Research Master's Programme Methodology and Statistics for the Behavioral, Biomedical and Social Sciences

Utrecht University, the Netherlands

Master Thesis Ruben van den Goorbergh (3870995)

TITLE:

The harm of SMOTE and other resampling techniques in clinical prediction models: a simulation study

May 2021

Supervisors:

Maarten van Smeden (Julius centrum Utrecht), Ben Van Calster (KU Leuven)

Preferred journal of publication: Statistics in Medicine

Word count: 4716

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

The harm of SMOTE and other resampling techniques in clinical prediction models: a simulation study

Ruben van den Goorbergh

Correspondence
Ruben van den Goorbergh, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht,Utrecht University, P.O. Box 85500, 3508GA Utrecht, The Netherlands. Email: r.vandengoorbergh@gmail.com

Summary

Methods to adjust for outcome imbalance, i.e. imbalance between the number of events and non-events, such as synthetic minority oversampling (SMOTE) are receiving increasing interest in the field of clinical prediction modelling. In this paper, the effect of imbalance correction methods in the form of resampling techniques on the performance of (ridge) logistic regression and random forest is examined, before and after re-calibration. The effect of resampling techniques is illustrated in the context of ovarian tumours. Then, Monte Carlo simulations are presented to evaluate the discriminative performance as well as the calibration of the different models with respect to different resampling techniques adjusting data imbalance. The results show that the calibration of the models is substantially affected by resampling techniques. The simulations further reveal that logistic regression models do not benefit from resampling techniques. The results suggest that Random forest may benefit from imbalance adjustments in terms of a slightly improved c-statistic, especially when the prevalence is very low. The results further show that random undersampling can lead to various problems that are associated with small data sets. Conclusively, clinical prediction modellers should refrain from applying data adjustment methods to imbalanced data, especially when the calibration of the model is of interest.

KEYWORDS:
Clinical prediction models, Logistic regression, Random forest, Calibration, Simulation

1 | INTRODUCTION

In the field of clinical prediction modelling, the prevalence of the event of interest is typically low, often leading to a development data set in which the number of non-events is far greater than the number of events. In the machine learning literature, this situation is referred to as class imbalance or imbalanced data.[1] The traditional approach for developing dichotomous clinical risk prediction models involves the use of logistic regression models,[2] which, in general, do not suffer from unevenly distributed classes.[3] However, imbalanced data may lead to problems for many other machine learning algorithms because of their accuracy-oriented design, resulting in overlooking the minority class.[4] Therefore, solutions to imbalanced data could be of great interest for clinical prediction modelers, as the use of complex machine learning models is growing.[2]

One category of solutions for dealing with imbalanced data that is advocated in the literature,[5,6,7] is to pre-process the data on which the clinical prediction model is developed using resampling techniques. These result in an artificial, more balanced

development data set, aiming to improve predictive accuracy of the algorithms fit to the more balanced data.[1,4] These techniques involve either undersampling, where a part of the cases belonging to the larger class (majority class) is discarded, or oversampling, where the size of the smaller class (minority class) is artificially increased.

In the machine learning literature, where artificially "correcting" imbalances in the data seems common practice,[1,4] performance of the models is often thought of in terms of classification accuracy. That is, models are assessed on their ability to classify patients, usually to one of two classes. However, in a medical setting, it is often of paramount importance that the clinical prediction model is not only able to accurately make a distinction between events and non-events, but also able to accurately estimate the risk of event (i.e. the calibration of the model is important). This means that there should be an overall agreement between the observed outcomes and risk estimates.[8] Combining the notion of calibration with the fact that, in medicine, the diseases and other health outcomes we want to predict often lead to imbalanced data sets, raises the question how adjustments made for class imbalance affect the performance of the models used for making these predictions from a perspective of the estimation of individual risks.

In this study, I will investigate the performance of two regression-based models (regular- and ridge logistic regression) and a tree-based model (random forest) for dichotomous risk prediction on imbalanced data. The study focuses on the discriminative performance, the risk calibration and predictive accuracy of these models when the imbalance is either adjusted for by one of the following resampling techniques: random undersampling (RUS), random oversampling (ROS) or Synthetic Minority Oversampling Technique (SMOTE).[9,10]

This article is structured as follows. In the next section, the class imbalance approaches, used models and performance metrics are described. Then, in section three, a case study is presented, illustrating the potential harm of resampling techniques. Section four and five describe a simulation study and in the last section, I provide a discussion of the results.

2 | MODELS AND IMBALANCE SOLUTIONS

2.1 | Solutions imbalanced data

In random oversampling (ROS), the size of the minority class is increased by resampling cases from the minority class of the original imbalanced data set, with replacement, until the minority class has the same size as the majority class. This results in an artificially balanced data set containing duplicate cases for the minority class. This process is illustrated for a two predictor situation by the plot in the lower left panel of Figure 1. In random undersampling (RUS), balance is achieved by reducing the size of the majority class to the size of the minority class by randomly discarding cases from the majority class, illustrated by the upper right panel of Figure 1.

Synthetic minority oversampling technique (SMOTE) is a form of oversampling that creates new, synthetic cases. Contrary to ROS, where the minority cases are simply duplicated and thus result in a data set with cases that are identical, SMOTE results in synthetic data points that are interpolations of the original minority class cases.[9,10] The procedure is as follows: for every minority class case, the k nearest minority class neighbours in the predictor space are determined, based on the Euclidean distance.[11,12] Then, the differences between the feature vector of the minority case and those of its k nearest neighbours are taken. These differences are then multiplied by a random number between 0 and 1 and added to the feature vector of the minority case. From the illustration in the lower right panel of Figure 1, it can be seen that these synthetic cases lie on the 'line' between two original minority cases. By creating synthetic data in this manner, there is more variation in the minority cases and hence, the models trained on this data set should be less prone to overfitting than when trained on ROS data.[9]

