	-0.0004 (0.0076)
MDCD	17	-0.0004 (0.0044)	0 (0.0062)	0 (0.0071)	0 (0.0068)
MDCR	19	0.0000 (0.0052)	0.0037 (0.0143)	0 (0.0057)	0 (0.0075)
IQVIA Germany	8	-0.0011 (0.0048)	0.0012 (0.0095)	-0.0045 (0.0204)	-0.0010 (0.0098)
All databases	58	-0.0004 (0.0053)	0.0008 (0.0099)	0 (0.0074)	0 (0.0081)

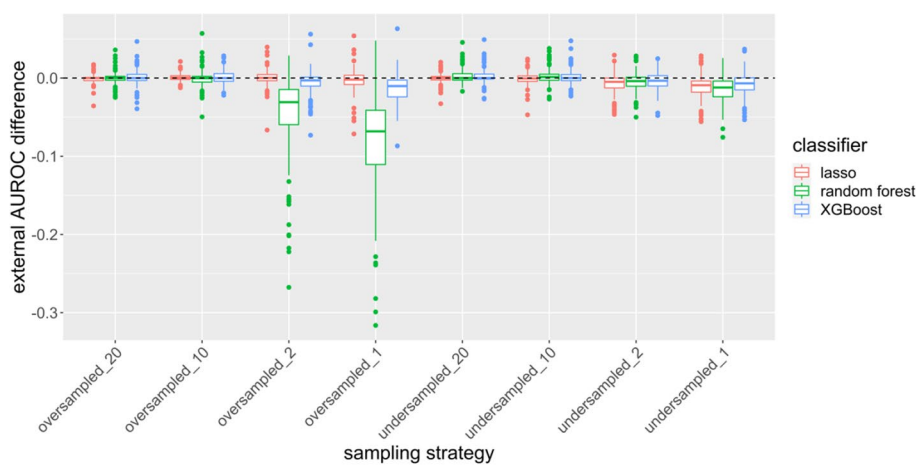


Fig. 6 External AUROC differences across all prediction problems and databases for each sampling strategy and classifier. A positive difference means original data model had a lower AUROC, and a negative difference means original data model had a higher AUROC

file 6. Consistent with internal validation, the calibration plots prior to recalibration indicate increased overestimation for random oversampling or random undersampling towards smaller target IRs. However, after recalibration, the calibration plots mostly resembled those of the original data models.

Discussion

In this study, we empirically investigated the impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. We developed models for various outcomes of interest within a target population of people with pharmaceutically treated depression. We varied the classifier (lasso logistic regression, random forest, XGBoost), the sampling strategy (random oversampling, random undersampling), and the target IR [1, 2, 9, 20] and applied each combination across 58 prediction tasks (each a combination of a prediction problem and one of the four databases). Overall, we found that random oversampling or random undersampling towards different imbalance ratios generally does not improve the performance of prediction models developed in large observational health databases.

On internal validation, the impact of random sampling on model discrimination in terms of an increase or a decrease in the AUROC appeared limited for most models. Only the models for random oversampling with random forest showed more variation in AUROC difference and generally a substantial decrease in the AUROC compared to the original data model. The combination of oversampling with random forest and a target IR of 1 showed the largest drop in test AUROC. The impact on the AUROC appeared to vary more for datasets with a lower number of outcome events. Inspection of the calibration plots allowed us to investigate the impact of random sampling on model calibration. When random oversampling or random undersampling is applied, the outcome proportion in the data used to train the classifier is modified, and we therefore expected miscalibration. In line with our expectations, both random oversampling and random undersampling resulted in overestimated risks. We found that this miscalibration could largely be corrected by recalibrating the models towards the original IRs; after recalibration, the calibration plots resembled those of the original data models, although for random oversampling with random forest the recalibrated models appeared to underestimate risks instead. This highlights that it is important to be aware of the impact of random sampling on model calibration. In our previous systematic review, we found that calibration was often not evaluated at all [3]; we consider it likely that many researchers applying random sampling are not aware of the impact on model calibration and the consequent need for recalibration. For example, several recently published papers on clinical prediction modelling applied random sampling to balance the data used for model development without assessing calibration [23–25].

