Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2019, Vol. 28(8) 2455–2474

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280218784726
journals.sagepub.com/home/smm
SAGE

Sample size for binary logistic prediction models: Beyond events per variable criteria

Maarten van Smeden,[1] Karel GM Moons,[1] Joris AH de Groot,[1] Gary S Collins,[2] Douglas G Altman,[2] Marinus JC Eijkemans[1] and Johannes B Reitsma[1]

Abstract

Binary logistic regression is one of the most frequently applied statistical approaches for developing clinical prediction models. Developers of such models often rely on an Events Per Variable criterion (EPV), notably EPV ≥10, to determine the minimal sample size required and the maximum number of candidate predictors that can be examined. We present an extensive simulation study in which we studied the influence of EPV, events fraction, number of candidate predictors, the correlations and distributions of candidate predictor variables, area under the ROC curve, and predictor effects on out-of-sample predictive performance of prediction models. The out-of-sample performance (calibration, discrimination and probability prediction error) of developed prediction models was studied before and after regression shrinkage and variable selection. The results indicate that EPV does not have a strong relation with metrics of predictive performance, and is not an appropriate criterion for (binary) prediction model development studies. We show that out-of-sample predictive performance can better be approximated by considering the number of predictors, the total sample size and the events fraction. We propose that the development of new sample size criteria for prediction models should be based on these three parameters, and provide suggestions for improving sample size determination.

Keywords

Logistic regression, prediction models, sample size, EPV, simulations, predictive performance

1 Introduction

Binary logistic regression modeling is among the most frequently used approaches for developing multivariable clinical prediction models for binary outcomes.[1,2] Two major categories are: diagnostic prediction models that estimate the probability of a target disease being currently present versus not present; and prognostic prediction models that predict the probability of developing a certain health state or disease outcome over a certain time period.[3] These models are developed to estimate probabilities for new individuals, i.e. individuals that were not part of the data used for developing the model,[3–5] which need to be accurate and estimated with sufficient precision to correctly guide patient management and treatment decisions.

One key contributing factor to obtain robust predictive performance of prediction models is the size of the data set used for development of the prediction model relative to the number of predictors (variables) considered for inclusion in the model (hereinafter referred to as candidate predictors).[4,6–10] For logistic regression analysis, sample size is typically expressed in terms of events per variable (EPV), defined by the ratio of the number of events, i.e. number of observations in the smaller of the two outcome groups, relative to the number of degrees of freedom (parameters) required to represent the predictors considered in developing the prediction model. Lower EPV values in the prediction model development have frequently been associated with poorer predictive performance upon validation.[6,7,9,11–13]

[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands
[2]Centre for Statistics in Medicine, Botnar Research Centre, University of Oxford, Oxford, UK

Corresponding author:
Maarten van Smeden, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands.
Email: M.van_Smeden@lumc.nl

In the medical literature, an EPV of 10 is widely used as the lower limit for developing prediction models that predict a binary outcome.[14,15] This minimal sample size criterion has also generally been accepted as a methodological quality item in appraising published prediction modeling studies.[2,14,16] However, some authors have expressed concerns that that the EPV ≥ 10 rule-of-thumb is not based on convincing scientific reasoning.[17] The rule also did not perform well in large-scale simulation studies.[18–20] Indeed, EPV ≥ 10 has been found too lenient when default stepwise predictor selection strategies are used for development of the prediction model.[11,13] EPV ≥ 50 may be needed when stepwise predictor selection with conventional type I error $\alpha = .05$ is applied.[11] Conversely, more recent work suggests that the EPV ≥ 10 criterion may be too strict in particular settings, showing several examples where prediction models developed with modern regression shrinkage techniques showed good out-of-sample predictive performance in settings with EPV $\ll 10$.[15,21]

Despite all the concerns and controversy, surprisingly few alternatives for considering sample size for logistic regression analysis have been proposed to move beyond EPV criteria, except those that have focused on significance testing of logistic regression coefficients.[22] Sample size calculations for testing single coefficients are of little interest when developing a prediction model to be used for new individuals where the predictive performance of the model as a whole is of primary concern.

Our work is motivated by the lack of sample size guidance and uncertainty about the factors driving the predictive performance of clinical prediction models that are developed using binary logistic regression. We report an extensive simulation study to evaluate out-of-sample predictive performance (hereafter shortened to predictive performance) of developed prediction models, applying several methods for model development. We examine the predictive performance of logistic regression-based prediction models developed using conventional Maximum Likelihood (ML), Ridge regression,[23] Least absolute shrinkage and selection operator (Lasso),[24] Firth's correction[25] and heuristic shrinkage after ML estimation.[26] Backwards elimination predictor selection using the conventional $p = .05$ and $p = .157$ (=AIC) stopping rules is also evaluated. Using a full-factorial approach, we varied EPV, the events fraction, number of candidate predictors, area under the ROC curve (model discrimination), distribution of predictor variables and type of predictor variable effects. The simulation results are summarized using metamodels.[27,28]

This paper is structured as follows. In section 2 we present models and notation. The design of the simulation study is presented in section 3, and the results are described in section 4. A discussion of our findings and its implications for sample size considerations for logistic regression is presented in section 5.

2 Developing a prediction model using logistic regression

2.1 General notation

We define a logistic regression model for estimating the probability of an event occurring ($Y = 1$) versus not occurring ($Y = 0$) given values of (a subset of) P candidate predictors, $\boldsymbol{X} = \{1, X_1, \ldots, X_P\}$. For an individual i ($i = 1, \ldots, N$), let $\pi_i = \Pr(Y = 1|\boldsymbol{x_i}) = 1 - \Pr(Y = 0|\boldsymbol{x_i})$). The logistic model assumes that π_i is an inverse logistic function of $\boldsymbol{x}_i$

$$\pi_i = \frac{1}{1 + \exp\{-(\boldsymbol{\beta}'\boldsymbol{x}_i)\}}$$

where the vector $\boldsymbol{\beta}$ contains an intercept, a scalar, and $P^* \leq P$ regression coefficients corresponding to the log odds ratios for a 1-unit increase in the corresponding predictor (hereinafter referred to as predictor effects), assuming a linear effect for each candidate predictor. At different steps in the prediction model development, the number of predictor effects estimated (P^*) may be smaller than the number of candidate predictors (P) due to predictor selection.

2.2 Maximum likelihood estimation and known finite sample properties

Conventionally, the $P^* + 1$ dimensional parameter vector $\boldsymbol{\beta}$ of the logistic model is estimated by ML estimation, which maximizes the log-likelihood function[29]

$$\log L(\boldsymbol{\beta}) = \sum_i y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i)$$

which are usually derived by iteratively solving the scoring equation: $\partial \log L / \partial \beta_p = 0, p = 0, \ldots, P^*$.

While ML logistic regression remains a popular approach to developing prediction models, ML is also known to possess several finite sample properties that can cause problems when applying the technique in small or sparse data. These properties can be classified into the following five separate and not mutually exclusive issues:

- Issue 1: ML estimators are not optimal for making model predictions of the expected probability (risk) in new individuals. In most circumstances shrinkage estimators can be defined that have lower expected error for estimating probabilities in new individuals than ML estimators.[30,31] The benefits of the shrinkage estimators over the corresponding ML estimators decreases with increasing EPV.[9]
- Issue 2: the predictor effects are finite sample biased.[32,33] The regression coefficients from ML logistic regression models estimate the (multivariable) log-odds ratios for the individual predictors, which are biased towards more extreme effects, i.e. creating optimistic estimates of predictor effects for individual predictors in smaller data sets. This bias reduces with increasing EPV,[34–36] but may not completely disappear even in large samples.[19]
- Issue 3: model estimation becomes instable when predictor effects are large or sparse (i.e. separation).[37,38] The estimated predictor effects tend to become infinitely large in value when a linear combination of predictors can be defined that perfectly discriminates between events and non-events. Extreme probability estimates close to their natural boundaries of 0 or 1 is an undesirable consequence. Separation becomes less likely with increasing EPV.[19]
- Issue 4: model estimation becomes instable when predictors are strongly correlated (i.e. collinearity).[9,39] If correlations between predictors are very strong, the standard errors for the predictor effects become inflated reflecting uncertainty about the effect of the individual predictor, although this has limited effect on the predictive performance of the entire model.[8] With increasing EPV, spurious predictor collinearity becomes less likely.
- Issue 5: commonly used automated predictor selection strategies (e.g. stepwise selection using p-values to decide on predictor inclusion[40]) cause distortions when applied in smaller data sets. In small datasets, predictor selection is known to: (i) lead to unstable models where small changes in the number of individuals – deletion or addition of individuals – can result in different predictors being selected[7,8,41,42]; (ii) cause bias in the predictor effects towards extremer values[10,11]; and (iii) reduce a model's predictive performance when applied in new individuals, due to omission of important predictors (underfitting) or inclusion of many unimportant predictors (overfitting).[9,11] The distortions due to predictor selection typically decrease with increasing EPV.

As these small and sparse data effects can affect the performance of a developed ML prediction model, and thus impact the required sample size for prediction model development studies, we additionally focus on four commonly applied shrinkage estimators for logistic regression. Each of these methods aims to reduce at least one of the aforementioned issues.