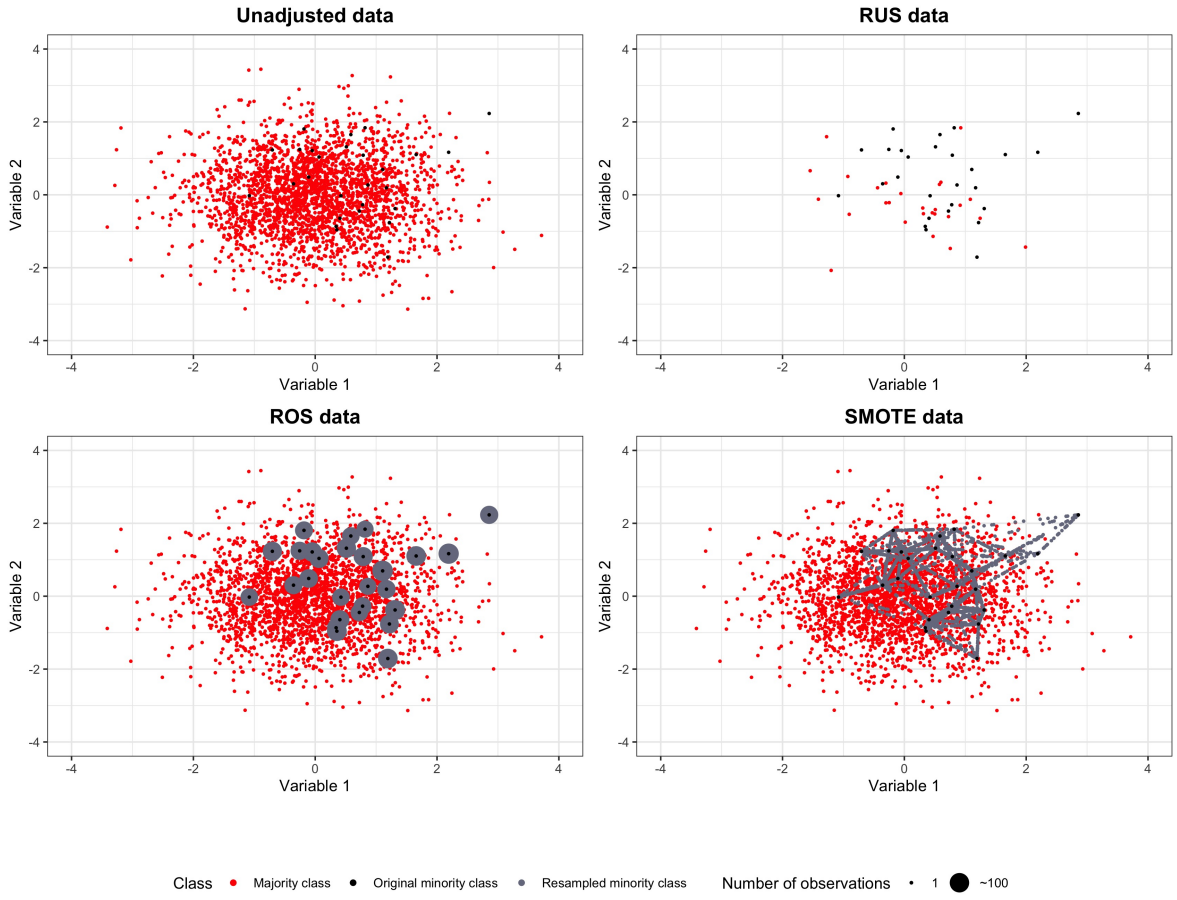

FIGURE 1 Imbalance adjustments. Illustration of the application of different imbalance adjustments on the same data set with an imbalance ratio of 1:100. ROS = Random Oversampling, RUS = Random Undersampling, SMOTE = Synthetic Minority Oversampling Technique

2.1.1 | Models

In logistic regression, the probability ($\pi_i(\mathbf{x}_i)$) of a positive outcome for person $i(y_i = 1)$ is modelled by a linear combination of R predictors. We define π_i as $P(Y = 1|\mathbf{x_i})$, with $i = 1, ..., n$ and $\mathbf{x_i} = (1, x_{1,i}, ..., x_{R,i})$. In the context of this paper, $\mathbf{x_i}$ can either be part of a balanced, or unbalanced data set. When assuming no interaction effects, the logistic regression has the following form

$$log\left(\frac{\pi_i}{1-\pi_1}\right) = \alpha + \sum_{j=i}^{R} \beta_j x_{ij} = \mathbf{x_i}\boldsymbol{\beta}$$

where $\pi_i = \frac{e^{\mathbf{x_i}\boldsymbol{\beta}}}{1+e^{\mathbf{x_i}\boldsymbol{\beta}}}$ and $\boldsymbol{\beta}$ is a column vector containing intercept α and the coefficients β_j. The regression coefficients are estimated by maximizing the log-likelihood function of the following form

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i log(\pi_i(\boldsymbol{\beta})) + (1 - y_i) log(1 - \pi_i(\boldsymbol{\beta}))\}.$$

In ridge logistic regression,[13,14] the following penalized version of the log likelihood function is used for estimating $\boldsymbol{\beta}$, tending to shrink the coefficients towards zero:

$$l(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{R} \beta_j^2.$$

The hyperparameter λ needs to be tuned. In this article, λ is estimated by minimizing the deviance using 10-fold cross-validation with a grid of 251 possible values for λ ranging from 0 (no shrinkage) to 64 (large shrinkage).

A random forest classifier[15] consists of a combination of K decision trees. Each individual tree is trained on a bootstrapped sample of the development data set, using a subset of the predictors to reduce correlation between the different trees. To get to a dichotomous classification, the predictions of all trees are combined by means of a majority vote. To get to a probability estimate, the fraction of trees that voted positive given a particular combination of predictor values is taken. To fit the model, the `randomForest`[16] package with default hyperparameters was used.

2.2 | Predictive performance measures

Models were assessed in terms of calibration (agreement between the observed outcome and the risk estimates), discrimination (ability to distinguish events from non-events) and predictive accuracy (agreement between predicted and observed classes). Calibration was quantified using the calibration intercept, also known as calibration-in-the-large (CIL), and the calibration slope. Discrimination was assessed by means of the c-statistic, also known as area under the receiver operating characteristic curve (AUC). Regarding predictive accuracy, sensitivity, specificity and accuracy were used as performance metrics. An overview of the performance measures and their interpretation can be found in Table 1.

TABLE 1 Performance measures

Performance measure	Interpretation
Accuracy $\left(\frac{TP+TN}{TP+FP+TN+FN}\right)$	Fraction correctly classified cases
Sensitivity $\left(\frac{TP}{TP+FN}\right)$	Fraction correctly classified positive cases over all positive cases
Specificity $\left(\frac{TN}{FP+TN}\right)$	Fraction correctly classified negative cases over all negative cases
c-statistic	c-statistic = 1: perfect discrimination c-statistic = 0.5: no discriminative ability
Calibration slope	Calibration slope < 1: overfitting Calibration slope > 1: underfitting
Calibration intercept	Calibration intercept = 0: mean calibration is perfect Calibration intercept < 0: probabilities are systematically too high Calibration intercept > 0: probabilities are systematically too low
Calibration curve	The further the curve from the diagonal, the worse the calibrative performance

Abbreviations: TP = True postive, TN = True negative, FP = False positive, FN = False negative

3 | CASE STUDY

To illustrate the effect of imbalance solutions on the performance of clinical prediction models, I present a case study applying all described models and imbalance solutions to a subset of the data from a clinical study of ovarian tumours of size $N = 3488$ with no missing data.[17] The aim is to develop a clinical prediction model to predict whether a tumor is malignant (n = 703, 20.2%) or benign (n = 2785, 79.8%). For illustration purposes, we only consider 3 predictors: the age of the patient, and the diameter of the ovary at two measurement points.

To investigate performance of all models in combination with the different imbalance solutions, the data was first split up into a validation and a development set using a 1:4 ratio. This yielded a development data set of $n = 2790$ with 557 events, and a validation data set of $n = 698$. Then, the development set was pre-processed using either ROS, RUS or SMOTE, resulting

in four different development sets: $D_{\text{unadjusted}}$, D_{ROS}, D_{ROS} and D_{SMOTE}. Subsequently, prediction models were developed using maximum likelihood logistic regression, Ridge logistic regression and random forest, resulting in 12 (4 x 3) different models. Ultimately, the models were assessed on their out of sample performance in the validation set, using various common measures of predictive performance (Table 1, Figure 2). For classification, the conventional risk threshold of 0.5 was used.

One may notice that all resampling methods cause the estimated probabilities to exceed the observed event fraction over the whole scale (Figure 2). That is, the probabilities estimated by the prediction models trained on data sets of which the class imbalance was adjusted, were systematically too high, resulting in an overestimation of individual patient risks. The c-statistic of the regression models remained the same over all resampling techniques. For the random forest models, the models developed on oversampled data resulted in slightly lower c-statistics than the models train on unadjusted or RUS data (Table 2). Additionally, all imbalance solutions affect the performance in terms of predictive accuracy: the overall classification accuracy decreases, but the sensitivity and specificity are more balanced.

