Universiteit Leiden
The Netherlands

Regression shrinkage methods for clinical prediction models do not guarantee improved performance: simulation study

Calster, B. van; Smeden, M. van; Cock, B. de; Steyerberg, E.W.

Citation

Calster, B. van, Smeden, M. van, Cock, B. de, & Steyerberg, E. W. (2020). Regression shrinkage methods for clinical prediction models do not guarantee improved performance: simulation study. *Statistical Methods In Medical Research*, *29*(11), 3166-3178. doi:10.1177/0962280220921415

Version: Publisher's Version

License: Creative Commons CC BY-NC 4.0 license

Downloaded from: https://hdl.handle.net/1887/3185124

Note: To cite this publication please use the final published version (if applicable).

Check for updates

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2020, Vol. 29(11) 3166–3178
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280220921415
journals.sagepub.com/home/smm
SAGE

Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study

Ben Van Calster[1,2], Maarten van Smeden[2,3], Bavo De Cock[1,4] and Ewout W Steyerberg[2]

Abstract

When developing risk prediction models on datasets with limited sample size, shrinkage methods are recommended. Earlier studies showed that shrinkage results in better predictive performance on average. This simulation study aimed to investigate the variability of regression shrinkage on predictive performance for a binary outcome. We compared standard maximum likelihood with the following shrinkage methods: uniform shrinkage (likelihood-based and bootstrap-based), penalized maximum likelihood (ridge) methods, LASSO logistic regression, adaptive LASSO, and Firth's correction. In the simulation study, we varied the number of predictors and their strength, the correlation between predictors, the event rate of the outcome, and the events per variable. In terms of results, we focused on the calibration slope. The slope indicates whether risk predictions are too extreme (slope < 1) or not extreme enough (slope > 1). The results can be summarized into three main findings. First, shrinkage improved calibration slopes on average. Second, the between-sample variability of calibration slopes was often increased relative to maximum likelihood. In contrast to other shrinkage approaches, Firth's correction had a small shrinkage effect but showed low variability. Third, the correlation between the estimated shrinkage and the optimal shrinkage to remove overfitting was typically negative, with Firth's correction as the exception. We conclude that, despite improved performance on average, shrinkage often worked poorly in individual datasets, in particular when it was most needed. The results imply that shrinkage methods do not solve problems associated with small sample size or low number of events per variable.

Keywords

Clinical risk prediction models, Firth's correction, logistic regression, maximum likelihood, penalized likelihood, shrinkage

1 Introduction

When developing clinical prediction models, the ultimate aim is to obtain risk estimates that work well on patients that were not used to develop the model.[1] To do so, we have to keep statistical overfitting under control. Assuming that data collection was done carefully, and according to standardized procedures and definitions, the values in a dataset reflect (1) true underlying distributions of and associations between variables, and (2) some amount of random variability. Overfitting occurs when a prediction model also captures these random idiosyncrasies of the development dataset, which by definition do not generalize to new data from the same population.[2]

[1]Department of Development and Regeneration, KU Leuven, Leuven, Belgium
[2]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands
[3]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, Netherlands
[4]Department of Accountancy, KU Leuven, Finance and Insurance, Leuven, Belgium

Corresponding author:
Ben Van Calster, Department of Development and Regeneration, KU Leuven, Herestraat 49 Box 805, Leuven 3000, Belgium.
Email: ben.vancalster@kuleuven.be

The risk of an overfitted model increases when the model building strategy is too ambitious for the available data, for example when the number of variables that are tested as potential model predictors is large given the available sample size.

A well-known rule of thumb for sample size for prediction models is to have at least 10 events per variable (EPV).[3–6] For binary outcomes, the number of events is the number of cases in the smallest of the two outcome levels. 'Variables' actually refers to the number of parameters that are considered for inclusion in the model (excluding intercepts). Some parameters may be checked but not included in the final model, and variables may be modeled using more than one parameter. Recent research has indicated that the EPV ≥ 10 rule is too simplistic, and highlights that there are no good rules of thumb regarding sample size.[7–11] Therefore, the use of shrinkage methods is recommended when sample size is small.[5,6] Several studies have suggested that model performance improves on average when shrinkage methods are applied.[5,9,12–17] Some have suggested that shrinkage may be needed for EPV values up to 20 if the model is prespecified.[1] When variable selection has to be performed to develop the model, the required EPV for reliable selection may increase to 50.[1]

Most regression shrinkage methods deliberately induce bias in the coefficient estimates, by shrinking them towards zero, in order to reduce the expected variance in the predictions. As a consequence, for models with a binary outcome, these methods aim to prevent predicted risks that are too extreme, i.e. where small risks are underestimated, and high risks overestimated. This leads to better expected mean squared error of the predictions.[18] Since prediction focuses on reliable predictions, inducing bias in the model coefficients is not a key concern. Therefore, it seems that the use of shrinkage methods is always good when sample size is limited. Moreover, standard maximum likelihood estimation suffers from small sample bias leading to exaggerated coefficient estimates (i.e. away from zero).[6,19] However, some observations are puzzling. Hans van Houwelingen already noted that 'it is surprising to observe that the estimated shrinkage factors can be quite off the mark and are negatively correlated with optimal shrinkage factor'.[20] This would imply that shrinkage methods shrink too little when it is really needed, and vice versa. However, van Houwelingen's paper included only small simulation study focusing on uniform shrinkage factors. It is of interest to see whether this also occurs with other approaches to regression shrinkage, such as LASSO, ridge, and Firth's correction.[19,21,22] Other studies suggest that some methods result in too much shrinkage on average, as indicated by an average calibration slope larger than one.[9,14,16,23] In Box 1, we present an illustration dealing with a prediction model for ovarian cancer diagnosis,[24] to illustrate that standard regression and regression shrinkage may be more variable in performance than many would think.