2.3 Regression shrinkage

2.3.1 Heuristic shrinkage logistic regression

Van Houwelingen and Le Cessie[43] proposed a heuristic shrinkage (HS) factor to be applied uniformly on all the predictor effects. The shrinkage factor is calculated as

$$\hat{c}_{heur} = \frac{G^2 - P^*}{G^2} \tag{1}$$

where G^2 is the ML logistic regression's likelihood ratio statistic: $-2(\log L(\mathbf{0}) - \log L(\boldsymbol{\beta}))$, with $L(\mathbf{0})$ denoting the likelihood under the intercept-only ML logistic model.[8] The predictor effects of the ML regression are subsequently multiplied with $\hat{c}_{heur}$ to obtain shrunken predictor effect estimates. After shrinkage, the intercept is recalculated by refitting the ML model by taking the shrunken regression coefficients as fixed (i.e. as offset terms).

The HS estimator was developed to improve a model's predictive performance over the ML estimator in smaller data sets (*issue 1*).[43] However, in cases of weak predictor effects the HS estimator can perform poorly, as can be seen from equation (1) that $\hat{c}_{heur}$ takes on a negative value if: $G^2 < P^*$, in which case each of the predictor effects switches sign and a different modeling strategy is recommended. As HS relies on estimating the ML model to calculate the shrinkage factor and intercept, it may be sensitive to ML estimation instability (*issue 3* and *issue 4*).[44]

2.3.2 Firth logistic regression

Firth's penalized likelihood logistic regression model[25,38,45] penalizes the model likelihood by $\log|I(\boldsymbol{\beta})|^{1/2}$, where $I(\boldsymbol{\beta})$ denotes Fisher's information matrix evaluated at $\boldsymbol{\beta}$, $I(\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$, with $\boldsymbol{W} = \mathrm{diag}\{\pi_i(1-\pi_i)\}$. Firth's logistic regression model is estimated by solving the modified scoring equation

$$\sum_i \{y_i - \pi_i + h_i(0.5 - \pi_i)\}x_{ip} = 0, \quad p = 0, \ldots, P^* \tag{2}$$

where h_i is the diagonal element i of matrix $\boldsymbol{W}^{1/2}\boldsymbol{X}\{\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\}^{-1}\boldsymbol{X}'\boldsymbol{W}^{1/2}$ (cf.[38]).

The Firth estimator was initially developed to remove the first-order finite sample bias (*issue 2*) in ML estimators of logistic regression coefficients and other exponential models with canonical links.[25] As a consequence of its penalty function, the regression coefficients remain finite in situations of separation (*issue 3*).[38] More recently, Puhr and colleagues[21] evaluated the Firth's estimator for improving predictive performance (*issue 1*), warning that it introduced bias in predicted probabilities toward $\pi_i = 0.5$, a consequence of the use the Fisher's matrix for penalization that maximizes at $\pi_i = 0.5$.[21]

2.3.3 Ridge logistic regression

Ridge regression penalizes the likelihood proportionally to the sum of squared predictor effects: $\sum_{j=1}^{P^*} \beta_j^2$. For estimation, the predictors are standardized to have mean zero and unit variance.[46] Then, for a particular value of the tuning parameter λ_1, the regression coefficients are estimated using a coordinate decent algorithm that minimizes

$$-L^{\star}(\boldsymbol{\beta}) - \lambda_1 \sum_{j=1}^{P^*} \beta_j^2 \tag{3}$$

where

$$L^{\star}(\boldsymbol{\beta}) \propto -\sum_{i=1}^{N} \tilde{\pi}_i(1-\tilde{\pi}_i)\left(z_i - \beta_0 - \sum_{p=1}^{P^*} x_{ip}\beta_p\right) \tag{4}$$

$$z_i = \tilde{\beta}_0 + \sum_{p=1}^{P^*} x_{ip}\tilde{\beta}_p + \frac{y_i - \tilde{\pi}_i}{\tilde{\pi}_i(1-\tilde{\pi}_i)} \tag{5}$$

with the tilde (˜) denoting that estimates are evaluated at their current value (cf.[47]). The optimal value for the tuning parameter λ_1 can be approximated using cross-validation optimized for a particular performance criterion. In this paper, we apply the commonly used 10-fold cross-validation (whenever possible) with minimal deviance as the performance criterion.

The Ridge estimator was originally developed to deal with collinearity (*issue* 4).[23,39] Due to its penalty function it can also deal with separation (*issue* 3). Moreover, the Ridge estimator has been shown to improve predictive performance in smaller data sets that do not suffer from collinearity or separation (*issue* 1), although in some circumstances it showed signs of underfitting.[15,48]

2.3.4 Least absolute shrinkage and selection operator (Lasso) regression

Lasso regression penalizes the likelihood proportional to the sum of the absolute value of predictor effects: $\sum_{j=1}^{df} |\beta_j|$. Estimating Lasso regression can be done using the same procedure as Ridge regression, where

$$-L^{\star}(\boldsymbol{\beta}) - \lambda_2 \sum_{j=1}^{P^*} |\beta_j| \tag{6}$$

with $L^{\star}(\boldsymbol{\beta})$ defined by equation (4). Similar to Ridge regression, in this paper we use 10-fold cross-validation with minimal deviance as the performance criterion to define the tuning parameter.

Lasso regression is attractive for developing prediction models as it simultaneously performs regression shrinkage (addressing *issue* 1) and predictor selection (by shrinking some coefficients to zero), while avoiding some of the adverse effects of regular automated predictor selection strategies (*issue* 5). It is also suited to handle separation (*issue* 3), but in the context of highly correlated predictors (*issue* 4), the Lasso has been reported to perform less well.[15,49]

3 Methods

This simulation study was set up to evaluate the predictive performance of various prediction modeling strategies in relation to characteristics of the development data. Our primary interest was in the size of the development data set relative to other data characteristics, such as the number of candidate predictors and the events fraction (i.e. $\Pr(Y=1)$). The various modeling strategies we considered are described in section 3 3.1, and the variations in data characteristics are described in section 3.2. A description of the predictive performance metrics and metamodels are given in sections 3.3.1 and 3.3.2, respectively. Software and error handling are given in section 3.4.

3.1 Modeling strategies

The predictive performance of the various logistic regression models as described in section 2 were evaluated on large sample validation data sets. These regressions correspond to different ways of applying regression shrinkage (ML regression applies none). For future reference, we collectively call these approaches "regression shrinkage strategies".

We also evaluated predictive performance after backward elimination predictor selection.[40] This procedure starts by estimating a model with all P candidate predictor variables and considering the p-values associated with the predictor effects. For some pre-specified threshold value, the variable with the highest p-value exceeding the threshold value is dropped. The model is then re-estimated without the omitted variable. This process is continued until all the p-values associated with the effects of the predictors in the model are below the threshold. In this paper, we consider two commonly used threshold p-values values for ML and Firth's regressions. We use conventional threshold $p=0.050$ and a more conservative threshold $p=0.157$. The latter is equivalent to the AIC criterion for selection of predictors. We collectively call the backwards elimination predictor selection approaches and Lasso (which performs predictor selection by means of shrinkage) "predictor selection strategies".

3.2 Design and procedure

We conducted a full factorial simulation study, examining six design factors. These six factors are: (1) EPV, ranging from 3 to 50; (2) events fraction ($\Pr(Y=1)$), ranging from 50% to 6%; (3) number of candidate predictors (P), ranging from 4 to 12; (4) model discrimination, defined by the area under the ROC curve (AUC[50]), ranging from 0.65 to 0.85; (5) distribution of the predictor variables, independent Bernoulli or multivariate normal with equal pairwise correlation ranging from 0 to 0.5; (6) type of predictor effect, ranging from equal predictor effects for all candidate predictors to half of the candidate predictors as noise variables. All factor levels are described in Table 1.

In total, 4032 unique simulation scenarios were investigated. For each of these scenarios, 5000 simulation runs were executed using the following steps:

(1) A development data set was generated satisfying the simulation conditions (Table 1). For each of $N=(\text{EPV}\times P)/\Pr(Y=1)$ hypothetical individuals, a predictor variable vector ($\boldsymbol{x}_i$) was drawn. For each individual, a binary outcome was generated as $y_i=\text{Bernoulli}(\pi_i)$ (i.e. the outcome was drawn conditional on the true risk for each individual, which depends on the true predictor effects and the individuals predictor values).
(2) Nine binary logistic prediction models with different regression shrinkage and predictor selection strategies were estimated on the development data generated at step 1. These approaches are described in Table 2.
(3) A large validation data set was generated with sample size, $N^*=5000/\Pr(Y=1)$ (i.e. data set with 5000 expected events, which is 25 times larger than the recommended minimum sample size for validation studies[51]), using the sampling approach of step 1.
(4) The performance of the prediction models developed in step 2 is evaluated on the validation data generated in step 3. The measures of performance are detailed in section 3.3.

Table 1. Design factorial simulation study (7 × 4 × 3 × 4 × 3 × 4).