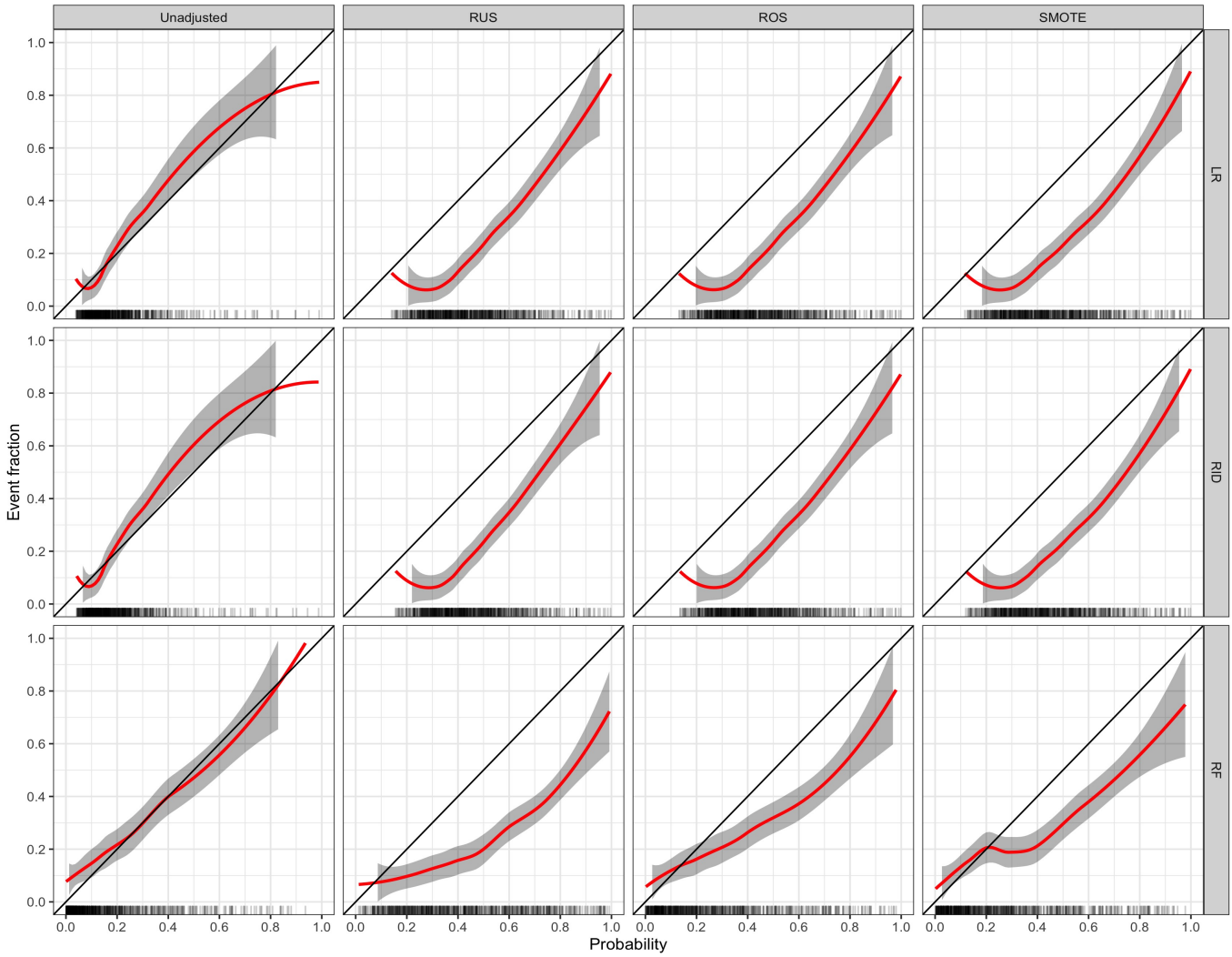

FIGURE 2 Calibration plots. The red line shows the loess curve fitted on the estimated probabilities. The gray band shows the 95% confidence interval of the loess probability estimates. The black diagonal shows the hypothetical situation of perfect estimated probabilities. Bottom rugs display the probability estimates. Abbreviations: LR = Maximum likelihood logistic regression, RID = Ridge logistic regression, RF = Random forest, RUS = Random undersampling, ROS = Random oversampling, SMOTE = Synthetic Minority Oversampling Technique

TABLE 2 Performance of models in combination with data adjustments for handling class imbalance

	Adjustment	Risk threshold = 0.5 Accuracy	Sensitivity	Specificity	c-statistic (CI)	Calibration intercept	Calibration slope
LR	Unadjusted	0.81	0.15	0.98	0.76 (0.71 — 0.80)	0.09 (-0.11 — 0.28)	1.18 (0.93 — 1.45)
	RUS	0.72	0.64	0.74	0.76 (0.71 — 0.80)	-1.28 (-1.48 — -1.09)	1.18 (0.92 — 1.45)
	ROS	0.72	0.64	0.74	0.76 (0.71 — 0.80)	-1.28 (-1.48 — -1.09)	1.13 (0.88 — 1.39)
	SMOTE	0.72	0.66	0.74	0.76 (0.71 — 0.80)	-1.29 (-1.49 — -1.09)	1.08 (0.85 — 1.33)
RID	Unadjusted	0.81	0.14	0.98	0.76 (0.71 — 0.80)	0.09 (-0.11 — 0.28)	1.22 (0.96 — 1.50)
	RUS	0.72	0.64	0.74	0.76 (0.71 — 0.80)	-1.28 (-1.48 — -1.09)	1.24 (0.98 — 1.53)
	ROS	0.72	0.64	0.75	0.76 (0.71 — 0.80)	-1.28 (-1.48 — -1.08)	1.14 (0.90 — 1.40)
	SMOTE	0.72	0.66	0.74	0.76 (0.71 — 0.80)	-1.29 (-1.49 — -1.09)	1.09 (0.86 — 1.35)
RF	Unadjusted	0.80	0.23	0.95	0.73 (0.68 — 0.77)	0.27 (0.05 — 0.48)	0.60 (0.46 — 0.74)
	RUS	0.70	0.64	0.71	0.73 (0.68 — 0.77)	-1.37 (-1.58 — -1.16)	0.67 (0.51 — 0.83)
	ROS	0.78	0.41	0.88	0.72 (0.67 — 0.76)	-0.48 (-0.69 — -0.27)	0.57 (0.43 — 0.71)
	SMOTE	0.78	0.42	0.88	0.71 (0.66 — 0.75)	-0.49 (-0.7 — -0.28)	0.57 (0.43 — 0.71)

Risk threshold of 0.5 used for classification.

Abbreviations: Abbreviations: LR = Maximum likelihood logistic regression, RID = Ridge logistic regression, RF = Random forest, ROS = Random oversampling, RUS = Random undersampling, SMOTE = Synthetic Minority Oversampling Technique, CI = Confidence Interval.

4 | SIMULATION STUDY: METHODS

4.1 | Aim

The aim of this study was to investigate the impact of different common solutions to imbalanced data on model performance. Performance was assessed in terms of discrimination, calibration and classification accuracy. The simulation was designed and reported in line with best practice.[18]

4.2 | Data generating mechanism

This study focused on the situation where a binary outcome variable was predicted using multiple continuous predictor variables. Twenty-four scenarios were investigated by fully crossing the following simulation factors.

- Sample size (N): 2500, 5000
- Number of predictors (R): 3, 6, 12, 24
- Outcome prevalence: 0.3, 0.1, 0.01

Candidate predictor variables were drawn from a multivariate standard normal distribution (with zero correlation between predictors). Then, the outcome probability of each case was computed by applying a logistic function to the generated predictors. The coefficients of this function were approximated numerically for each scenario (Appendix A), such that the predictors were of equal strength, the c-statistic of the data generating model was 0.75 and the outcome prevalence expected in accordance with the simulation condition. The outcome variable was sampled from a binomial distribution, using the computed probabilities.

4.3 | Methods

For each generated development data set, four prediction model development data sets were created: (i) unadjusted; (ii) ROS; (iii) RUS; and (iv) SMOTE. On each of these data sets, the following models were fit: (i) maximum likelihood logistic regression; (ii) ridge logistic regression; and (iii) random forest. This resulted in 12 (3×4) different prediction models per simulation scenario. As miscalibration due to the use of resampling techniques could be expected, I also implemented a logistic re-calibration approach for the models developed on adjusted data, resulting in another 9 (3×3) models. This re-calibration was done by fitting a logistic regression model with the probability estimated by the initial model as an offset variable and the intercept as the only free parameter, making the average predicted probability equal to the overall observed event-rate:[8]

$$log\left(\frac{\pi_{re-calibrated}}{1-\pi_{re-calibrated}}\right) = \alpha_{new} + log\left(\frac{\pi_{original}}{1-\pi_{original}}\right).$$

To convert the estimated probabilities into a dichotomous prediction, the conventional decision threshold of 0.5 was used. For models trained on unadjusted development data sets, performance was also assessed using a second decision threshold, being the true prevalence. In total, this resulted in the assessment of 24 different models per simulation scenario; 12 models resulting from crossing different versions of the development data set with all models, 9 re-calibrated versions of the models trained on the adjusted data sets and 3 models trained on the unadjusted development data sets using an alternative decision threshold.