The aim of this simulation study was to investigate the performance of various modern shrinkage approaches for the validity of clinical prediction models that are developed with small number of predictors relative to the total sample size (low dimensional). This implies a situation in which some preselection of potentially important predictors has been done before the modeling (e.g. by expert opinion or based on previous studies). We address the performance on average, as well as performance for individual simulation runs. The latter is done by evaluating the between-sample variability in the amount of shrinkage provided by various methods, and the correlation between estimated shrinkage and optimal shrinkage.

Box 1 Clinical illustration: prediction model for ovarian cancer diagnosis.

In 2005, the International Ovarian Tumor Analysis group published its first ultrasound-based risk prediction models to diagnose ovarian malignancy in patients that are selected for surgery.[24] The dataset of 1066 patients were randomly split in a development part of 754 (191 with a malignancy) and a validation part of 312 (75 with a malignancy). For the model, over 40 predictors were considered, totaling 52 parameters. The EPV was 3.7 (191/52). Data-driven variable selection was used in the context of standard logistic regression (no shrinkage), leading to a model with 12 predictors. Using the dataset from this study, the model had a calibration intercept of 0.007 and a calibration slope of 1.09 on the validation part. Contrary to expectation, the observed slope suggested mild underfitting: the estimated risks were too close to the overall outcome prevalence. If likelihood-based uniform shrinkage factor were used,[25] predictors coefficients would have been multiplied by 0.89. This implies a shrinkage of 11%, which seems little given the data-driven selection among 52 parameters in a dataset of moderate size. With this method, the calibration slope on the validation part would have been 1.22. Hence, shrinkage worsened the calibration of the model. Obviously, the small size of the validation part set implied considerable random variation. Nevertheless, this illustrates that a thorough assessment of the variability of standard logistic regression and alternatives based on shrinkage is important.

2 Materials and methods

2.1 Regression methods

We will apply standard logistic regression based on maximum likelihood estimation, and compare this to a collection of shrinkage methods within the context of logistic regression. We apply likelihood-based and bootstrap-based uniform shrinkage methods,[12,25] methods that directly shrink coefficient estimates without or with variable selection,[21,22,26–29] and Firth's penalized likelihood.[19,30] We will discuss each method in what follows.

2.1.1 Standard logistic regression

This is the reference method, in which coefficients are determined by maximum likelihood (ML). Hence, no shrinkage is applied here. When the outcome variably Y equals 1 for an event and 0 for a non-event, the probability of an event ($Y = 1$) for patient i (π_i) is estimated based on a weighted combination of p predictor variables X_j. We define π_i as $P(Y = 1|\mathbf{x_i})$, with $i = 1, \ldots, n$, and $\mathbf{x_i} = (1, x_{1,i}, \ldots, x_{p,i})'$. Assuming only linear effects and no interactions between the predictors, the logistic regression has the following form

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

where $\pi_i = \frac{e^{\mathbf{x_i'\beta}}}{1+e^{\mathbf{x_i'\beta}}}$, and $\boldsymbol{\beta}$ a column vector containing the intercept α and the coefficients β_j. Coefficient estimates $\hat{\alpha}$ and $\hat{\beta}_j$ are obtained by finding the maximum of the log-likelihood function

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i log(\pi_i(\boldsymbol{\beta})) + (1-y_i) log(1-\pi_i(\boldsymbol{\beta})) \right\}$$

2.1.2 Likelihood-based uniform shrinkage (LU)

This method uses the likelihood-ratio statistic to compute a uniform shrinkage factor

$$s_{LU} = \frac{\chi^2_{model} - df}{\chi^2_{model}}$$

where χ^2_{model} is the likelihood-ratio statistic of the fitted model based on standard maximum likelihood and df is the degrees of freedom for the number of candidate predictors considered for the model.[25] The shrunk model coefficients are then calculated as $\hat{\beta}_{j,LU} = s_{LU}\hat{\beta}_j$. After adjusting the coefficients, we re-estimated the intercept to guarantee that the average predicted risk equaled the event rate.

2.1.3 Bootstrap-based uniform shrinkage (BU)

The uniform shrinkage factor s can also be computed using a bootstrap procedure:[12]

1. A bootstrap sample is taken from the original data sample, that is, a random sample with replacement of the same size as the original sample.
2. If a selection procedure was used to select variables this is also applied in the bootstrap samples. The regression coefficients are estimated again on the bootstrap sample, $\hat{\boldsymbol{\beta}}_{\mathbf{bt}}$.
3. The linear predictor for each of the observations in the original sample is calculated using $\hat{\boldsymbol{\beta}}_{\mathbf{bt}}$.
4. In the original sample, the linear predictor obtained in the previous step is used to predict the outcome using maximum likelihood. Retain the coefficient for the regression of the linear predictor.
5. Repeat the procedure, steps 1–4, and the average coefficient from step 4 provides the shrinkage factor s_{BU}. We used 200 repetitions.
6. The shrunk coefficients are calculated as $\hat{\beta}_{j,BU} = s_{BU}\hat{\beta}_j$.
7. Re-estimate the intercept using maximum likelihood while keeping $\hat{\beta}_{j,BU}$ fixed.

2.1.4 Classical ridge logistic regression

Regression shrinkage is implemented via the ridge penalty, also known as the quadratic or L2-penalty.[21] Ridge regression was extended to logistic regression initially by Schaefer and colleagues, and later by Le Cessie and Van Houwelingen.[26,27] The following penalized version of the log-likelihood function is maximized

$$\ell(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} \beta_j^2$$

The tuning parameter, λ, controls the amount of shrinkage. Ridge regression shrinks the estimated coefficients towards zero on average, with higher values of λ leading to more shrinkage. This implicitly induces bias in the coefficients. Note that coefficients will not be shrunk to zero and that the intercept term is not penalized. The shrinkage parameter λ is a hyperparameter that has to be estimated ('tuned'). We used 10-fold cross-validation to find the value for λ that minimized the deviance, using a grid of 251 possible values between 0 (no shrinkage) and 64 (very large shrinkage). The 250 non-null values were equidistant on logarithmic scale. We used the glmnet R package to implement ridge logistic regression.[31]