Simulation factors		Factor levels	
1.	Events per variable (EPV)	3, 5, 10, 15, 20, 30, 50	
2.	Events fraction	1/2, 1/4, 1/8, 1/16	
3.	Number of candidate predictors (*P*)	4, 8, 12	
4.	Model discrimination (AUC)	.65,.75,.85	
5.	Distribution of predictor variables	B(0.5):	Independent Bernoulli with success probability.5.
		MVN(0.0):	Normal (means = 0, variances = 1, covariances = 0.0)
		MVN(0.3):	Normal (means = 0, variances = 1, covariances = 0.3)
		MVN(0.5):	Normal (means = 0, variances = 1, covariances = 0.5)
6.	Predictor effects	Equal effect:	$\beta_1 = \cdots = \beta_P$
		1 strong:	$3\beta_1 = \beta_2 = \cdots = \beta_P$
		1 noise:	$\beta_1 = 0, \beta_2 = \cdots = \beta_P$
		1/2 noise:	$\beta_1 = \cdots = \beta_{P/2} = 0, \beta_{P/2+1} = \cdots = \beta_P$

Table 2. Prediction models: parameter shrinkage and variable selection strategies.

Model	Parameter shrinkage	Variable selection	Abbreviation
Maximum likelihood (full model)	No	No	ML
Maximum likelihood (backward 1)	No	Yes, $p < 0.050$	ML_p
Maximum likelihood (backward 2)	No	Yes, $p < 0.157$	ML_{AIC}
Heuristic shrinkage	Yes	No	HS
Firth's penalized likelihood (full model)	Yes	No	Firth
Firth's penalized likelihood (backward 1)	Yes	Yes, $p < 0.050$	Firth_p
Firth's penalized likelihood (backward 2)	Yes	Yes, $p < 0.157$	Firth_{AIC}
Ridge penalized likelihood	Yes	No	Ridge
Lasso penalized likelihood	Yes	Yes	Lasso

More details about the development of the simulation scenarios appear in Web Appendix A.

3.3 Simulation outcomes

3.3.1 Predictive performance metrics

Model discrimination was evaluated by the average (taken over all validation simulation samples) loss in the area under the ROC-curve (ΔAUC). ΔAUC was defined by the average difference between the AUCs estimated on the generated data and the AUC of the data generating model (the AUC defined by simulation factor number 5, Table 1). ΔAUCs were expected to be negative, with higher values (closer to zero) indicating better discriminative performance.

Model calibration performance was evaluated by the median of calibration slopes (CS) and average calibration in the large (CIL). CS values closer to 1 and CIL closer to 0 indicate better performance. CS was estimated using standard procedures.[52–54] Due to the expected skewness of slope distributions for smaller sized development data sets, medians rather than means and interquartile ranges rather than standard deviations were calculated. CS < 1 indicates model overfitting, CS > 1 indicates underfitting. CIL was calculated by average differences between the generated events fraction $\left(\text{i.e.} \frac{\sum_i y_i}{N}\right)$ and average estimated probabilities $\left(\text{i.e.} \frac{\sum_i \hat{\pi}_i}{N}\right)$. Values of CIL < 0 indicates systematic underestimation of estimated probabilities, CIL > 0 indicates systematic overestimation of estimated probabilities.

The prediction error was evaluated by the average of Brier scores[55] (Brier), the square root of the mean squared prediction error (rMPSE) and mean absolute prediction error (MAPE). The rMPSE and MAPE are based on the distance between the estimated probabilities ($\hat{\pi}_i$) and the true probabilities (π_i, which can be calculated under the data generating model using the generated predictor variable vector ($\boldsymbol{x}_i$)), by the square root of the average squared distance and the absolute distance, respectively. Lower values for Brier, rMSPE and MAPE indicate better performance.

3.3.2 Metamodels

Variation in simulation results across simulation conditions was studied by using metamodels.[27,28] The metamodels were used to quantify the relative impact of the various development data characteristics (taken as covariates in the metamodels) on a particular predictive performance simulation outcome (the outcome variable in the metamodel).

We considered the following covariates in the metamodel: development sample size (N), events fraction ($\Pr(Y=1)$), number of candidate predictor (P), true area under the characteristic curve (AUC), binary predictor variables (Bin, coding: $0=\text{no}$, $1=\text{yes}$), predictor pairwise correlations (Cor), and noise variables (Noise, coding: $0=\text{no}$, $1=\text{yes}$). Metamodels were developed for the following outcomes: natural log transformed MSPE (=rMPSE^2), natural log transformed MAPE, natural log transformed Brier, ΔAUC ($\times 100$ for notational convenience) and CS. These models were developed separately for each of the shrinkage and predictor selection strategies.

To facilitate interpretation, three separate metamodels were considered: i) a full model with all metamodel covariates, ii) a simplified model with only the development data size, events fraction and the number of candidate predictors, and for comparison: iii) a model with only development data EPV as a covariate. Metamodel ii was conceptualized before the start of the simulation study based on the notion that it would incorporate the same type of information as needed for estimating EPV *before* data collection, that is information available at the design phase of a prediction model development study (i.e. before the actual number of events are known).

The metamodels were estimated using linear regression with a Ridge penalty (i.e. Gaussian Ridge regression) specifying only linear main effects of the metamodel covariates. While more complex models (e.g. for interactions and non-linear effects) are possible, we found that linear main effects to be sufficient for constructing the metamodels. The Ridge metamodel tuning parameter was chosen based on 10-fold cross-validation that minimized mean squared error.

3.4 Software and estimation error handling

All simulations were performed in R (version 3.2.2)[56] executed on a central high-performance computing facility running on a CentOS Linux operating system. We used the CRAN packages: `GLMnet`[47] for estimating the Ridge and Lasso regression models (version 2.0-5, with an expanded grid of tuning parameters of 100 additional λ values that were smaller than the lowest value of the default), package `logistf` (version 1.21) for estimating ML, HS and Firth's model and to perform backward selection and package `MASS` (version 7.3-45)[57] for generating predictor data. Estimation errors were closely monitored (details in Web Appendix B). A summary of the estimation errors and their handling is given in Table 3. The Web Appendix also presents detailed simulation results focusing only on the ML model (Web Appendix C) and the relative rankings of the various model strategies with respect to the observed predictive performance (Web Appendix D).

4 Results

4.1 Predictive performance by relative size of development data

Figure 1 shows the average predictive performance of the various prediction models as a function of EPV and the events fraction. The impact of EPV and events fraction was consistent across the prediction models. There was improved predictive performance (i.e. reduction in average value for rMSPE and MAPE; ΔAUC closer to zero) when EPV increased (while keeping events fraction constant), and when the events fraction decreased (while keeping EPV constant). Differences between events fraction conditions decreased when EPV increased. Brier consistently improved (i.e. reduction in average value) with decreasing events fractions across prediction models, but showed little association with EPV beyond an EPV of 20.

Close to perfect average values (a value of 0) were observed for CIL for all models across all studied conditions (Figure 1), except for the Firth regressions with and without predictor selection (Figure 1).

Table 3. Simulation estimation errors and consequences.

	No. (%)	Consequences
Development datasets generated	20,160,000 (100%)	
Simulation conditions	4,032 (100%)	
Separation detected	90,846 (0.45%)	The separated cases are left in (to avoid selective missing data).
Degenerate distributions		
<3 events or <3 non-events generated	211 (0.001%)	Data sets are treated as missing data sets.
<8 events or <8 non-events generated	68,048 (0.34%)	Leave-one-out cross-validation is used for estimating Lasso and Ridge tuning parameters.
Degenerate predictor variable generated	0 (0%)	
Heuristic shrinkage factor inestimable	2,470,118 (12.25%)	For HS, results are replaced by ML results.
Degenerated linear predictor (no variables selected)		
ML_p	650,133 (3.22%)	
ML_{AIC}	179,638 (0.89%)	
Firth_p	718,194 (3.56%)	
Firth_{AIC}	204,617 (1.01%)	
Lasso	744,575 (3.69%)	

This miscalibration-in-the-large occurred in lower EPV settings, and did not occur in the conditions where the events fraction was 1/2. CS improved (i.e. average values closer to 1) with increasing EPV and decreasing events fraction for all models. On average, all models except the Ridge regression showed signs of overfitting (CS values below 1). The Ridge regression consistently showed signs of underfitting (CS values above 1). For all models, improved CS values were observed when EPV increased (while keeping the events fraction constant) and the events fraction decreased (while keeping EPV constant).

4.1.1 Performance of regression shrinkage strategies by relative size of development data

Unsurprisingly, the impact of shrinkage lessened with increasing EPV as depicted in Figure 2. The active regression shrinkage strategies (Ridge, Lasso, HS, Firth) showed lower median rMPSE and MAPE values than the non-shrunken ML regression at EPV = 5 and EPV = 10. In those settings, Ridge, Lasso and HS regression showed more variability between simulation scenarios than Firth and ML. For simulation scenarios at EPV = 50, the differences between shrinkage strategies were smaller.

In Figure 2, Brier and CIL outcomes showed little variation between shrinkage strategies. Notice that for this figure the events fraction was kept constant at 1/2, miscalibration-in-the-large was therefore not observed for the Firth regression. Poor CIL and rMPSE for the HS model was observed in some conditions with a high rate of separation (results not shown). Only the Ridge regression showed superior performance on the outcome ΔAUC, with little differences between HS, Firth and ML, and slightly less favorable and more variable performance of the Lasso regression at EPV = 5 and EPV = 10. The Lasso regression yielded CS closest to optimal (value of 1).