4.4 | Performance

For each scenario, 2000 development sets were generated. To evaluate the performance of all models for a given scenario, one validation set per scenario was generated of size $N = 100{,}000$. For the assessment of the models, the same measures for predictive performance where used as in the case study (Table 1)

4.5 | Software and error handling

All analyses were performed using `R` (version 3.6.2)[19] executed on a high-performance computing facility running on a Linux-based Operating System (CentOS7). To fit the regression and machine learning models, the R packages `glmnet` (Version

4.0-2)[20] and `randomForest` (Version 4.6-14)[16] were used. To implement SMOTE and simulate data from a multivariate normal distribution, I respectively used the `smotefamily` (Version 1.3.1)[11] and the `MASS` (Version 7.3-51.5)[21] R packages. Errors in the generation of the development data sets and estimation of the models were closely monitored (details in Appendix B). A summary of the data sets in which data separation occured is given in in Table 3

5 | RESULTS

Because the simulation results differed little between scenarios, the main document focuses on the scenario with R= 12 predictors and sample size $N = 5000$. Detailed results of the other scenarios are provided in the supplementary tables and figures. In Table 4 and Table 5, the results averaged over different scenarios can be found. Averaging was done by taking the arithmetic mean over the median performance measures of all scenarios. By using the median values, the influence of outliers was reduced.

5.1 | Average performance before re-calibration

Figure 3 summarizes the performance of the different models before re-calibration. The left panels show that for all models, the resampling techniques result in negative calibration intercepts, indicating that probabilities are systematically overestimated. This effect progresses as the prevalence gets lower, i.e., as the class imbalance is more extreme. Ergo, the more extreme the class imbalance, the more the risk of individual patients is overestimated when resampling techniques are used. The middle column of Figure 3 shows that the average predictive performance of the ridge- and maximum likelihood logistic regression does not improve on any of the predictive performance measures. Where SMOTE and ROS have no noticeable effect on the average c-statistic of those models, regression models trained on RUS data have a lower average c-statistic than the models trained on the unadjusted development data sets. The contrary seems to be true for the random forest models. Especially RUS shows to improve the discriminative performance of the random forest model, as the prevalence decreases. The calibration slopes in the right panels of Figure 3 show that maximum likelihood logistic regression and random forest tend to overfit when the prevalence is low. Especially the random forest model seems to estimate too extreme probabilities in this situation, as indicated by the calibration slopes. Ridge logistic regression shows to slightly underfit in this situation on unadjusted and RUS data, and to overfit on SMOTE and ROS data. Contrary to the random forest models, the regression models seem to approach the ideal calibration slope value of 1, regardless of the data adjustment technique employed, as the prevalence gets closer to 0.5.

5.2 | Average performance after re-calibration

To overcome the CIL problems caused by resampling techniques, all models were re-calibrated by means of logistic re-calibration. For the logistic regression models, this method seems to work properly, as the calibration intercepts in the left panels of Figure 4 are close to the ideal value of zero. Re-calibration did not appear to affect the calibration slopes, since these did not change. The c-statistic also remained the same for the re-calibrated models, as by re-calibrating the intercept the rank order of the probabilities does not change.

Random forest models developed on ROS data with an original prevalence of 0.01, show to have extremely large negative calibration intercepts (Table 5). This is caused by the fact that these models estimate a large part of the probabilities to be 0. This leads to problems in logistic re-calibration, because the logarithm of 0 is undefined. The process of how is dealt with probability estimates of 0 is described in appendix B. The same problems occur for random forest in combination with a low prevalence and SMOTE, but less frequently. Hence, the median calibration intercept of these models is not as extreme as it is using ROS. For random forest models developed on ROS data with an original prevalence of 0.1, re-calibration also led to undesirable results in the form of a positive calibration intercept (Figure 4), indicating a systematic underestimation of the individual patient risks. For the random forest models developed on unadjusted data, the results show no effect of re-calibrating the model on any of the performance measures. For random forest in combination with RUS, the simulation results suggest that logistic re-calibration may improve the CIL.

5.3 | Variability in performance and data separation

Figure 5 and Figure 6 show the variability of the c-statistic, calibration intercept and calibration slope under different models estimated on development data with a prevalence of 0.01. Figure 5 displays that RUS results in performance measures that vary much more over different development data sets than ROS, SMOTE or unadjusted data. The same pattern is shown by the calibration slopes in Figure 6. For visualization purposes, the most extreme slopes are not included in Figure 6, but extreme calibration slopes of >600 have been observed in models using RUS data. This indicates that the probability estimates in these cases are almost identical for all cases, regardless of whether the cases are events or non-events.

Table 3 shows the cases of data separation in each scenario where the prevalence was 0.01. It can be seen that data separation is not a problem in the unadjusted and oversampled development data sets. However, data separation occurs in the development data sets that are undersampled. In these data sets of size $N = 2500$, in combination with the large number of predictors of $R = 24$, data separation occurred more than 50% of the time. In the scenarios where the prevalence was 0.1 or 0.3, data separation did not occur in the development data sets.

TABLE 3 Data separation

	$n = 2500$				$n = 5000$			
# predictors	3	6	12	24	3	6	12	24
Unadjusted	0	0	0	0	0	0	0	0
RUS	0	3	52	1238	0	0	0	9
ROS	0	0	0	0	0	0	0	0
SMOTE	0	0	0	0	0	0	0	0

Number of separated development data sets for prevalence = 0.01, total number of generated development data sets per cell is 2000.
Abbreviations: RUS = Random undersampling, ROS = Random oversampling, SMOTE = Synthetic Minority Oversampling Technique

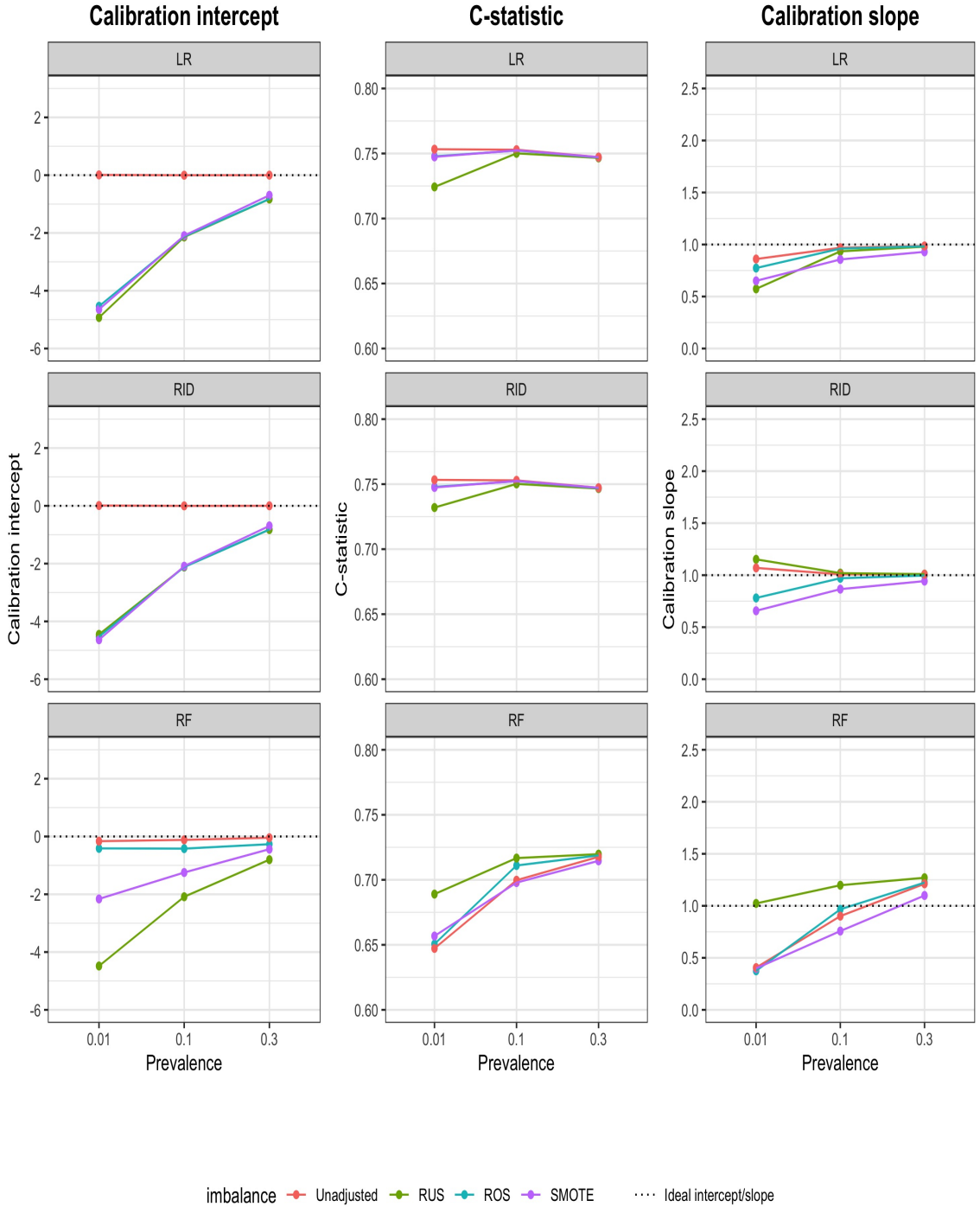

FIGURE 3 Model performance before re-calibration. Median c-statistics, calibration intercepts and calibration slopes. Sample size $N = 5000$, number of predictors $R = 12$. Abbreviations: LR = Maximum likelihood logistic regression, RID = Ridge logistic regression, RF = Random forest, RUS = Random undersampling, ROS = Random oversampling, SMOTE = Synthetic Minority Oversampling Technique