2.1.5 General penalized maximum likelihood estimation

Ridge logistic regression is a special case of penalized maximum likelihood (PML) that maximizes the following function[28]

$$\ell(\boldsymbol{\beta}) - 0.5\lambda \sum_{j=1}^{p} \left(s_j \beta_j\right)^2$$

where s_j are scaling factors that allow more flexibility than classical ridge. In our study, will apply the method as suggested by Harrell.[28] We set the scale factors to the standard deviation of the predictor. As our predictors are simulated as standard normal variable, and we standardize the variables before fitting models, this approach does not differ from classical ridge. However, Harrell suggests to tune the shrinkage parameter based on a Akaike Information Criterion instead of cross-validation, because it is faster and performs slightly better.[28] Following Harrell's suggestion, the tuning parameter was chosen using the corrected Akaike Information Criterion using a similar grid as for classical ridge.[28,32] The rms R package was used to implement this method. In tables and figures, we refer to this method with the abbreviation PML, and to classical ridge regression with the abbreviation L2.

2.1.6 Classical LASSO logistic regression

LASSO is similar to ridge, but uses the L1-penalty that poses a constraint on the sum of the absolute value of the estimated coefficients.[22] For logistic regression, the LASSO optimizes the following function

$$\ell(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} |\beta_j|$$

The L1-penalty allows coefficients to be shrunk to zero, and hence LASSO performs variables selection as well. The shrinkage parameter was tuned using cross-validation in the same way as for classical ridge logistic regression. The glmnet R package was used.

2.1.7 Adaptive LASSO (AL)

The Adaptive LASSO is a variant of the LASSO where a weight is given for each parameter in the penalty term, in order to obtain variable selection consistency.[29] The optimized function is

$$\ell(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} w_j |\beta_j|$$

where $w_j = \frac{1}{\left|\hat{\beta}_j^{init}\right|^{\gamma}}$ contains adaptive weights. The $\hat{\beta}_j^{init}$ are initial coefficient estimates for the predictors. We used the maximum likelihood estimate $\hat{\beta}_j$ as $\hat{\beta}_j^{init}$, and fixed γ at unity.[15,29] Adaptive LASSO shrinks higher absolute values of $\hat{\beta}_j^{init}$ less than lower values. We tuned the shrinkage parameter using cross-validation as for classical LASSO. The glmnet R package was used.

2.1.8 Firth's penalized likelihood

Firth developed a procedure to remove the first-order bias in the regression coefficients based on maximum likelihood.[19,30] To do so, modified score functions are used to estimate model coefficients. This avoids problems with separation, but also shrinks the coefficients. In addition, Firth's correction reduces the variance. In terms of the log-likelihood, Firth's correction optimizes

$$\ell(\boldsymbol{\beta}) + 0.5 log|I(\boldsymbol{\beta})|$$

where $I(\boldsymbol{\beta})$ is the Fisher information matrix evaluated at $\boldsymbol{\beta}$. We used the logistf R package to implement this method. For making predictions based on Firth's correction, the intercept has to be corrected.[16] We used the same intercept re-estimation procedure as for the uniform shrinkage methods.

2.2 Simulation setup

We simulated data to predict a binary outcome. We used a full factorial simulation setup varying the following factors: EPV, the number and strength of predictors, the correlation between predictors, and the outcome event rate (Table 1). In total, this gave us 60 simulation scenarios. In the setting with five true predictors, the true coefficients of the predictors were 0.2, 0.2, 0.2, 0.5, and 0.8. These values were based on the Cohen's d measure of effect size, and would correspond to having three weak predictors (odds ratio 1.22), one moderate predictor (odds ratio 1.65), and one strong predictor (odds ratio 2.23).[33] In the setting with 10 true predictors, six had a coefficient of 0.2, two had a coefficient of 0.5 and two had a coefficient of 0.8. Noise predictors had coefficients of 0. The chosen values of the simulation factors had an impact on the true c-statistic (i.e. area under the receiver operating characteristic curve) of the model, the sample size of the simulated datasets, and the number of cases with an event (Table 2).

For every scenario, the simulations were performed as follows. First, for each of 1,000,000 individuals, the predictor values were generated by draws from a standard multivariate normal distribution, with equal pairwise correlations. The true model formula (linear predictor) was applied to each patient, with the intercept chosen to obtain the target event rate (Table 2). The inverse logit of the linear predictor was the true risk for that individual. Then, the outcome for each patient was generated through a Bernoulli trial using the true risk. A different dataset, but also with 1,000,000 individuals, was generated for model validation. Predictors and outcomes were generated analogous to the development population, which means that our out-of-sample performance corresponds to a large sample internal validation setting.

We executed 1000 simulation runs per simulation condition. For each run, we generated a development dataset of the appropriate size (Table 2) by randomly drawing without replacement from the development population. The event rate was fixed at the target value in each development dataset by applying stratified sampling. Next, the predictor variables were standardized, and all types of models were fitted. When using standard maximum likelihood, separation was suggested when R warned for fitted probabilities of zero or one, or when the model did not converge. In these circumstances, results for standard maximum likelihood were replaced with results based on Firth's correction, because this is a situation where the use of the method is indicated.[19,30] For LU and BU, the shrinkage factor s was calculated for the model using Firth's correction (with bootstrap models for BU also based on Firth's correction). Harrell's suggested PML method often resulted in an error when there was the suggestion of separation. In these cases, we used Firth's correction instead of the PML algorithm. In this way, we could avoid the exclusion of samples that were suggestive of separation.[34] For logistic regression with bootstrap uniform shrinkage, bootstrap models suggestive of separation were replaced by other bootstrap replicates without separation.

The resulting models were validated on the accompanying full validation dataset. We calculated the c-statistic and the calibration slope. Because the development and validation data are based on identical populations, the calibration intercept was of little interest and therefore not calculated.[35] At internal validation (i.e. when the

Table 1. Overview of the simulation factors in the full factorial simulation design.