4.1.2 Performance of predictor selection strategies by relative size of development data

Backwards elimination (ML_p, ML_{AIC}, Firth_p and Firth_{AIC}) produced higher median rMPSE and MAPE than ML and Firth regressions that did not perform predictor selection (Figure 2). Median rMPSE and MAPE were more favorable for ML_{AIC} and Firth_{AIC} than ML_p and Firth_p. Backwards elimination also showed more variable MAPE and rMPSE values across the different simulation scenarios. The patterns were noticeable for the EPV = 5 and EPV = 10 conditions but did not completely disappear even at EPV = 50. Lasso regression had lower MAPE and rMPSE than the backwards elimination strategies and less variable results between conditions for the whole considered range of EPV.

Brier and CIL showed little variation between predictor selection strategies (Figure 2). For the predictor selection strategies, median ΔAUC were least favorable and more variable for Firth_p and ML_p, followed by ML_{AIC} and Firth_{AIC}, followed by Lasso. Lasso also yielded closer to optimal CS, with little differences

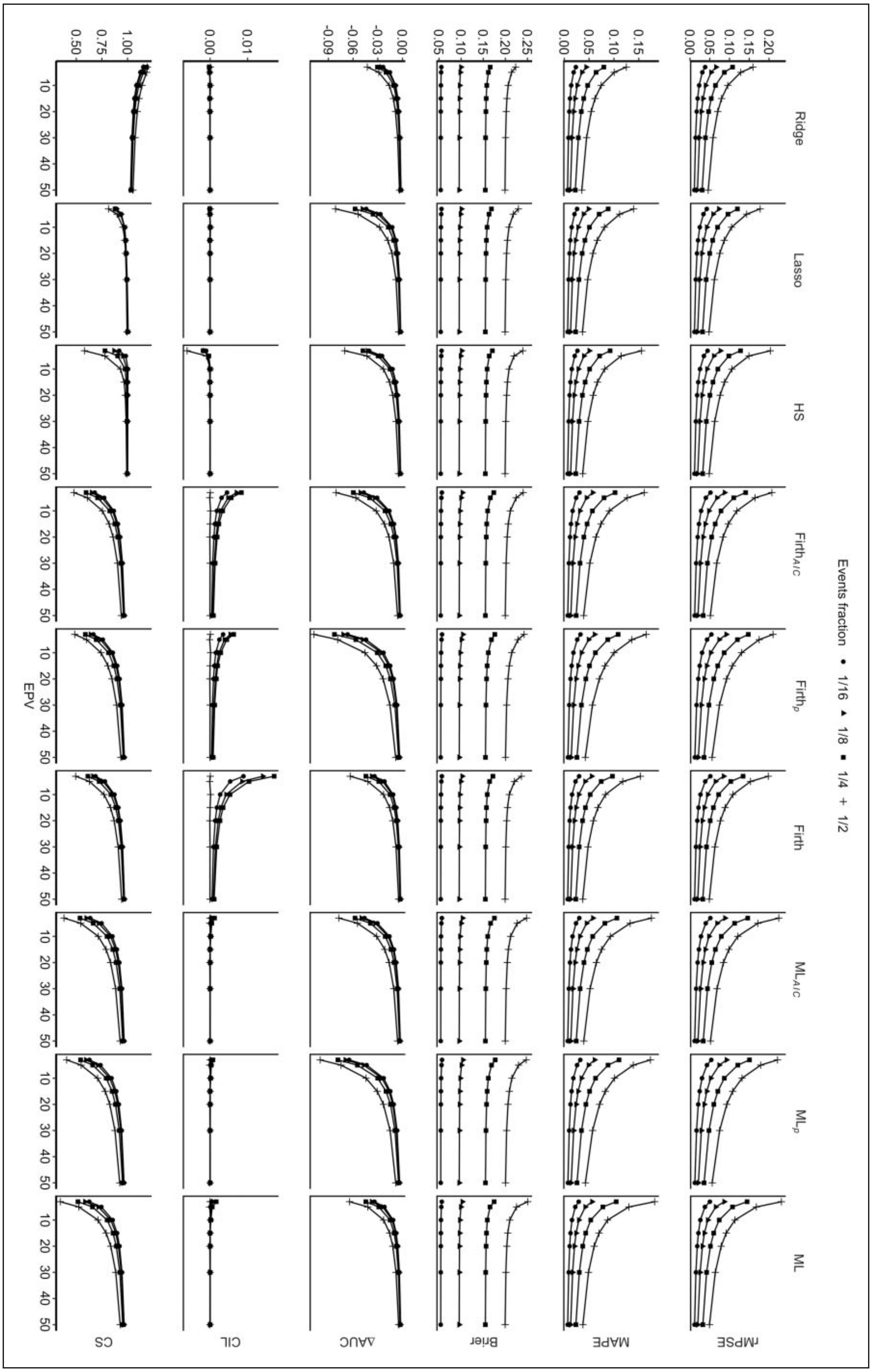

Figure 1. Marginal out-of-sample predictive performance.

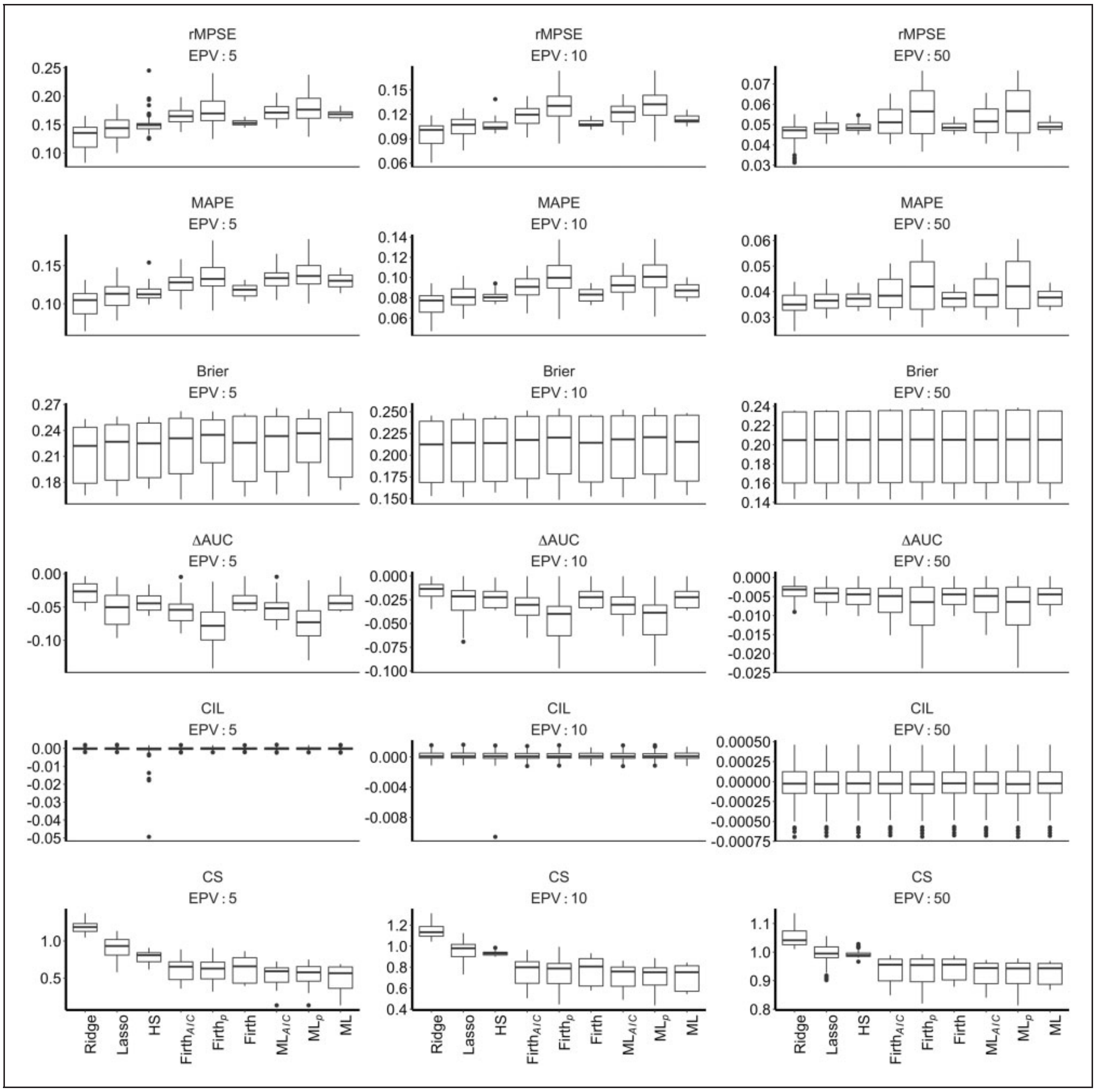

Figure 2. Boxplot distribution of out-of-sample predictive performance outcomes (restricted to conditions with events fraction = 1/2).

observed between the backwards elimination strategies. These patterns were observed consistently across the considered EPV range.