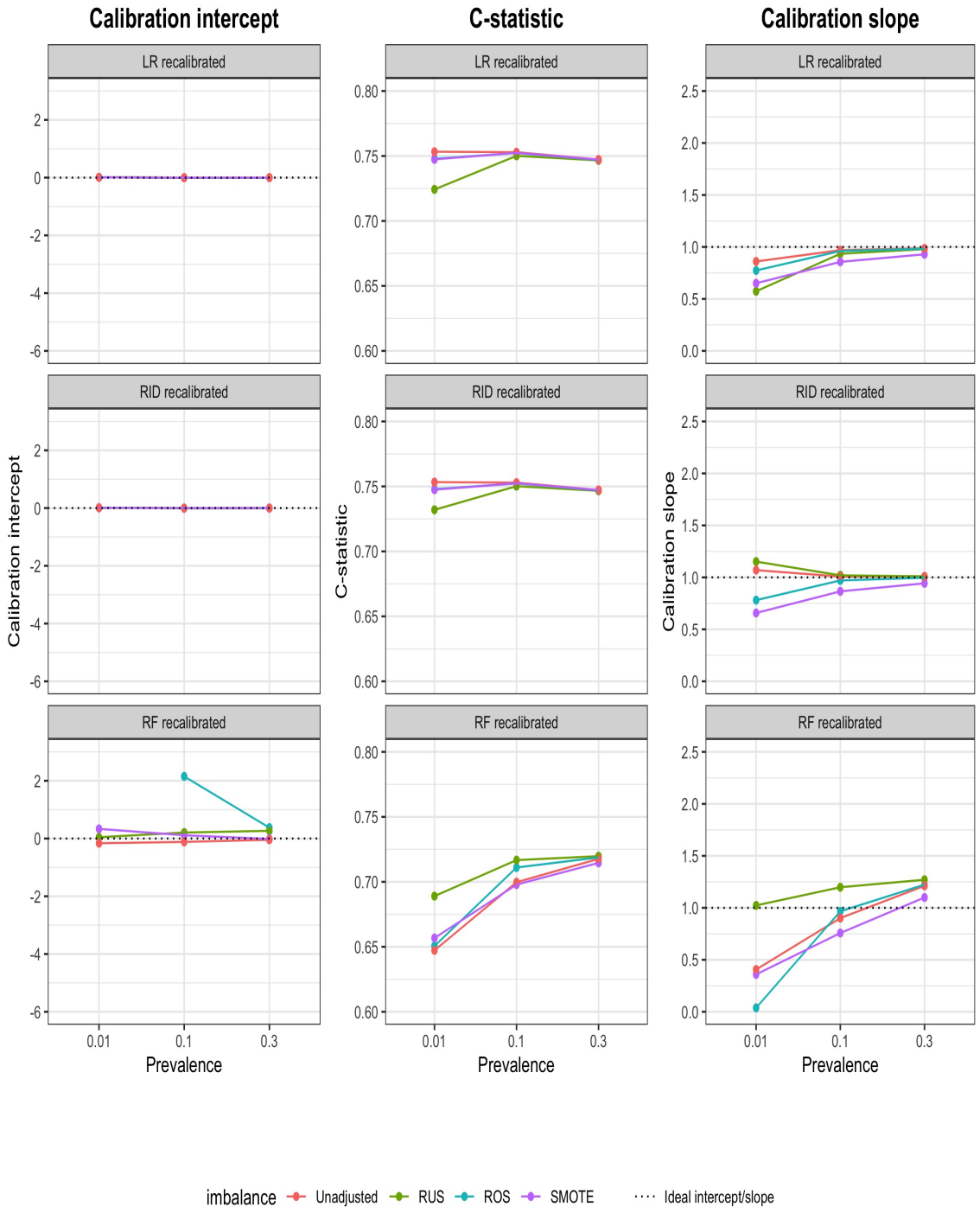

FIGURE 4 Model performance after re-calibration. Median c-statistics, calibration intercepts and calibration slopes. Sample size $N = 5000$, number of predictors $R = 12$. Calibration intercepts are winsorized at -6 and 2.5 for visualization purposes. Abbreviations: LR = Maximum likelihood logistic regression, RID = Ridge logistic regression, RF = Random forest