Simulation factor	Factor levels
Events per variable	3, 5, 10, 20, 50
Predictors	Five true predictors; 10 true predictors; five true and five noise predictors
Correlation between predictors	0, 0.5
Outcome event rate	0.1, 0.5

Table 2. Overview of the characteristics of the 60 simulation scenarios.

Predictors	Correlation	Event rate	EPV	Events	Sample size	True c statistic	True model intercept
Five true predictors, or Five true + five noise predictors	0	0.1	3	15	150	0.75	−2.57
			5	25	250		
			10	50	500		
			20	100	1000		
			50	250	2500		
		0.5	3	15	30	0.74	0
			5	25	50		
			10	50	100		
			20	100	200		
			50	250	500		
	0.5	0.1	3	15	150	0.83	−2.98
			5	25	250		
			10	50	500		
			20	100	1000		
			50	250	2500		
		0.5	3	15	30	0.81	0
			5	25	50		
			10	50	100		
			20	100	200		
			50	250	500		
Ten true predictors	0	0.1	3	30	300	0.82	−2.88
			5	50	500		
			10	100	1000		
			20	200	2000		
			50	500	5000		
		0.5	3	30	60	0.80	0
			5	50	100		
			10	100	200		
			20	200	400		
			50	500	1000		
	0.5	0.1	3	30	300	0.93	−4.34
			5	50	500		
			10	100	1000		
			20	200	2000		
			50	500	5000		
		0.5	3	30	60	0.91	0
			5	50	100		
			10	100	200		
			20	200	400		
			50	500	1000		

underlying population is the same), the calibration slope measures bias of risk predictions in terms of spread.[35,36] A slope below unity suggests that predictions are too extreme: low risks are underestimated, high risks are overestimated. A slope above unity suggests the opposite. We calculated median slopes to assess the deviation from the target value of unity. To investigate the variability in the slope, we calculated the median absolute deviation (MAD) of the log(slope). To combine bias (deviation of slope from unity on average) and variability, we calculated root mean squared distance from the target value (RMSD) of the log(slope) over the 1000 runs. We used the logarithm of the slope to acknowledge its asymmetry. A slope of 0.5 (half the target) corresponds to a similar quantitative deviation to a slope of two (double the target), but in opposite directions. The RMSD was calculated as the square root of the mean of $\left(log(1) - log(slope)\right)^2$ over the 1000 runs. Finally, we calculated the Spearman correlation between the estimated shrinkage and the optimal shrinkage over the 1000 simulation runs. The optimal shrinkage was defined as $log(1) - log(\text{slope}_{\text{ML}})$, with slope_{ML} the slope for the standard maximum likelihood model. The estimated shrinkage for a specific shrinkage approach was defined as $log\left(\text{slope}_{\text{shrinkage}}\right) - log(\text{slope}_{\text{ML}})$. To calculate MAD, RMSD, and correlations, we winsorized slopes at 0.01 to avoid problems with rare instances of negative calibration slopes. When no variables were selected by (adaptive) LASSO, the calibration slope was arbitrarily set at 10 to reflect the extreme amount of underfitting.

R code used for the simulations can be found at GitHub (https://github.com/benvancalster/shrinkagesim/).

3 Results

There were few runs where separation was suggested (Table S1), except in the scenario with three EPV, 10 true predictors, 0.5 correlation and 0.5 event rate. Generally, results differed little between the five predictor and 10 prediction scenarios, therefore we focus here on the scenarios with five true predictors for the main document. Detailed results for all scenarios are provided in supplementary tables and figures.

3.1 Performance on average

The median calibration slope approached unity for all methods as EPV increased (Figure 1, Figure S1, Table S2). The standard maximum likelihood model yielded the lowest median calibration slopes. For classical ridge regression, the median slope at lower EPV values was consistently above unity, suggesting too much shrinkage on average. Harrell's PML and LASSO were better, but in many scenarios showed median slopes above unity as well. Other methods generally had median slopes below unity, with bootstrap uniform shrinkage usually having median slopes closest to unity. The use of Firth's correction was slightly better than maximum likelihood.

The average c-statistics also converged to their respective true values as EPV increased (Figure S2). By design, uniform shrinkage had the same c-statistics as regular maximum likelihood. When predictors were correlated, classical ridge and Harrell's PML had highest c-statistics. When predictors were uncorrelated and no noise predictors were present, LASSO had lower c-statistics than the maximum likelihood model. Adaptive LASSO only had better discrimination than maximum likelihood when noise predictors were present. Firth's correction did not improve the c-statistic.

3.2 Variability in the applied shrinkage

For the scenarios with five true predictors, pairwise correlations of 0.5 between predictors, and an event rate of 50%, box plots of the calibration slopes over the 1000 simulation runs are shown in Figure 2. For all scenarios, box plots are given in Figure S3, and MAD in Figure S4. The variability of the calibration slope after shrinkage was larger than the variability based on maximum likelihood, except when Firth's correction was used. Firth's correction consistently reduced variability (Figure S4). This increased variability was particularly strong when EPV is low, and correlations between predictors were low. Only when there were 10 true predictors with high intercorrelations, most shrinkage methods had lower variability than maximum likelihood.

Generally, shrinkage methods improved the RMSD relative to the maximum likelihood model (Table S3, Figure 3, Figure S5). However, LASSO, adaptive LASSO, classical ridge and Harrell's PML often had higher RMSD than maximum likelihood when predictors were uncorrelated and EPV or sample size was low. Classical ridge and Harrell's PML often showed higher RMSD than other methods when predictors were correlated and EPV was high. Two methods, the bootstrap uniform shrinkage and Firth's correction, always had lower RMSD than maximum likelihood.

Box plots of the c-statistics also showed high between-sample variability for all methods (Figure S6).