4.2 Predictive performance by other development data characteristics

Figure 3 describes the average performances of prediction models. We left out Brier (only noticeable changes occurred when varying the AUC of the data generating mechanism) and CIL (close to optimal for all but Firth regressions) for this presentation.

Lower AUC of the data generating mechanism was associated with poorer CS and ΔAUC outcomes. In conditions with AUC = 0.65, Ridge regression was superior in terms of rMPSE, MAPE and ΔAUC, while HS was superior in terms of CS. We also observed improved predictive performance as the number of predictors increased. This is partly due to a doubling of the development data size in our simulations when going from 4 to 8 predictors and three-fold increase in sample size when going from 4 to 12 predictors, a direct consequence of EPV as one of the chosen simulation factors.

With respect to the individual effects of the predictor variables (Figure 3), the average predictive performance of the variable selection strategies was best in conditions with one strong predictor. Effects of noise variables on the

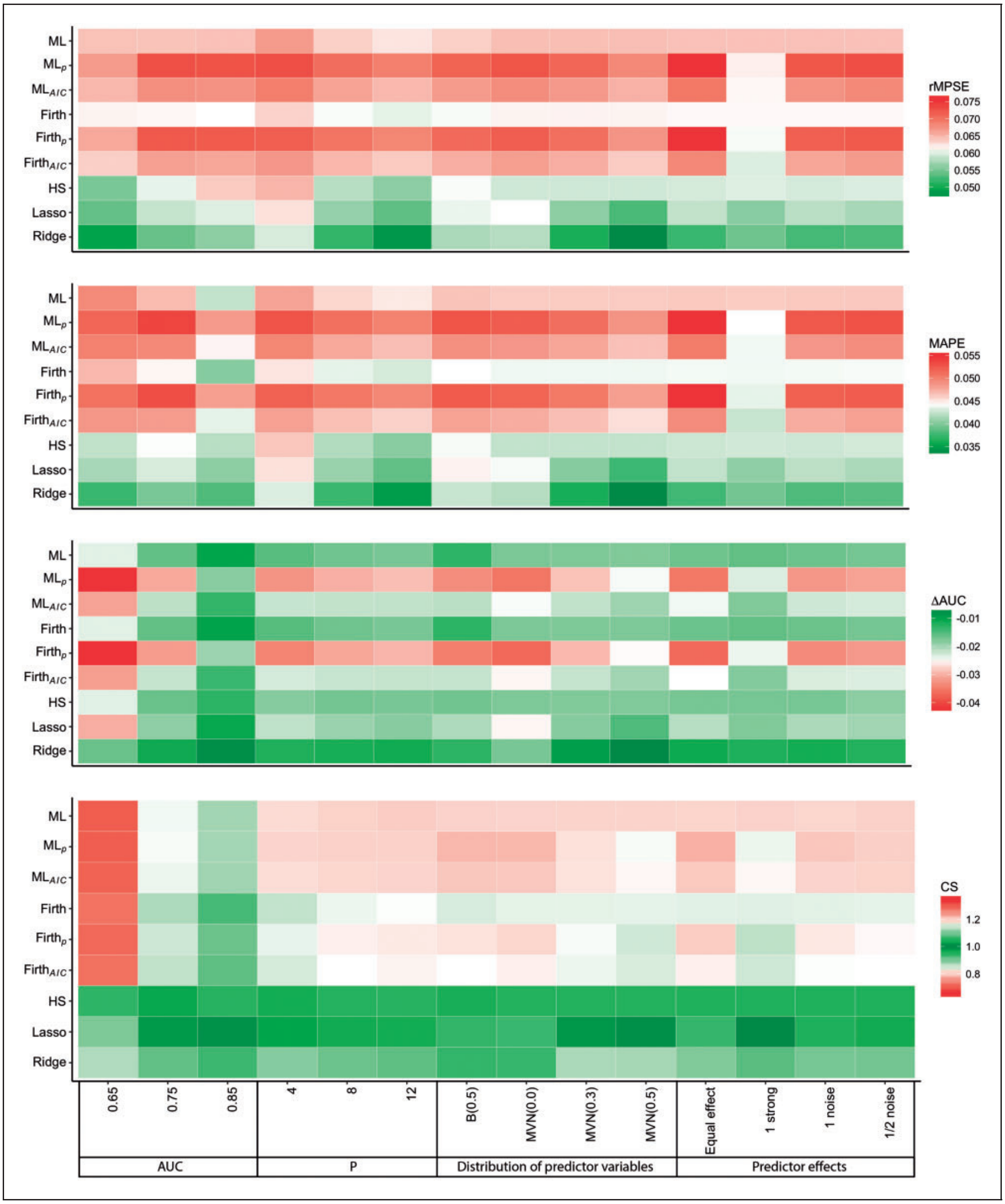

Figure 3. Average relative out-of-sample performances of modeling strategies per simulation factor level.

performances were negligible. Higher pairwise correlations between the predictors improved rMPSE, MAPE and ΔAUC for Ridge and Lasso and CS for Lasso. Higher correlations also increased the signs of underfitting of the Ridge regression ($CS > 1$).

4.3 Metamodels results

Table 4 presents the fitted results of the metamodels (linear regressions subject to a Ridge penalty).

The metamodels showed similar results for the outcomes natural log transformed MSPE (ln(MSPE)) and MAPE (ln(MAPE)) outcomes (Tables 4 and 5). For the metamodels that included all eight covariates as linear

Table 4. Results of simulation meta models: Outcome: ln(MSPE).

			Natural log transformed					Original scale			
Meta model		Int	EPV	N	Events fraction	P	AUC	Cor	Bin	Noise	R^2
Full	ML	−0.55	.	−1.06	0.36	0.94	0.40	0.00	0.05	0.00	0.993
Simplified	ML	−0.59	.	−1.06	0.36	0.94	.	.	.	.	0.992
EPV only	ML	−3.29	−1.06	.	.	.	.	.	.	.	0.432
Full	Firth	−0.84	.	−1.03	0.33	0.93	0.31	0.00	0.04	0.00	0.993
Simplified	Firth	−0.86	.	−1.03	0.33	0.93	.	.	.	.	0.992
EPV only	Firth	−3.42	−1.03	.	.	.	.	.	.	.	0.438
Full	HS	−0.39	.	−0.97	0.44	0.74	1.17	0.00	−0.01	0.00	0.985
Simplified	HS	−0.75	.	−0.97	0.44	0.74	.	.	.	.	0.977
EPV only	HS	−3.64	−0.97	.	.	.	.	.	.	.	0.385
Full	Lasso	−0.59	.	−0.93	0.46	0.68	0.97	−0.48	0.04	0.03	0.983
Simplified	Lasso	−0.86	.	−0.93	0.46	0.68	.	.	.	.	0.973
EPV only	Lasso	−3.78	−0.93	.	.	.	.	.	.	.	0.371
Full	Ridge	−0.39	.	−0.88	0.50	0.49	1.33	−0.85	0.03	−0.02	0.979
Simplified	Ridge	−0.93	.	−0.88	0.50	0.49	.	.	.	.	0.952
EPV only	Ridge	−4.08	−0.88	.	.	.	.	.	.	.	0.337
Full	ML$_p$	−0.85	.	−1.02	0.40	0.95	0.34	0.03	0.07	0.17	0.943
Simplified	ML$_p$	−0.57	.	−1.03	0.40	0.96	.	.	.	.	0.939
EPV only	ML$_p$	−3.18	−1.03	.	.	.	.	.	.	.	0.393
Full	ML$_{AIC}$	−0.74	.	−1.05	0.38	0.95	0.35	0.00	0.06	0.10	0.977
Simplified	ML$_{AIC}$	−0.59	.	−1.05	0.38	0.95	.	.	.	.	0.975
EPV only	ML$_{AIC}$	−3.25	−1.05	.	.	.	.	.	.	.	0.417
Full	Firth$_p$	−0.94	.	−1.01	0.39	0.95	0.34	0.02	0.07	0.17	0.939
Simplified	Firth$_p$	−0.66	.	−1.01	0.39	0.95	.	.	.	.	0.935
EPV only	Firth$_p$	−3.22	−1.01	.	.	.	.	.	.	.	0.392
Full	Firth$_{AIC}$	−0.90	.	−1.03	0.37	0.95	0.32	0.00	0.06	0.10	0.975
Simplified	Firth$_{AIC}$	−0.74	.	−1.03	0.37	0.95	.	.	.	.	0.973
EPV only	Firth$_{AIC}$	−3.32	−1.03	.	.	.	.	.	.	.	0.418

Full: metamodel with all eight meta-model covariates; Simplified: model with covariates N, events fraction and P, EPV only: meta model with EPV as a covariate. Int: Intercept; EPV: Events per variable; N: Sample size; P: number of candidate predictors; AUC: Area under the ROC-curve; Cor: Predictor pairwise correlations.

main effects, the percentage of explained variance (R^2) was 99.3% for the outcome ln(rMSPE) and $R^2 = 99.6\%$ for ln(MAPE). Using the simplified metamodel with three covariates, the R^2 dropped to between 93.5% and 99.2% indicating that these factors – N, events fraction and P – explained a sizable amount of the variance between simulation conditions. R^2 was similar for ML and Firth regression, but lower for Ridge, Lasso, HS and after backwards elimination. Using only EPV as covariate in the metamodel yielded R^2 between 28.5% and 43.2% for the ln(MSPE) and ln(MAPE) outcomes.