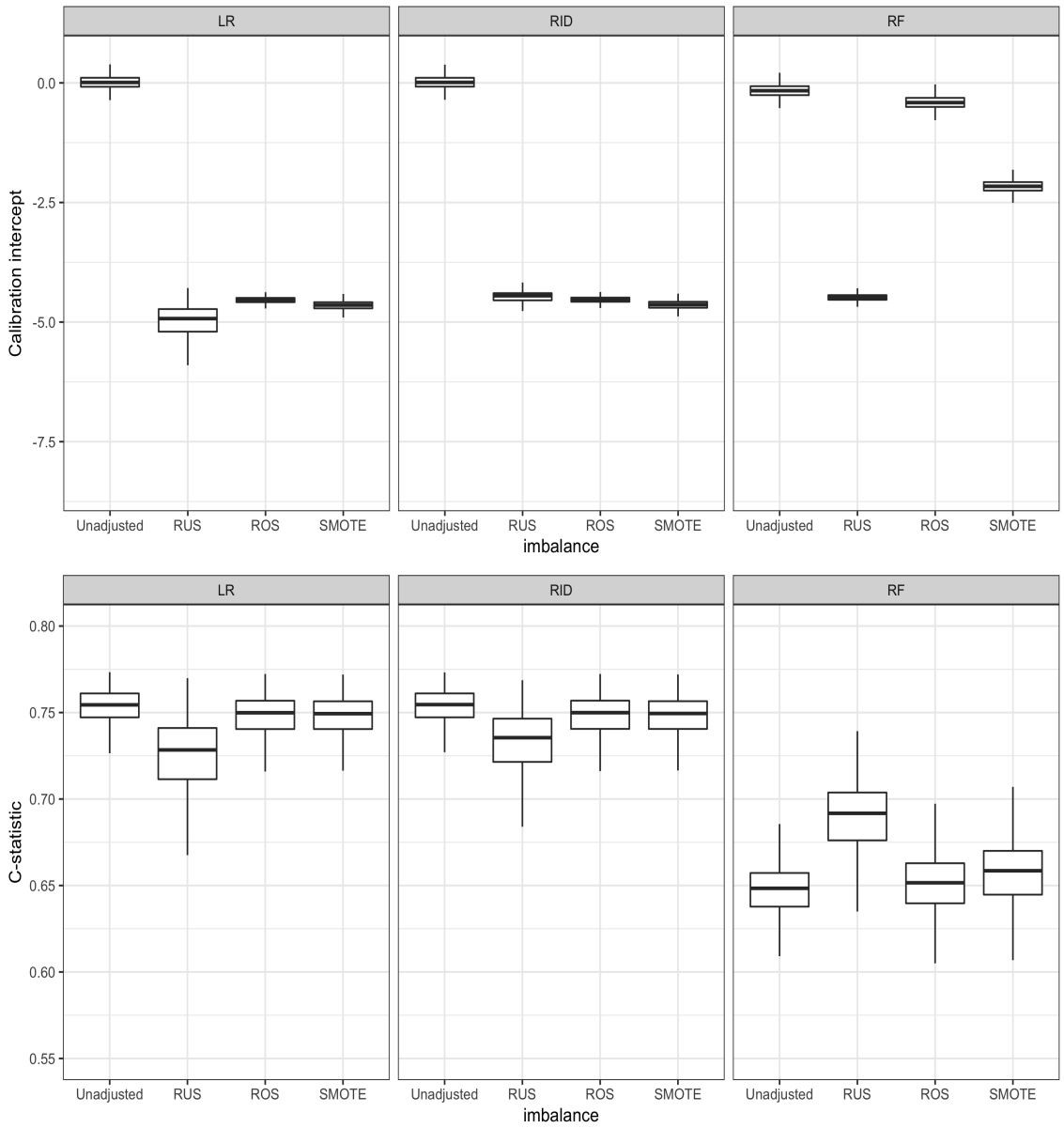

FIGURE 5 Boxplots of the C-statistics and calibration intercepts over 2000 simulation iterations. The length of the whiskers is 1.5 times the interquartile range. Sample size $N = 5000$, number of predictors $R = 12$, prevalence = 0.01. Abbreviations: LR = Maximum likelihood logistic regression, RID = Ridge logistic regression, RF = Random forest, RUS = Random undersampling, ROS = Random oversampling, SMOTE = Synthetic Minority Oversampling Technique

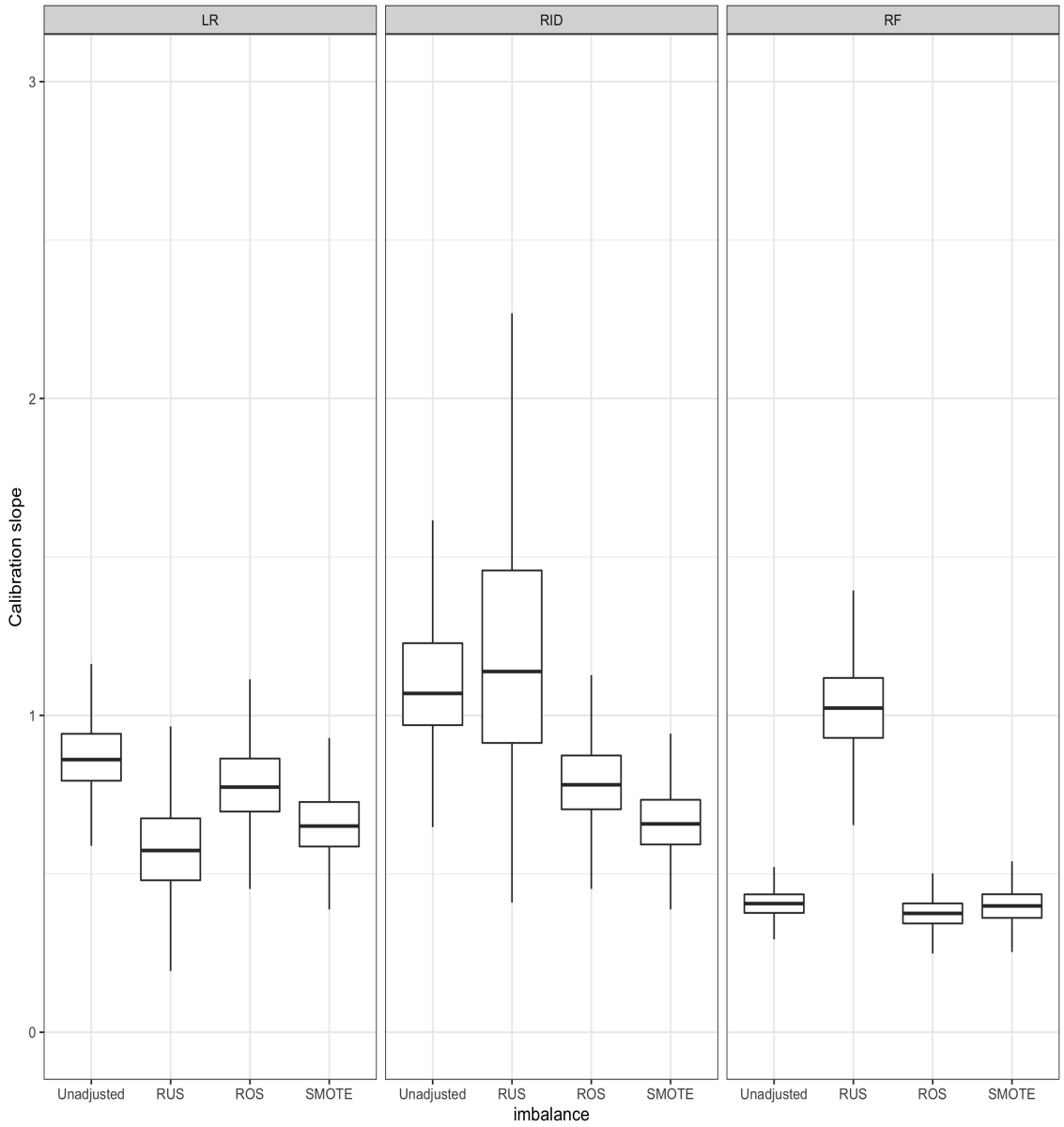

FIGURE 6 Boxplots of the Calibration slopes over 2000 simulation iterations. The length of the whiskers is 1.5 times the interquartile range. Calibration slopes are winsorized at 0 and 3 for visualization purposes. Sample size $N = 5000$, number of predictors $R = 12$, prevalence = 0.01. Abbreviations: LR = Maximum likelihood logistic regression, RID = Ridge logistic regression, RF = Random forest, RUS = Random undersampling, ROS = Random oversampling, SMOTE = Synthetic Minority Oversampling Technique

TABLE 4 Performance of models averaged over scenarios with different number of predictors and sample size

Model	Adjustment	Prevalence	Accuracy	Sensitivity	Specificity	C-statistic	Calibration intercept	Calibration slope
LR	Unadjusted	0.01	0.99	0.00	1.00	0.73	0.00	0.79
		0.1	0.90	0.03	1.00	0.74	-0.01	0.95
		0.3	0.73	0.33	0.91	0.74	0.00	0.98
	RUS	0.01	0.64	0.64	0.64	0.69	-5.94	0.51
		0.1	0.67	0.67	0.67	0.74	-2.18	0.91
		0.3	0.67	0.68	0.67	0.74	-0.84	0.97
	ROS	0.01	0.71	0.61	0.71	0.72	-4.68	0.70
		0.1	0.68	0.67	0.68	0.74	-2.16	0.94
		0.3	0.67	0.68	0.67	0.74	-0.83	0.97
	SMOTE	0.01	0.72	0.60	0.72	0.72	-4.77	0.60
		0.1	0.70	0.65	0.70	0.74	-2.12	0.86
		0.3	0.69	0.62	0.72	0.74	-0.70	0.93
RID	Unadjusted	0.01	0.99	0.00	1.00	0.73	0.00	1.08
		0.1	0.90	0.03	1.00	0.74	0.00	1.01
		0.3	0.73	0.32	0.92	0.74	0.00	1.01
	RUS	0.01	0.64	0.65	0.64	0.71	-4.54	1.24
		0.1	0.67	0.67	0.67	0.74	-2.15	1.03
		0.3	0.67	0.68	0.67	0.74	-0.83	1.01
	ROS	0.01	0.71	0.61	0.71	0.72	-4.66	0.70
		0.1	0.68	0.67	0.68	0.74	-2.15	0.95
		0.3	0.67	0.68	0.67	0.74	-0.83	0.99
	SMOTE	0.01	0.72	0.60	0.72	0.72	-4.78	0.60
		0.1	0.70	0.65	0.70	0.74	-2.12	0.87
		0.3	0.69	0.62	0.73	0.74	-0.69	0.95
RF	Unadjusted	0.01	0.99	0.00	1.00	0.62	-0.13	0.31
		0.1	0.89	0.02	1.00	0.68	-0.09	0.72
		0.3	0.72	0.24	0.92	0.70	-0.04	1.05
	RUS	0.01	0.61	0.62	0.61	0.66	-4.67	0.80
		0.1	0.64	0.64	0.64	0.69	-2.19	1.02
		0.3	0.65	0.65	0.65	0.70	-0.85	1.12
	ROS	0.01	0.99	0.00	1.00	0.62	-0.43	0.29
		0.1	0.89	0.05	0.99	0.69	-0.43	0.78
		0.3	0.71	0.34	0.87	0.70	-0.28	1.06
	SMOTE	0.01	0.96	0.06	0.97	0.63	-2.61	0.30
		0.1	0.83	0.24	0.90	0.68	-1.34	0.63
		0.3	0.70	0.43	0.81	0.70	-0.45	0.96