Most previous studies that investigated the impact of class imbalance methods on the performance of clinical prediction models only evaluated model discrimination using threshold-specific measures such as sensitivity, specificity, and positive predictive value. Thresholds are typically carefully selected within the specific clinical context, which makes it difficult to compare models based on threshold-specific measures. We were interested in investigating the impact of random oversampling and random undersampling on model performance across various outcomes of interest and therefore evaluated model discrimination using the AUROC, which provides a summary measure across all possible thresholds; this makes it difficult for us to directly compare our findings with previous literature. We are not aware of any previous study that has systematically identified a positive impact of random oversampling and random undersampling on the performance of prediction models developed in large observational health databases. One previous study investigated various class imbalance methods using data of cancer patients and suggests that a higher test AUROC could be found amongst these class imbalance methods compared to when no class imbalance method was applied [5]. However, the authors did not consistently identify the same method that would result in a higher AUROC, and it is unclear from this study whether the best class imbalance method could be identified prior to evaluating the performance on the test set. Additionally, calibration was not assessed. We investigated whether the best sampling strategy in terms of AUROC could be identified prior to evaluating the internal validation performance by selecting the sampling strategy with the highest AUROC during CV, and we generally found no improvement in the test AUROC.

Our findings were in line with a recent study focusing on logistic regression that found that completely balancing the data did not result in models with better performance [6]. More specifically, the authors found in a simulation study and a case study that random oversampling, random undersampling, and SMOTE did not improve model discrimination in terms of AUROC. Different from this previous study, our study investigated the impact of random oversampling and random undersampling on model performance for multiple imbalance ratios and multiple classifiers, using large and high-dimensional datasets from multiple observational health databases, and evaluated both internal and external validation. Our findings therefore allow us to extend the findings for random oversampling and random undersampling from this previous study to models developed using lasso logistic regression, random forest and XGBoost in large observational health databases. The authors similarly highlighted the miscalibration resulting from random sampling. SMOTE was proposed for continuous features and our datasets only contained binary features as candidate predictors; we were therefore not able to investigate SMOTE using our data [26].

Finally, we investigated the impact of random oversampling and random undersampling on external validation performance. To the best of our knowledge, no previous study has investigated whether random sampling would result in models with better generalizability and robustness by assessing external validation performance across various databases. We found that consistent with internal validation, on external validation the models for random oversampling with random forest showed more variation in AUROC difference and generally a substantial decrease in the AUROC. Otherwise, the AUROC differences were relatively small. Overall, the results suggest that random oversampling and random undersampling do not result in models with better generalizability and robustness.

A potential limitation of our study is that our results were based on outcomes of interest within a target population of people with pharmaceutically treated depression; we cannot guarantee that these findings will generalize across all prediction problems. Furthermore, a potential limitation of our study is that our results did not account for AUROC uncertainty that may occur due to a low outcome event count in the test set. Nevertheless, to the best of our knowledge, this is the first study that has empirically investigated the impact of random oversampling and random undersampling on the internal and external validation performance of prediction models developed in large observational health databases. By developing and validating models using data mapped to the OMOP CDM, we were able to develop a total of 1,566 prediction models and empirically investigate the impact of random oversampling and random undersampling on internal and external validation performance across four databases. Based on our findings, we do not recommend applying random oversampling or random undersampling when developing prediction models in large observational health databases. Future research could extend our research to other class imbalance methods.

Conclusions

In this study, we empirically investigated the impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. We developed models for various outcomes of interest within a

target population of people with pharmaceutically treated depression across four large observational health databases. Overall, we found that random oversampling or random undersampling towards different imbalance ratios generally does not improve the performance of prediction models developed in large observational health databases. Based on our findings, we do not recommend applying random oversampling or random undersampling when developing prediction models in large observational health databases.

Abbreviations

AUROC	Area Under the Receiver Operating characteristic Curve
CCAE	IBM MarketScan [®] Commercial Claims and Encounters Database
CDM	Common Data Model
CV	Cross-validation
EHR	Electronic Health Record
IQVIA Germany	IQVIA Disease Analyser Germany EMR
IR	Imbalance Ratio
MDCD	IBM MarketScan [®] Multi-State Medicaid Database
MDCR	IBM MarketScan [®] Medicare Supplemental Database
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
PLP	Patient-Level Prediction
SMOTE	Synthetic Minority Oversampling Technique
USA	United States of America

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40537-023-00857-7>.

Additional file 1. Database descriptions.