As expected, MSPE and MAPE were negatively related to N and positively related to P. The positive relation between the events fraction (events fraction $\leq 1/2$) and MPSE/MAPE can be explained by a shift of the average of estimated probabilities π_i towards zero as the event rate decreases (assuming the model is appropriately calibrated). These lower probabilities have lower expected variance, considering that the variance of a Bernoulli is asymptotically $\pi(1 - \pi)$. Similar findings were observed for the outcome ln(Brier) (Table 6). There was a strong relation between the simplified model covariates and ln(Brier) ($R^2 > 92.0\%$). Little variation between the fitted metamodel coefficients and R^2 was observed for the different regression models. For all models, N was negatively related to ln(Brier), while the events fraction and P were positively related to ln(Brier). In contrast, EPV had only weak relation to ln(Brier) with $R^2 < 1.0\%$.

The outcomes ΔAUC (Table 7) and CS (Table 8) were less well predicted by the eight covariate metamodel and varied considerably between the prediction models (R^2 between 84.8% and 49.6%). Similarly, for the simplified metamodel with three covariates, R^2 was between 70.0% and 19.0%. R^2 dropped even further for metamodels with EPV as the only predictor, R^2 was between 63.3% and 18.0%. Largest R^2 was observed for the ML regression. Similar to other metamodels, ΔAUC and CS improved with higher N and decreasing P. The direction of effect of

Table 5. Results of simulation meta models: Outcome: ln(MAPE).

			Natural log transformed					Original scale			
Meta model		Int	EPV	*N*	Events fraction	*P*	AUC	Cor	Bin	Noise	R^2
Full	ML	−0.60	.	−0.53	0.31	0.48	−0.50	0.00	−0.01	0.00	0.996
Simplified	ML	−0.48	.	−0.53	0.31	0.48	.	.	.	.	0.992
EPV only	ML	−2.03	−0.53	.	.	.	.	.	.	.	0.355
Full	Firth	−0.74	.	−0.51	0.29	0.47	−0.51	0.00	−0.01	0.00	0.996
Simplified	Firth	−0.61	.	−0.51	0.29	0.47	.	.	.	.	0.991
EPV only	Firth	−2.10	−0.51	.	.	.	.	.	.	.	0.357
Full	HS	−0.55	.	−0.49	0.33	0.39	−0.15	0.00	−0.03	0.00	0.991
Simplified	HS	−0.56	.	−0.49	0.33	0.39	.	.	.	.	0.991
EPV only	HS	−2.19	−0.49	.	.	.	.	.	.	.	0.326
Full	Lasso	−0.59	.	−0.48	0.34	0.35	−0.19	−0.24	−0.01	0.01	0.989
Simplified	Lasso	−0.59	.	−0.48	0.34	0.35	.	.	.	.	0.983
EPV only	Lasso	−2.24	−0.48	.	.	.	.	.	.	.	0.314
Full	Ridge	−0.48	.	−0.45	0.36	0.26	0.03	−0.43	−0.02	−0.01	0.986
Simplified	Ridge	−0.61	.	−0.45	0.36	0.26	.	.	.	.	0.970
EPV only	Ridge	−2.39	−0.45	.	.	.	.	.	.	.	0.285
Full	ML_p	−0.75	.	−0.52	0.31	0.49	−0.58	0.03	−0.01	0.09	0.951
Simplified	ML_p	−0.45	.	−0.52	0.31	0.49	.	.	.	.	0.942
EPV only	ML_p	−1.95	−0.52	.	.	.	.	.	.	.	0.334
Full	ML_{AIC}	−0.70	.	−0.53	0.31	0.49	−0.55	0.01	−0.01	0.06	0.982
Simplified	ML_{AIC}	−0.48	.	−0.53	0.31	0.49	.	.	.	.	0.975
EPV only	ML_{AIC}	−2.00	−0.53	.	.	.	.	.	.	.	0.348
Full	Firth_p	−0.79	.	−0.52	0.30	0.50	−0.56	0.02	−0.01	0.09	0.947
Simplified	Firth_p	−0.49	.	−0.52	0.30	0.50	.	.	.	.	0.938
EPV only	Firth_p	−1.96	−0.52	.	.	.	.	.	.	.	0.335
Full	Firth_{AIC}	−0.78	.	−0.52	0.30	0.49	−0.55	0.01	−0.01	0.06	0.979
Simplified	Firth_{AIC}	−0.55	.	−0.52	0.30	0.49	.	.	.	.	0.973
EPV only	Firth_{AIC}	−2.03	−0.52	.	.	.	.	.	.	.	0.348

Full: metamodel with all eight meta-model covariates; Simplified: model with covariates *N*, events fraction and *P*, EPV only: meta model with EPV as a covariate. Int: Intercept; EPV: Events per variable; *N*: Sample size; *P*: number of candidate predictors; AUC: Area under the ROC-curve; Cor: Predictor pairwise correlations.

the events fraction is in the opposite direction as compared to MSPE, MAPE and brier, showing decreased performance with decreasing event fraction.

5 Discussion

This paper has investigated the impact of EPV and other development data characteristics in relation to modelling strategies on the (out-of-sample) predictive performance of prediction models developed with logistic regression. We showed that the EPV fails to have a strong relation with metrics of predictive performance across modelling strategies. Given our findings, it is clear that EPV is not an appropriate sample size criterion for binary prediction model development studies. Below we discuss our simulation results, followed by a discussion of the implications for sample size determination for prediction model development. A new strategy for such sample size consideration is proposed.

5.1 Simulation findings

Our study confirms previous findings that predictive performance can be poor for prediction models developed using conventional maximum likelihood binary logistic regression in data with a small number of subjects relative to the number of predictors. As expected, predictive performance generally improved when regression shrinkage strategies were applied, while backwards elimination predictor selection strategies generally worsened the predictive accuracy of the prediction model. These tendencies were observed consistently for discrimination

Table 6. Results of simulation meta models: Outcome: ln(Brier).

			Natural log transformed					Original scale			
Meta model		Int	EPV	N	Events fraction	P	AUC	Cor	Bin	Noise	R^2
Full	ML	−1.23	.	−0.04	0.62	0.04	−1.02	0.00	0.01	0.00	0.969
Simplified	ML	−0.91	.	−0.04	0.62	0.04	.	.	.	.	0.925
EPV only	ML	−2.06	−0.04	.	.	.	.	.	.	.	0.005
Full	Firth	−1.27	.	−0.03	0.62	0.03	−1.02	0.00	0.01	0.00	0.969
Simplified	Firth	−0.95	.	−0.03	0.62	0.03	.	.	.	.	0.923
EPV only	Firth	−2.08	−0.03	.	.	.	.	.	.	.	0.003
Full	HS	−1.23	.	−0.03	0.62	0.02	−0.98	0.00	0.00	0.00	0.969
Simplified	HS	−0.93	.	−0.03	0.62	0.02	.	.	.	.	0.927
EPV only	HS	−2.08	−0.03	.	.	.	.	.	.	.	0.003
Full	Lasso	−1.27	.	−0.03	0.62	0.02	−1.00	−0.02	0.01	0.00	0.969
Simplified	Lasso	−0.96	.	−0.03	0.62	0.02	.	.	.	.	0.925
EPV only	Lasso	−2.10	−0.03	.	.	.	.	.	.	.	0.002
Full	Ridge	−1.29	.	−0.02	0.62	0.01	−1.00	−0.02	0.01	0.00	0.968
Simplified	Ridge	−0.98	.	−0.02	0.62	0.01	.	.	.	.	0.924
EPV only	Ridge	−2.12	−0.02	.	.	.	.	.	.	.	0.002
Full	ML_p	−1.19	.	−0.04	0.62	0.04	−0.96	−0.02	0.01	0.01	0.969
Simplified	ML_p	−0.89	.	−0.04	0.62	0.04	.	.	.	.	0.929
EPV only	ML_p	−2.04	−0.04	.	.	.	.	.	.	.	0.006
Full	ML_{AIC}	−1.22	.	−0.04	0.62	0.04	−0.99	−0.01	0.01	0.00	0.969
Simplified	ML_{AIC}	−0.91	.	−0.04	0.62	0.04	.	.	.	.	0.927
EPV only	ML_{AIC}	−2.05	−0.04	.	.	.	.	.	.	.	0.005
Full	Firth_p	−1.20	.	−0.04	0.62	0.04	−0.96	−0.02	0.01	0.01	0.968
Simplified	Firth_p	−0.90	.	−0.04	0.62	0.04	.	.	.	.	0.929
EPV only	Firth_p	−2.05	−0.04	.	.	.	.	.	.	.	0.005
Full	Firth_{AIC}	−1.24	.	−0.04	0.62	0.04	−1.00	−0.01	0.01	0.00	0.969
Simplified	Firth_{AIC}	−0.93	.	−0.04	0.62	0.04	.	.	.	.	0.926
EPV only	Firth_{AIC}	−2.06	−0.04	.	.	.	.	.	.	.	0.004

Full: metamodel with all eight meta-model covariates; Simplified: model with covariates *N*, events fraction and *P*, EPV only: meta model with EPV as a covariate. Int: Intercept; EPV: Events per variable; *N*: Sample size; *P*: number of candidate predictors; AUC: Area under the ROC-curve; Cor: Predictor pairwise correlations.