Abbreviations: LR = Maximum likelihood logistic regression, RID = Ridge logistic regression, RF = Random forest, RUS = Random undersampling, ROS = Random oversampling, SMOTE = Synthetic Minority Oversampling Technique

TABLE 5 Performance of re-calibrated models and models with prevalence as risk threshold averaged over scenarios with different number of predictors and sample size

Model	Adjustment	Prevalence	Accuracy	Sensitivity	Specificity	C-statistic	Calibration intercept	Calibration slope
LR	Unadjusted	0.01	0.69	0.63	0.70	0.73	-0.01	0.80
		0.1	0.66	0.69	0.66	0.74	0.00	0.95
		0.3	0.67	0.68	0.67	0.74	0.00	0.98
	RUS	0.01	0.99	0.01	1.00	0.69	-0.05	0.51
		0.1	0.90	0.04	1.00	0.74	0.00	0.91
		0.3	0.73	0.33	0.91	0.74	0.00	0.97
	ROS	0.01	0.99	0.00	1.00	0.72	0.00	0.69
		0.1	0.90	0.03	1.00	0.74	0.00	0.94
		0.3	0.73	0.33	0.91	0.74	0.00	0.97
	SMOTE	0.01	0.99	0.00	1.00	0.72	0.00	0.60
		0.1	0.90	0.05	0.99	0.74	0.00	0.86
		0.3	0.73	0.34	0.91	0.74	-0.01	0.93
RID	Unadjusted	0.01	0.64	0.68	0.64	0.73	0.01	1.08
		0.1	0.66	0.70	0.65	0.74	0.00	1.01
		0.3	0.67	0.68	0.67	0.74	0.00	1.01
	RUS	0.01	0.99	0.00	1.00	0.71	-0.02	1.24
		0.1	0.90	0.02	1.00	0.74	0.00	1.03
		0.3	0.73	0.31	0.92	0.74	0.00	1.01
	ROS	0.01	0.99	0.00	1.00	0.72	-0.01	0.70
		0.1	0.90	0.03	1.00	0.74	0.00	0.95
		0.3	0.73	0.32	0.91	0.74	-0.00	0.99
	SMOTE	0.01	0.99	0.00	1.00	0.72	-0.01	0.60
		0.1	0.90	0.05	0.99	0.74	0.00	0.87
		0.3	0.73	0.33	0.91	0.74	0.00	0.95
RF	Unadjusted	0.01	0.72	0.44	0.73	0.62	-0.13	0.31
		0.1	0.60	0.66	0.59	0.68	-0.09	0.72
		0.3	0.63	0.68	0.61	0.70	-0.04	1.05
	RUS	0.01	0.99	0.00	1.00	0.66	0.01	0.80
		0.1	0.90	0.01	1.00	0.69	0.20	1.02
		0.3	0.71	0.14	0.96	0.70	0.26	1.12
	ROS	0.01	0.99	0.00	1.00	0.62	-3.12e+15	0.03
		0.1	0.90	0.00	1.00	0.68	2.21	0.76
		0.3	0.71	0.17	0.95	0.70	0.38	1.10
	SMOTE	0.01	0.99	0.00	1.00	0.63	-0.02	0.23
		0.1	0.89	0.04	0.98	0.67	0.03	0.64
		0.3	0.71	0.31	0.88	0.70	-0.03	0.98

Abbreviations: LR = Maximum likelihood logistic regression, RID = Ridge logistic regression, RF = Random forest, RUS = Random undersampling, ROS = Random oversampling, SMOTE = Synthetic Minority Oversampling Technique

6 | DISCUSSION

This paper investigates the effects of resampling techniques on the performance of clinical prediction models and points out the pitfalls of these techniques. Imbalanced data as a result of a low disease prevalence has gained increasing attention in the literature,[6,7,22] possibly under influence of the focus on imbalanced data in other fields such as computer science. To illustrate the effect of resampling methods on logistic regression models and on an accuracy-oriented classifier (random forest), an elaborate simulation study was conducted. I will here discuss separately: i) the effect of resampling methods on the performance of all models; ii) data related problems that may arise due to the use of resampling techniques.

6.1 | Resampling techniques and performance

The results show that resampling techniques have an unequivocal effect on the calibration in the large of the models examined, indicating a systematic overestimation of the risk estimates of individuals. An explanation is that, by using resampling techniques, the unconditional probability of an event in the development set is altered by changing the outcome prevalence.[23,24] This can be seen as a form of sample selection bias,[25] which results in violating the assumption of all prediction models that the development data set is representative for the target population. Not adjusting for this bias therefore leads to overestimation of the probabilities. One way of correcting for this bias is by re-calibration of the intercept. While re-calibration worked for the regression models in the simulations, it did not yield a performance benefit over the unadjusted data. For random forest, the re-calibration of models developed on ROS -and to a lesser degree SMOTE- data with a low prevalence was also ineffective, which may be due to many probability estimates of zero. These probability estimates may occur due to the fact that oversampling techniques place a lot of weight on very specific regions in the sampling space, especially when the number of minority cases is small. This problem is known as overfitting.[26] This also explains why this problem occurs more frequently on ROS data than on SMOTE data, as models trained on ROS data are more prone to overfit.[9] However, re-calibration was effective for random forest models developed on undersampled data: it substantially reduced the negative effect of RUS on the calibration intercept, while maintaining the improvement of the c-statistic.

Another important result of this study is that resampling methods do not show to improve the discriminative ability of the logistic regression models. This result could be expected, as the logistic regression models do not have an accuracy based loss function. This means that logistic regression models do not suffer in discriminative performance from class imbalance by design. Random forest, on the other hand, does seem to benefit from resampling approaches, especially when the outcome prevalence is low. This is in line with previous research on the discriminative performance of random forest in combination with resampling methods in classification tasks in the biomedical domain.[5,27]

6.2 | Overfitting and small data set problems

The probability estimates by the regression models are affected by the use of resampling techniques. The simulations show that by using oversampling techniques, the estimated probabilities get too extreme when the prevalence is low. This effect can be explained by the fact that by using these techniques, we act as we have more information than that we actually have, due to adding non-cases that are either direct copies of existing non-cases (ROS) or by creating similar non-cases (SMOTE). Conversely, with undersampling we remove information. This results in too extreme probability estimates using maximum likelihood logistic regression, for ridge logistic regression the probability estimates appear to be too moderate. Typically, by removing a large part of the data, the penalty in the penalized log-likelihood function used to estimate the ridge logistic regression model gets a relatively higher weight, thus increasing the amount of shrinkage of the regression coefficients. In the most extreme cases, this leads to a regression model where the coefficients of predictors that are truly related to the outcome, are estimated to be (close to) zero, explaining the extreme observed calibration slopes. Another effect of discarding a large part of the data using RUS, is that development data sets can get very small, especially when the prevalence is low. In the case of an original sample size of $N = 2500$ and a prevalence of 0.01, using RUS results in a development data set of only size $n = 50$. Because smaller data sets are more likely to be separated data sets,[28] using RUS may have the undesirable side effect of separating the development data, besides the obvious issues for estimating precision.

This study also has some limitations. First, only the performance of the models when the imbalanced data were adjusted to a 1:1 ratio was investigated. Previous research suggests that adjusting to a different ratio might yield other results.[29] Second, both SMOTE and random forest allow for extensive tuning of the hyperparameters.[9,15] In this study, however, default settings

were used. Therefore, this research does not give an indication of the full potential of these techniques and thus should not be interpreted as such. Variations may increase or reduce the problems with imbalance adjustments that I showed here, but I believe they are likely to remain a significant problem for calibration of the prediction model.