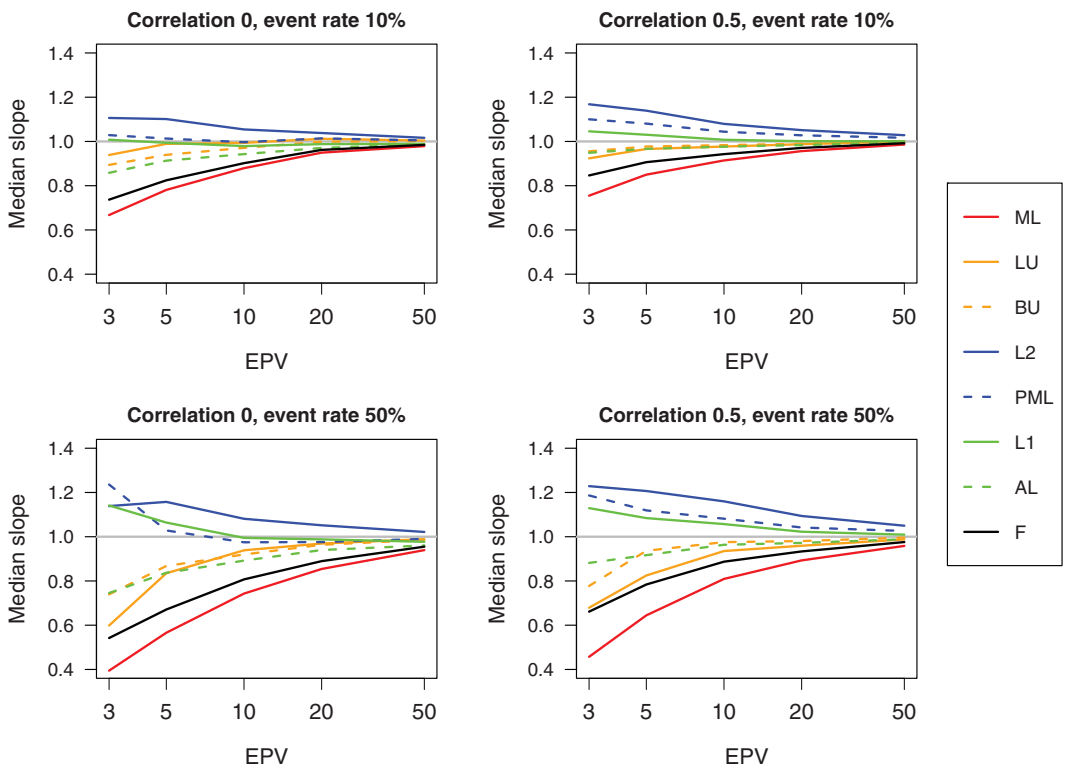

Figure 1. Median calibration slopes for the scenarios with five true predictors. ML: maximum likelihood; LU: uniform shrinkage based on likelihood; BU: uniform shrinkage based on bootstrapping; L2: classical ridge regression; PML: Harrell's penalized maximum likelihood; L1: LASSO regression; AL: adaptive LASSO; F: logistic regression with Firth's correction.

3.3 Correlation between estimated and optimal shrinkage

The Spearman correlation between estimated and optimal shrinkage was typically negative (Figure 4, Table S4, Figures S7–S8). Firth's correction was the exception with consistently positive correlations. LASSO-based methods typically had the lowest negative correlations (closest to zero). For these methods, correlations were highest, and in particular cases even positive, in settings with more highly correlated predictors. The highest positive correlations between estimated and optimal shrinkage were found when there were 10 true predictors, there was non-zero true correlation between the predictors and the EPV was low.

3.4 Results for coefficient estimates and variable selection

Coefficient estimates of true predictors were exaggerated when the maximum likelihood model was used (Figure S9). The bias decreased with increasing EPV. Using Firth's correction removed the bias. All other shrinkage methods induced negative bias and consistently underestimated the coefficients. With respect to noise predictors, classical ridge, Harrell's PML, LASSO, and adaptive LASSO had positive bias in the estimated coefficients when there was correlation between predictors (Figure S10).

Regarding variable selection, adaptive LASSO selected less predictors than standard LASSO implementations (Figure S11). In simulation scenarios with noise predictors, these predictors were selected more often with increasing EPV, except when adaptive LASSO was used (Figure S12). Table S5 summarizes how often these methods selected no variables at all.

4 Discussion

In this paper, we assessed the performance of various shrinkage methods for clinical risk prediction models using simulations. Our key results were the following. First, shrinkage led to calibration slopes that were on average closer to the ideal value of unity than maximum likelihood. Firth's correction improved the slope least among the