(ΔAUC), calibration slopes (CS) and prediction error (rMPSE, MAPE, and Brier) outcomes. Calibration in the large was near ideal for all models in all simulation settings, except for Firth regression that showed upward biased estimation of probability towards $\pi = 0.5$. Some more recent refinements to the Firth's correction have shown promising results in circumventing the issues with calibration in the large.[21,48,58,59]

With larger sample sizes, the benefits (in terms of predictive performance) of the regression shrinkage strategies gradually declined, but predictive performance after shrinkage remained slightly superior or equivalent to ML regression even for larger sample sizes. Between the regression shrinkage strategies, the Ridge regression showed best discrimination (lowest average ΔAUC) and lowest prediction error (lowest average rMSPE, MAPE and Brier) performance when compared to Firth, Lasso and HS. Median CS of the HS and Lasso regression were closer to optimal than the Ridge regression, the latter showing signs of underfitting. The observed tendency to underfitting of Ridge regression is consistent with other recent simulation studies.[16,48] In smaller samples, backwards elimination with conventional $p = 0.050$ and AIC criteria, generally performed worse than an equivalent regression without predictor selection or Lasso, even when only half of the predictor variables were randomly associated to the outcome. For conditions with EPV as large as 50, backwards elimination was found to yield higher rMSPE and MAPE than the equivalent model with all variables left in. Between the backward elimination criteria, the more conservative AIC criterion was found to produce better average predictive performance than $p = 0.050$, in accordance with earlier work.[9,11]

The metamodels fitted on the simulation results revealed that between simulation variation of (r)MPSE, MAPE and Brier could largely be explained by a linear model with three covariates: sample size, the events fraction and the number of candidate predictors. The joint effect of these three covariates on prediction error tended to become

Table 7. Results of simulation meta models: Outcome: ΔAUC × 100.

			Natural log transformed					Original scale			
Meta model		Int	EPV	N	Events fraction	P	AUC	Cor	Bin	Noise	R^2
Full	ML	−3.63	.	1.47	0.92	−1.66	5.02	−0.03	−0.46	−0.06	0.821
Simplified	ML	−6.01	.	1.47	0.92	−1.66	.	.	.	.	0.700
EPV only	ML	−5.44	1.47	.	.	.	.	.	.	.	0.633
Full	Firth	−3.56	.	1.46	0.92	−1.66	5.05	−0.03	−0.47	−0.06	0.822
Simplified	Firth	−5.97	.	1.46	0.92	−1.66	.	.	.	.	0.698
EPV only	Firth	−5.42	1.46	.	.	.	.	.	.	.	0.632
Full	HS	−5.60	.	1.63	1.11	−1.53	4.05	−0.03	−0.08	−0.07	0.665
Simplified	HS	−7.05	.	1.63	1.11	−1.53	.	.	.	.	0.614
EPV only	HS	−5.96	1.63	.	.	.	.	.	.	.	0.571
Full	Lasso	−6.11	.	1.93	1.24	−1.63	7.30	2.11	−0.45	−0.08	0.713
Simplified	Lasso	−8.73	.	1.93	1.24	−1.63	.	.	.	.	0.580
EPV only	Lasso	−6.95	1.93	.	.	.	.	.	.	.	0.528
Full	Ridge	−3.14	.	0.98	0.62	−0.91	3.38	2.18	−0.42	−0.03	0.684
Simplified	Ridge	−4.47	.	0.98	0.62	−0.91	.	.	.	.	0.515
EPV only	Ridge	−3.70	0.98	.	.	.	.	.	.	.	0.468
Full	ML$_p$	−9.03	.	2.62	1.75	−2.29	8.89	2.18	−0.23	−0.23	0.764
Simplified	ML$_p$	−11.94	.	2.62	1.75	−2.29	.	.	.	.	0.645
EPV only	ML$_p$	−9.80	2.62	.	.	.	.	.	.	.	0.597
Full	ML$_{AIC}$	−5.79	.	1.91	1.25	−1.87	6.64	0.99	−0.33	−0.17	0.797
Simplified	ML$_{AIC}$	−8.37	.	1.91	1.25	−1.87	.	.	.	.	0.680
EPV only	ML$_{AIC}$	−7.13	1.92	.	.	.	.	.	.	.	0.626
Full	Firth$_p$	−9.92	.	2.75	1.81	−2.33	8.72	2.36	−0.22	−0.22	0.751
Simplified	Firth$_p$	−12.72	.	2.75	1.81	−2.33	.	.	.	.	0.646
EPV only	Firth$_p$	−10.25	2.75	.	.	.	.	.	.	.	0.592
Full	Firth$_{AIC}$	−6.18	.	1.98	1.28	−1.89	6.61	1.10	−0.33	−0.17	0.785
Simplified	Firth$_{AIC}$	−8.73	.	1.98	1.28	−1.89	.	.	.	.	0.677
EPV only	Firth$_{AIC}$	−7.34	1.98	.	.	.	.	.	.	.	0.621

Full: metamodel with all eight meta-model covariates; Simplified: model with covariates *N*, events fraction and *P*, EPV only: meta model with EPV as a covariate. Int: Intercept; EPV: Events per variable; *N*: Sample size; *P*: number of candidate predictors; AUC: Area under the ROC-curve; Cor: Predictor pairwise correlations.

slightly weaker when regression shrinkage or variable selection strategies were applied. ΔAUC and CS were found to be more unpredictable by the metamodel regression. ΔAUC and CS were found to be particularly sensitive to the prediction model development strategy employed (e.g. whether regression shrinkage or predictor selection was used) and, importantly, dependent on the AUC of the data generating mechanism.

Some limitations apply to our study. The broad setup of our simulations, with over 4000 unique scenarios, does allow for a generalization of the findings to a large variety of prediction modeling settings. However, as with any simulation study, the number of investigated scenarios was finite and extrapolation of our findings far beyond the investigated regions is not advised. A total of nine prediction modeling strategies were investigated. In practice, we expect that other approaches to regression shrinkage and predictor selection than we considered may sometimes be preferable (e.g. Elastic Net,[49] non-negative Garrotte,[60] random forest[61]). Finding optimal strategies for developing clinical prediction models in small or sparse data was not the main objective of the current study but is a worthwhile topic for future research.

5.2 Implications for sample size considerations

There is general consensus on the importance of having data of adequately size when developing a prediction model.[2] However, consensus is lacking on the criteria to determine what size would count as adequate.

Our results showed that the recommended minimal EPV criteria for prediction model development, notably the EPV ≥ 10 rule,[34] falls short of providing appropriate sample size guidance. Earlier critiques on EPV as a sample

Table 8. Results of simulation meta models: Outcome: CS.

			Natural log transformed					Original scale			
Meta model		Int	EPV	N	Events fraction	P	AUC	Cor	Bin	Noise	R^2
Full	ML	0.50	.	0.15	0.09	−0.16	0.65	0.00	0.00	0.00	0.848
Simplified	ML	0.31	.	0.15	0.09	−0.16	.	.	.	.	0.689
EPV only	ML	0.40	0.15	.	.	.	.	.	.	.	0.616
Full	Firth	0.73	.	0.12	0.08	−0.15	0.77	0.00	−0.01	0.00	0.835
Simplified	Firth	0.50	.	0.12	0.08	−0.15	.	.	.	.	0.556
EPV only	Firth	0.52	0.12	.	.	.	.	.	.	.	0.505
Full	HS	0.73	.	0.07	0.02	−0.08	0.03	0.00	−0.01	0.00	0.496
Simplified	HS	0.71	.	0.07	0.02	−0.08	.	.	.	.	0.495
EPV only	HS	0.77	0.07	.	.	.	.	.	.	.	0.368
Full	Lasso	0.98	.	0.04	0.03	−0.05	0.43	0.12	0.00	−0.01	0.513
Simplified	Lasso	0.85	.	0.04	0.03	−0.05	.	.	.	.	0.190
EPV only	Lasso	0.85	0.04	.	.	.	.	.	.	.	0.180
Full	Ridge	1.19	.	−0.05	−0.03	0.03	−0.25	0.14	0.01	0.00	0.823
Simplified	Ridge	1.31	.	−0.05	−0.03	0.03	.	.	.	.	0.488
EPV only	Ridge	1.23	−0.05	.	.	.	.	.	.	.	0.418
Full	ML$_p$	0.51	.	0.15	0.09	−0.15	0.65	0.09	0.00	−0.01	0.832
Simplified	ML$_p$	0.33	.	0.15	0.09	−0.15	.	.	.	.	0.652
EPV only	ML$_p$	0.42	0.15	.	.	.	.	.	.	.	0.588
Full	ML$_{AIC}$	0.52	.	0.15	0.09	−0.15	0.63	0.06	0.00	−0.01	0.848
Simplified	ML$_{AIC}$	0.33	.	0.15	0.09	−0.15	.	.	.	.	0.682
EPV only	ML$_{AIC}$	0.42	0.15	.	.	.	.	.	.	.	0.611
Full	Firth$_p$	0.67	.	0.13	0.08	−0.15	0.74	0.09	−0.01	−0.01	0.826
Simplified	Firth$_p$	0.44	.	0.13	0.08	−0.15	.	.	.	.	0.575
EPV only	Firth$_p$	0.49	0.13	.	.	.	.	.	.	.	0.522
Full	Firth$_{AIC}$	0.69	.	0.13	0.08	−0.15	0.73	0.05	−0.01	−0.01	0.846
Simplified	Firth$_{AIC}$	0.46	.	0.13	0.08	−0.15	.	.	.	.	0.595
EPV only	Firth$_{AIC}$	0.50	0.13	.	.	.	.	.	.	.	0.537