7 | CONCLUSION

This paper showed the impact of three popular resampling techniques applied to the context of clinical risk prediction. Where adjusting for imbalanced data is common practice in the field of machine learning, the results clearly show that clinical prediction modellers should be very careful employing these resampling methods, especially when the calibration of the model is of interest. Logistic regression models show not to benefit from resampling techniques in terms of an improved discriminative performance. This is an important result, as logistic regression remains one of the most popular methods in clinical prediction modelling.[2]

8 | SUPPLEMENTARY MATERIAL

All supplementary tables and figures are available online via: https://github.com/Goorbergh/resampling_techniquesCPM.git

APPENDIX

A COEFFICIENT ESTIMATION

The intercepts and coefficients that result in the desired true AUC and prevalence were estimated by numerical optimization using the `optim()` function from the `stats` package. The method used for optimization was `BFGS`. The function to be minimized was the sum of the difference between the observed AUC and the desired AUC squared and the difference between the observed prevalence and the desired prevalence squared. As this method can lead to slightly varying results, the minimization procedure was deployed 20 times after which the median coefficient values were chosen. This process was repeated for 20 generated data sets of size $N = 10^5$, after which again the median coefficient values of all 20 repetitions were chosen. The final coefficient values validated on independently generated data sets of size $N = 10^5$.

B ERROR HANDLING

Errors in the generation of the development data sets were closely monitored, a table summarizing the error occurrence per simulation cell is included in the article. Data separation[30] in the development data sets was assumed when the apparent AUC in the development data set is equal to 1, based on the maximum likelihood logistic regression model. Because in practice clinical prediction modellers should not develop prediction models on separated data, separated data sets were removed from the analysis. Very few cases of data separation, or no cases at all, were expected in the generated development data sets, given that the true AUC will be 0.75 and the minimum sample size is 2,500. However, data separation was likely to occur when random undersampling is used.

If a development data set contained cases of only one class, this data set was excluded from the analysis; the simulation results are based on complete case analysis. Development data sets with fewer than 8 events or non-events can cause severe problems in estimating tuning parameter λ in ridge regression using 10-fold cross validation. In such cases, leave one out cross validation was used to estimate λ. When there are less than 6 minority-class events in the development data set, the SMOTE-algorithm fails when using the default setting because it searches for the k=5 nearest neighbors. In such cases k was set to the number of minority-class events minus 1. Given that the smallest prevalence is 0.01 and the minimum sample size is 2,500, the probability of generating a data set with < 8 events is 0.00002. Hence, none or very few, cases where these errors occur.

Probabilities that were estimated to be 0 or 1 in the validation data set, lead to the inability to estimate the calibration slope and CIL because the log-odds can not be defined. To circumvent this problem, these probabilities were adjusted in the following way

$$\pi(\mathbf{x_i}) = 0 \rightarrow \pi(\mathbf{x_i}) = \frac{\pi_{min}}{2}$$
$$\pi(\mathbf{x_i}) = 1 \rightarrow \pi(\mathbf{x_i}) = 1 - \frac{1 - \pi_{max}}{2}$$

where π_{min} is the smallest non-zero probability estimate, and π_{max} is the largest non-one probability estimate in the development data set. The same strategy was be used in the process of re-calibrating models using maximum likelihood logistic regression when probabilities of 0 or 1 were encountered. Probabilities of 0 and 1 were expected to occur frequently, given the strength of the predictors and the sample size of the validation set.

References

1. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* 2009; 21(9): 1263–1284.
2. Chrisodoulou E, Jie M, Collins G, Steyerberg E, Verbakel J, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019; 110: 12–22. doi: https://doi.org/10.1016/j.jclinepi.2019.02.004
3. Oommen T, Baise L, Vogel R. Sampling Bias and Class Imbalance in Maximum-likelihood Logistic Regression. *Math Geosci* 2011; 43(1): 99–120. doi: https://doi.org/10.1007/s11004-010-9311-8
4. Fernández A, García S, Galar M, Krawczyk B, Herrera F. *Learning from Imbalanced Data Sets*. Springer . 2018.
5. Lee P. Resampling Methods Improve the Predictive Power of Modeling in Class-Imbalanced Datasets. *INT J ENV RES PUB HE* 2014; 11(9): 9776–9789. doi: 10.3390/ijerph110909776
6. Staartjes V, Schröder M. Letter to the Editor. Class imbalance in machine learning for neurosurgical outcome prediction: are our models valid?. *SPINE* 2018; 29(5). doi: 10.3171/2018.5.SPINE18543
7. Kernbach JM, Staartjes VE. Machine learning-based clinical prediction modeling – A practical guide for clinicians. 2020.
8. Steyerberg E. *Clinical Prediction Models: A Practical Approach to Development, Validation and Updating*. Springer International Publishing. 2th ed. 2019.
9. Chawla N, Bowyer K, Hall L, Kegelkeyer W. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res* 2002; 16(1): 321–357. doi: https://doi.org/10.1613/jair.953
10. Fernández A, García S, Herrera F, Chawla N. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *J Artif Intell Res* 2018; 61: 863–905. doi: https://doi.org/10.1613/jair.1.11192
11. Siriseriwan W. *smotefamily: A Collection of Oversampling Techniques for Class Imbalance Problem Based on SMOTE*. 2019. R package version 1.3.1.
12. Beygelzimer A, Kakadet S, Langford J, Arya S, Mount D, Li S. *FNN: Fast Nearest Neighbor Search Algorithms and Applications*. 2019. R package version 1.1.3.
13. Schaefer R, Roi L, Wolfe R. A ridge logistic estimator. *Commun Stat Theory Meth* 1984; 13(1): 99–113. doi: https://doi.org/10.1080/03610928408828664
14. Cessie SL, Houwelingen JCV. Ridge Estimators in Logistic Regression. *J R Stat Soc C* 1992; 41(1): 191–201. doi: https://doi.org/10.2307/2347628
15. Breiman L. Random Forests. *MACH LEARN* 2001; 45(1): 5–32. doi: https://doi.org/10.1023/A:1010933404324
16. Liaw A, Wiener M. Classification and Regression by randomForest. *R News* 2002; 2(3): 18-22.

17. Van Calster B, Van Hoorde K, Valentin L, et al. Evaluating the risk of ovarian cancer before surgery using the ADNEX model to differentiate between benign, borderline, early and advanced stage invasive, and secondary metastatic tumours: prospective multicentre diagnostic study. *BMJ* 2014; 349. doi: 10.1136/bmj.g5920

18. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat. Med.* 2019; 38(11): 2074-2102. doi: https://doi.org/10.1002/sim.8086

19. R Core Team . *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; Vienna, Austria: 2019.

20. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J. Stat. Softw* 2011; 39(5): 1–13. doi: 10.18637/jss.v039.i05

21. Venables WN, Ripley BD. *Modern Applied Statistics with S*. New York: Springer. fourth ed. 2002. ISBN 0-387-95457-0.

22. Mostafizur Rahman M, Davis D. Addressing the Class Imbalance Problem in Medical Datasets. *Int J Mach Learn Comput* 2013; 3(2): 224–228. doi: 10.7763/IJMLC.2013.V3.307

23. Pozzolo AD, Caelen O, Johnson RA, Bontempi G. Calibrating Probability with Undersampling for Unbalanced Classification. In: Proceedings of IEEE Symposium Series on Computational Intelligence; 2015

24. Appice A, Rodrigues PP, Costa VS, Soares C, Gama J, Jorge A., eds.*Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing . 2015

25. Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND., eds.*Dataset Shift in Machine Learning*. The MIT Press . 2008

26. Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter* 2004; 6(1): 20–29. doi: 10.1145/1007730.1007735

27. Dittman DJ, Khoshgoftaar TM, Napolitano A. The Effect of Data Sampling When Using Random Forest on Imbalanced Bioinformatics Data. In: IEEE; 2015

28. van Smeden M, de Groot JAH, Moons KGM, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol* 2016; 16(1). doi: 10.1186/s12874-016-0267-3

29. Estabrooks A, Jo T, Japkowicz N. A Multiple Resampling Method for Learning from Imbalanced Data Sets. *Comput. Intell* 2004; 20(1): 18–36. doi: 10.1111/j.0824-7935.2004.t01-1-00228.x

30. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 1984; 71(1): 1-10. doi: 10.1093/biomet/71.1.1