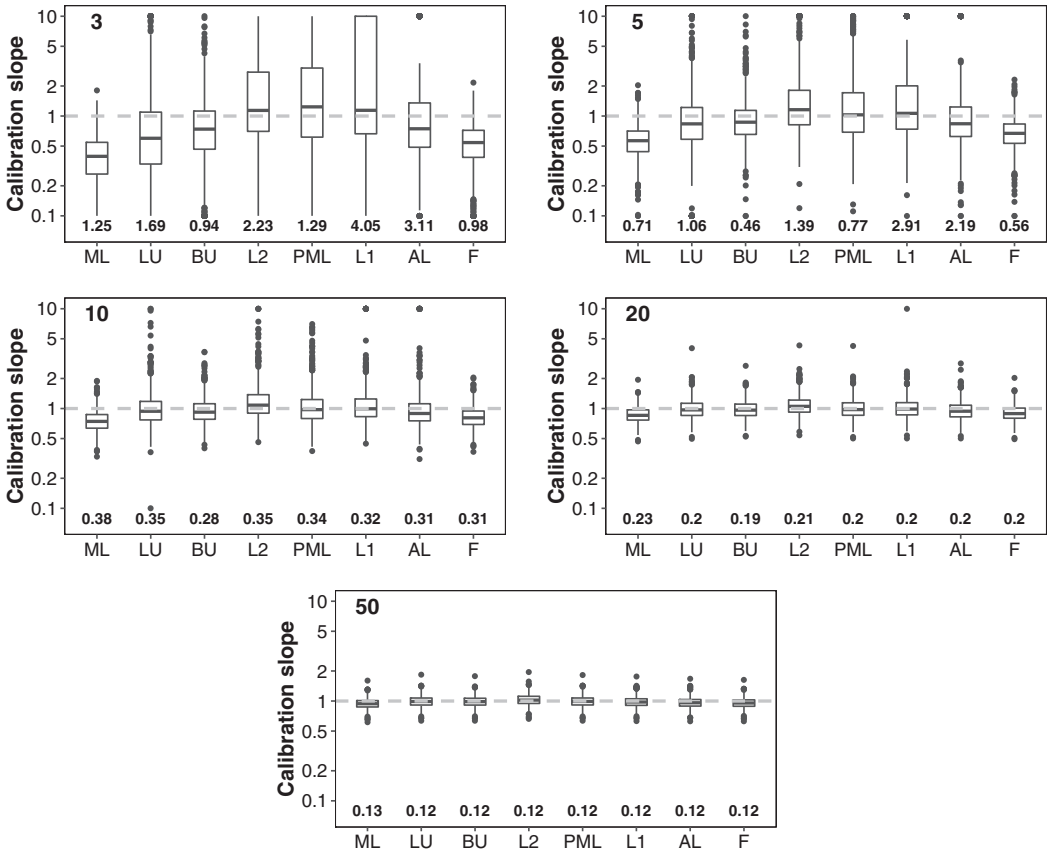

Figure 2. Box plots of the calibration slope over the 1 000 simulation runs for scenarios with five true predictors, no correlation between predictors, and 50% event rate. The events per variable are indicated in the top left. The numbers at the bottom are the root mean squared distances (RMSD) of the log of the calibration slopes. The length of the whiskers is at most 1.5 times the interquartile range. Calibration slopes are winsorized at 0.1 and 10 for visualization purposes. ML: maximum likelihood; LU: uniform shrinkage based on likelihood; BU: uniform shrinkage based on bootstrapping; L2: classical ridge regression; PML: Harrell's penalized maximum likelihood; L1: LASSO regression; AL: adaptive LASSO; F: logistic regression with Firth's correction.

considered methods. Classical ridge, and to a lesser extent Harrell's PML and LASSO, tended to shrink too much overall. Second, the performance of the shrinkage methods was highly variable, especially when sample size was relatively low. The exception was Firth's correction, which showed remarkably stable performance. Despite the increased variance, the RMSD of the calibration slopes was usually lower for shrinkage methods compared to standard maximum likelihood. This was notably the case for Firth's correction, due to its limited variability, but also for bootstrap uniform shrinkage. Third, we commonly observed that the estimated shrinkage was inversely correlated with the optimal shrinkage. This corroborated the early observation by van Houwelingen,[20] and implies that shrinkage often does least when it is needed most. Firth's correction was again the exception, with consistently positive correlations. Fourth, there were differences between the shrinkage methods. A key parameter to this end is the RMSD, because it combines bias in and variability of the calibration slope. Based on RMSD, Firth's correction and bootstrap uniform shrinkage would be the preferred methods. Shrinkage using the bootstrap uniform shrinkage factor performed remarkably well, perhaps because this method explicitly uses the calibration slope for shrinkage estimation. Firth's penalized likelihood almost surely improved performance over maximum likelihood, with low variability and positive correlation with optimal shrinkage. Important advantages of Firth's correction that lead to its stability are that it does not require the estimation of a tuning parameter, and that it shrinks extreme risk estimates. However, the magnitude of shrinkage was small.

These results have implications. Although shrinkage works on average by bringing the calibration slope closer to unity, it may not work as anticipated for any given dataset. The variability in the estimated shrinkage was particularly high when sample size was low. Thus, the use of shrinkage does not justify using lower sample size for the development of prediction models. When sample size is low, it may even be advisable not to build a

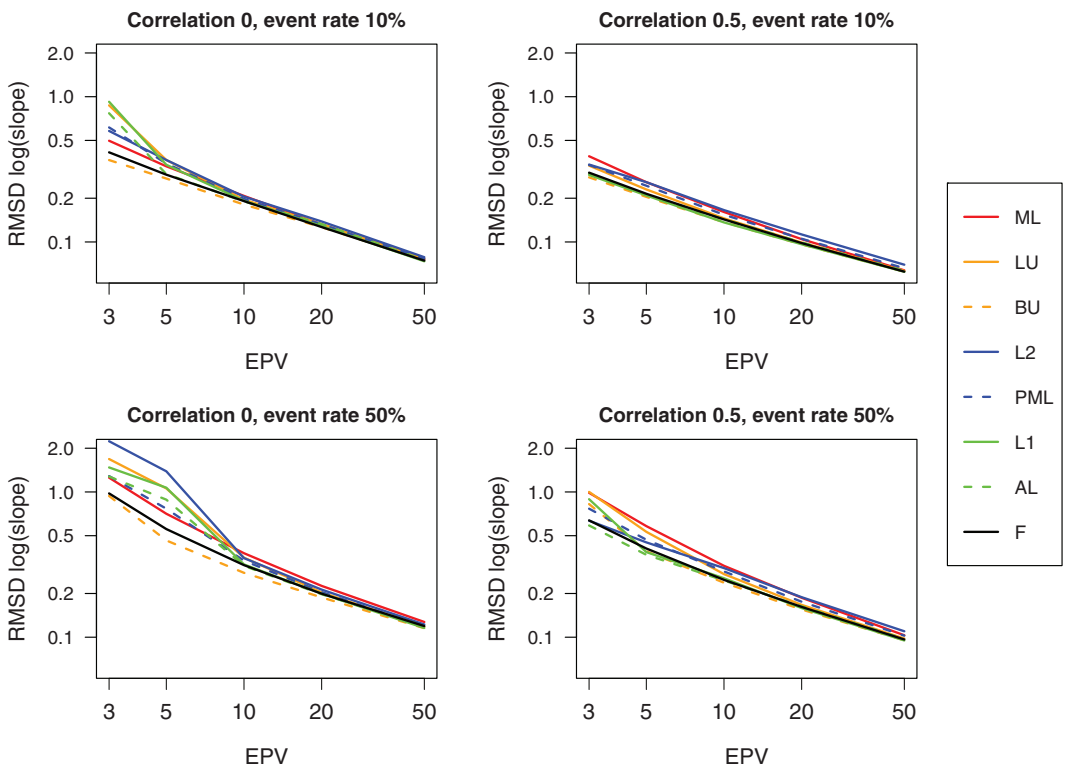

Figure 3. Root mean squared distance (RMSD) of the logarithm of the calibration slope over 1000 simulation runs for scenarios with five true predictors. ML: maximum likelihood; LU: uniform shrinkage based on likelihood; BU: uniform shrinkage based on bootstrapping; L2: classical ridge regression; PML: Harrell's penalized maximum likelihood; L1: LASSO regression; AL: adaptive LASSO; F: logistic regression with Firth's correction.