Additional file 2. Outcome event counts and proportions.

Additional file 3. Candidate predictors per database.

Additional file 4. Differences in maximum F1-score.

Additional file 5. Internal calibration plots per outcome.

Additional file 6. External calibration plots per database.

Acknowledgements

Not applicable.

Author contributions

CY conducted investigation and formal analysis of the study and developed the initial manuscript. EAF, JAK, JMR, and PRR guided conceptualization, investigation, and formal analysis of the study. All authors reviewed, edited, and approved the manuscript before submission.

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.

Availability of data and materials

The data that support the findings of this study are available from IBM and IQVIA, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. The IBM data that support the findings of this study are available from IBM MarketScan Research Databases (contact at: <https://www.ibm.com/products/marketscan-research-databases/databases>). The IQVIA data that support the findings of this study are available from IQVIA (contact at: <https://www.iqvia.com/solutions/real-world-evidence/real-world-data-and-insights>).

Declarations

Ethics approval and consent to participate

All patient data included in this study were deidentified. The New England Institutional Review Board (IRB) determined that studies conducted in these data are exempt from study-specific IRB review, as these studies do not qualify as human subjects research. No experiments were conducted on humans in this study. The research methods were conducted in accordance with appropriate guidelines.

Consent for publication

Not applicable.

Competing interests

JMR is an employee of Janssen Research and Development and shareholder of Johnson and Johnson. CY, EAF, JAK and PRR work for a research group who received unconditional research grants from Janssen Research and Development, none of which relate to the content of this work.

Received: 27 January 2023 Accepted: 14 December 2023

Published online: 03 January 2024

References

1. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng.* 2009;21(9):1263–84.
2. Branco P, Torgo L, Ribeiro RP. A survey of predictive modeling on imbalanced domains. *ACM Comput Surv.* 2016;49(2):Article31.
3. 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;29:983–9.
4. 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.
5. Fotouhi S, Asadi S, Kattan MW. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *J Biomed Inform.* 2019;90:103089.
6. van Goorbergh Rvd 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.
7. 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. *J Am Med Inform Assoc.* 2018;25(8):969–75.
8. Khalid S, Yang C, Blacketer C, Duarte-Salles T, Fernández-Bertolín S, Kim C, et al. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. *Comput Methods Programs Biomed.* 2021;211: 106394.
9. Reps JM, Williams RD, You SC, Falconer T, Minty E, Callahan A, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. *BMC Med Res Methodol.* 2020;20(1):102.
10. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc.* 2012;19(1):54–60.
11. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med.* 2016;35(2):214–26.
12. Leevy JL, Khoshgoftaar TM, Bauder RA, Seliya N. A survey on addressing high-class imbalance in big data. *J Big Data.* 2018;5(1):42.
13. Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010;33(1):1–22.
14. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825–30.
15. 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.*
16. Reps JM, Ryan P, Rijnbeek P. Investigating the impact of development and internal validation design when training prognostic models using a retrospective cohort in big US observational healthcare data. *BMJ Open.* 2021;11(12): e050146.
17. Blagus R, Lusa L. Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models. *BMC Bioinform.* 2015;16:363.
18. Sun X, Xu W. Fast implementation of DeLong's Algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Process Lett.* 2014;21(11):1389–93.
19. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17(1):230.
20. 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.
21. Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating.* New York: Springer, New York; 2008.
22. Williams RD, Reps JM, Kors JA, Ryan PB, Steyerberg E, Verhamme KM, et al. Using iterative pairwise external validation to contextualize prediction model performance: a use case predicting 1-year heart failure risk in patients with diabetes across five data sources. *Drug Saf.* 2022;45(5):563–70.
23. Chiew CJ, Liu N, Wong TH, Sim YE, Abdullah HR. Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission. *Ann Surg.* 2020;272(6):1133–9.
24. Liu L, Ni Y, Zhang N, Nick Pratap J. Mining patient-specific and contextual data with machine learning technologies to predict cancellation of children's Surgery. *Int J Med Inform.* 2019;129:234–41.
25. Makino M, Yoshimoto R, Ono M, Itoko T, Katsuki T, Koseki A, et al. Artificial intelligence predicts the progression of diabetic kidney disease using big data machine learning. *Sci Rep.* 2019;9(1):11862.
26. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* 2002;16:321–57.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.