prediction model. Alternatively, a less complicated model can be considered, for example by discarding many predictors a priori. In a previous study in the context of survival prediction models,[14] the authors suggested that it may be possible to develop an acceptable model with EPV of 2.5 if methods like ridge or LASSO are used, although acknowledged that more work was required.[14] We cannot defend this suggestion based on our results.

We have to be careful about recommendations with respect to specific shrinkage approaches, because the study was not designed to inform fully on their relative merits. For example, classical ridge with tuning based on 10-fold cross-validation led to poorer median calibration slopes than Harrell's PML estimation with tuning based on the corrected Akaike Information Criterion, but had less variability in the calibration slope (Figure S4). More research should study the impact of specific combinations of shrinkage and tuning methods.

The first limitation of our study was the focus on low-dimensional settings for which predictors were largely pre-specified. It would be relevant to investigate the issues of high variability and negative correlation in high-dimensional settings, settings where both sample size and the number of potential predictors are large (such as in some electronic health record studies). Second, we focused on normally distributed predictors, although typical applications also contain non-normal predictors such as skewed continuous predictors or categorical predictors. However, this does not invalidate the key results of our paper. We anticipate the performance variability to become even be larger when a mixture of predictor types is used. Third, we deliberately fixed event rates in simulated datasets, because in prediction model applications one has to go by the event rate that is observed in the data at hand. A downside of this choice is that our results ignore sampling variability in the event rate in observational cross-sectional or cohort studies. Such variability in the observed event rate may further worsen variability in performance. Finally, we investigated many well-known shrinkage methods. Nevertheless, it may be interesting to investigate whether our findings can be confirmed in other approaches, such as elastic net, smoothly clipped absolute deviation (SCAD), weighted fusion, or machine learning methods.[37–39]

Our results are in line with previous work. In line with a large recent simulation study, model performance in our study was related to event rate even when EPV was fixed.[9] Further, the results are consistent with the recommendation to base sample size on a maximal expected level of shrinkage.[10] In accordance with earlier

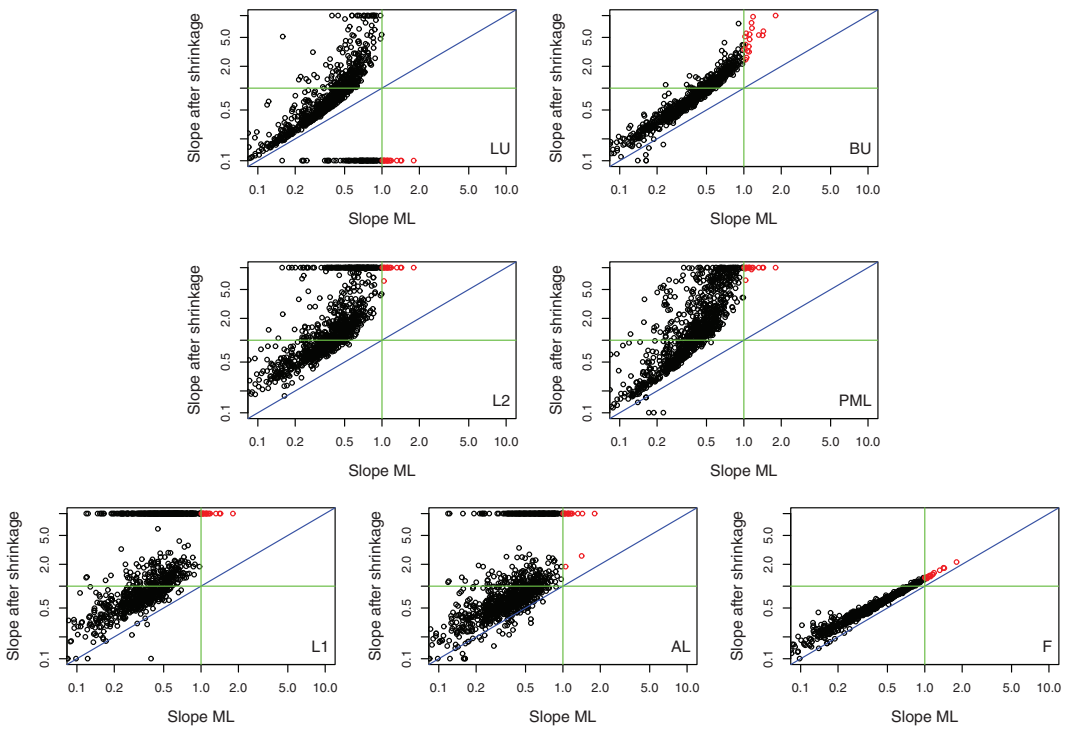

Figure 4. Scatter plots of the slope after shrinkage versus the slope based on maximum likelihood (no shrinkage) for the scenario with five true predictors, no correlation between predictors, 50% event rate, and three events per variable. Each point represents one of the 1000 simulation runs. The blue line is the diagonal, where both slopes are the same. The green lines show the ideal slope (unity). Red circles refer to simulation runs where maximum likelihood resulted in a slope above unity. LU: uniform shrinkage based on likelihood; BU: uniform shrinkage based on bootstrapping; L2: classical ridge regression; PML: Harrell's penalized maximum likelihood; L1: LASSO regression; AL: adaptive LASSO; F: logistic regression with Firth's correction.

work, we observed that methods like ridge or LASSO may have the tendency to shrink too much on average.[8,14–16] Perhaps the use of cross-validation may contribute to this, because shrinkage parameter tuning is based on datasets with reduced sample size. However, in contrast with earlier claims,[6,14] the bootstrap uniform shrinkage method performed relatively well in our simulations. These claims were based on simulations with 2.5 EPV, which is lower than the values considered in our study. Our results do not support the development of prediction models with such low EPV with any method, although more work on settings with very low event rates may be of interest.

In conclusion, shrinkage improves performance on average. The larger variability in calibration slope with the use of shrinkage methods, and the negative correlation between estimated and optimal shrinkage suggest that shrinkage may not work well for any given dataset. Firth's correction is a notable exception, with reduced variability and a positive correlation between estimated and optimal shrinkage. However, the amount of shrinkage it applied was modest. Overall, the use of shrinkage is not a solution to the problem of low sample size or low EPV. In such cases, more fundamental changes are needed, such as refraining from the development of a model, increasing sample size, or reducing a priori the number of predictors if this is clinically acceptable.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Research Foundation – Flanders (FWO) [grant number G0B4716N]; and the Internal Funds KU Leuven [grant number C24/15/037].

ORCID iDs

Ben Van Calster https://orcid.org/0000-0003-1613-7450
Maarten van Smeden https://orcid.org/0000-0002-5529-1541

Supplemental material

Supplemental Material for this article is available online.

References

1. Steyerberg EW. *Clinical prediction models*. 2nd ed. New York, NY: Springer, 2019.
2. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med* 2004; **66**: 411–421.
3. Harrell FE, Lee KL, Califf RM, et al. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984; **3**: 143–152.
4. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996; **49**: 1373–1379.
5. Steyerberg EW, Eijkemans MJ, Harrell FE Jr, et al. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med* 2000; **19**: 1059–1079.
6. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. *BMJ* 2015; **351**: h3868.
7. Courvoisier DS, Combescure C, Agoritsas T, et al. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. *J Clin Epidemiol* 2011; **64**: 993–1000.
8. van Smeden M, de Groot JA, Moons KG, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Med Res Methodol* 2016; **16**: 163.
9. van Smeden M, Moons KG, de Groot JA, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat Meth Med Res* 2019; **28**: 2455–2474.
10. Riley RD, Snell KI, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Stat Med* 2019; **38**: 1276–1296.
11. Ogundimu EO, Altman DG and Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. *J Clin Epidemiol* 2016; **76**: 175–182.
12. Steyerberg EW. Eijkemans MJC and Habbema JDF. Application of shrinkage techniques in logistic regression analysis: a case study. *Stat Neerl* 2001; **55**: 76–88.
13. Steyerberg EW. Eijkemans MJC, Harrell FE Jr, et al. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making* 2001; **21**: 45–56.
14. Ambler G, Seaman S and Omar RZ. An evaluation of penalised survival methods for developing prognostic models with rare events. *Stat Med* 2012; **31**: 1150–1161.
15. Pavlou M, Ambler G, Seaman SR, et al. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Stat Med* 2016; **35**: 1159–1177.
16. Puhr R, Heinze G, Nold M, et al. Firth's logistic regression with rare events: accurate effect estimates and predictions? *Stat Med* 2017; **36**: 2302–2317.
17. De Jong VMT, Eijkemans MJC, Van Calster B, et al. Sample size considerations and predictive performance of multinomial logistic prediction models. *Stat Med* 2019; **38**: 1601–1619.
18. Copas JB. Regression, prediction and shrinkage. *J R Stat Soc B* 1983; **45**: 311–354.
19. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; **80**: 27–38.
20. van Houwelingen JC. Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Stat Neerl* 2001; **55**: 17–34.
21. Hoerl AE and Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970; **12**: 55–67.
22. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc B* 1996; **58**: 267–288.
23. Musoro JZ, Zwinderman AH, Puhan MA, et al. Validation of prediction models based on lasso regression with multiply imputed data. *BMC Med Res Methodol* 2014; **14**: 116.
24. Timmerman D, Testa AC, Bourne T, et al. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group. *J Clin Oncol* 2005; **23**: 8794–8801.
25. Van Houwelingen JC and le Cessie S. Predictive value of statistical models. *Stat Med* 1990; **9**: 1303–1325.
26. Schaefer RL, Roi LD and Wolfe RA. A ridge logistic estimator. *Commun Stat Theory Meth* 1984; **13**: 99–113.
27. Le Cessie S and van Houwelingen JC. Ridge estimators in logistic regression. *J R Stat Soc C* 1992; **41**: 191–201.
28. Harrell FE Jr. *Regression modeling strategies*. 2nd ed. New York, NY: Springer, 2015.
29. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc* 2006; **101**: 1418–1429.

30. Heinze G and Schemper M. A solution to the problem of separation in logistic regression. *Stat Med* 2002; **21**: 2409–2419.
31. Friedman JH, Hastie T and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010; **33**: 1.
32. Hurvich CM and Tsai CL. Regression and time series model selection in small samples. *Biometrika* 1989; **76**: 297–307.
33. Pencina MJ, D'Agostino RB Sr, Pencina KM, et al. Interpreting incremental value of markers added to risk prediction models. *Am J Epidemiol* 2012; **176**: 473–481.
34. Mansournia MA, Geroldinger A, Greenland S, et al. Separation in logistic regression: causes, consequences, and control. *Am J Epidemiol* 2018;**187**:864–870.
35. Van Calster B, Nieboer D, Vergouwe Y, et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016; **74**: 167–176.
36. Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958; **45**: 562–565.
37. Zou H and Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B* 2005; **67**: 301–320.
38. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001; **96**: 1348–1360.
39. Daye ZJ and Jeng XJ. Shrinkage and model selection with correlated variables via weighted fusion. *Comput Stat Data Anal* 2009; **53**: 1284–1298.