Full: metamodel with all eight meta-model covariates; Simplified: model with covariates N, events fraction and P, EPV only: meta model with EPV as a covariate. Int: Intercept; EPV: Events per variable; N: Sample size; P: number of candidate predictors; AUC: Area under the ROC-curve; Cor: Predictor pairwise correlations.

size criterion has identified its weak theoretical and empirical underpinning,[17–20] and has shown that the EPV ≥ 10 rule can be too lenient[11,13] or too strict,[15,21] depending on the modelling approach taken. The current study also showed that EPV fails the minimal requirement of strong relation to (at least one aspect of) predictive performance. In itself EPV was found to have only a weak relation with outcomes of prediction error and a mediocre relation with calibration and discrimination. The EPV ≥ 10 rule also does not adequately account for changes in events fraction. The implied relation by the EPV ≥ 10 rule between required sample size (N) and the events fraction is described by the function: $N = 10 \times P/$ events fraction, where the events fraction $\leq 1/2$ (trivially recoding of Y can ensure the events fraction not exceeds $1/2$). This relation is depicted in Figure 4. The figure shows that the relation between the events fraction and required N is in the same direction but much steeper for EPV $= 10$ than the relation between the required sample size when keeping expected CS and ΔAUC constant. The relation between prediction error measures is in the opposite direction.

The search for new minimal sample size criteria inherently calls for abandoning EPV as the sole sample size criterion. Alternatives for sample size must have a predictable relationship with future predictive performance and be on a scale that is interpretable for users. It is our view that general single threshold values should be avoided. Instead, sample size determination should be based on threshold values on an interpretable scale that ensure predictive performance that is fit for purpose. What counts as fit for purpose varies from application to application (e.g. clinical prediction models for informing short-term high-risk treatment decisions may differ from the requirements for long-term low-risk decisions). It is the duty of the researcher to define what constitutes as fit for purpose in context and explain how the sample size was arrived at (see also: the TRIPOD statement[2,3]).

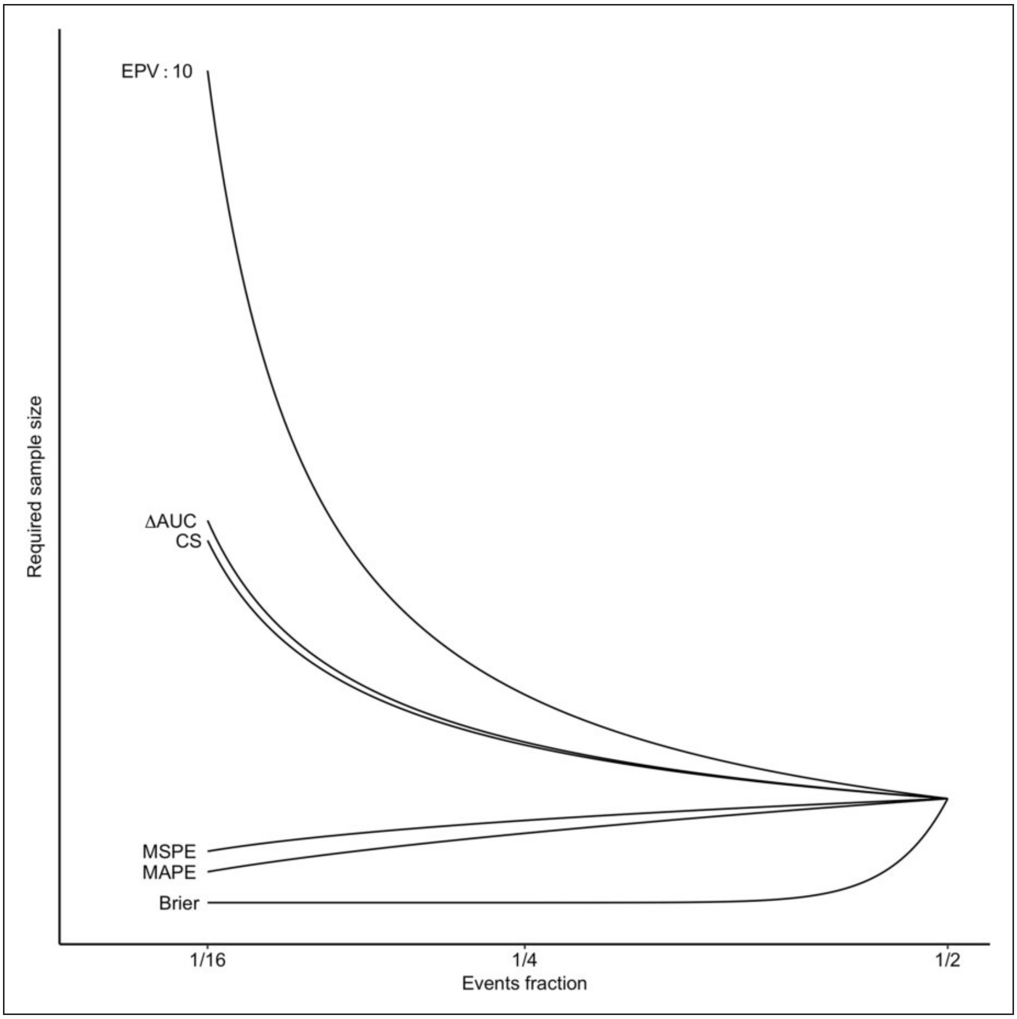

Figure 4. Relation required sample size and events fraction. Calculations based on metamodels with criterion values that were kept constant. For illustration purposes, the criterion values were chosen such that they would intersect at events fraction = 1/2.

5.3 New sample size criteria

Out-of-sample (r)MSPE and MAPE are natural metrics to determine sample size adequacy of prediction models, as they define the expected distance (squared or absolute) for new individuals between the estimated probabilities for new patients and their unobservable "true" values. Because clinical prediction models are primary used to estimate probabilities for new individuals,[3–5] rMSPE and MAPE have direct relevance when developing a prediction model.

The out-of-sample rMPSE and MAPE can be approximated via simulations as we have done in this paper. Our simulation code is available via GitHub (https://github.com/MvanSmeden/Beyond-EPV). Alternatively, rMSPE and MAPE may be approximated via the results of our metamodels (Table 4). For instance, at a sample size of $N = 400$, with $P = 8$ candidate predictors and an expected event fraction of 1/4, the predicted out-of-sample rMPSE is 0.065 when ML model (without variable selection) is applied and 0.053 for Ridge regression; MAPE is 0.045 for the ML model and 0.038 for the Ridge regression. Obviously, whether or not these expected "average" prediction errors on the probability scale are acceptable or not depends on the intended use of the prediction model (i.e. $N = 400$ may not be sufficient for accurate estimation of probability for high risk treatment decisions, even though for this example EPV = 20).

We warn readers that these out-of-sample performance predictions from the simulation metamodels have not been externally validated and that approximations may not work well far outside the range of investigated simulation settings. In particular, using these approximations for sample size calculations with very low events fractions may yield unacceptably poor discrimination and calibration performances (see Figure 4).

6 Conclusion

The currently recommended sample size criteria for developing prediction models, notably the EPV ≥ 10 rule-of-thumb, are insufficient to warrant appropriate sample size decisions. EPV criteria fail to take into account the intended use of the prediction model and have only a weak relation to out-of-sample predictive performance of the prediction model. Instead, sample size should be determined based on a meaningful out-of-sample predictive performance scale, such as the rMPSE and MAPE. The results of our study can be used to inform sample size considerations when developing a binary prediction model given the required predictive performance in new individuals.

Acknowledgement

We thank Prof Dr Richard D Riley for his constructive comments on an earlier version of this manuscript.

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 Netherlands Organisation for Scientific Research (project 918.10.615).

ORCID iD

Maarten van Smeden http://orcid.org/0000-0002-5529-1541
Gary S Collins http://orcid.org/0000-0002-2772-2316

References

1. Bouwmeester W, Zuithoff NP, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012; **9**: e1001221.
2. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; **162**: W1–W73.
3. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015; **162**: 55.
4. Altman DG and Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000; **19**: 453–473.
5. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012; **98**: 691–698.
6. Harrell FE, Lee KL, Califf RM, et al. Regression modelling strategies for improved prognostic prediction. *Stat Med* 1984; **3**: 143–152.
7. Harrell FE, Lee KL and Mark DB. Tutorial in biostatistics – multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; **15**: 361–387.
8. Harrell FE. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York, NY: Springer, 2001.
9. Steyerberg EW, Eijkemans MJC, Harrell FE, et al. Prognostic modeling with logistic regression analysis. *Med Decis Mak* 2001; **21**: 45–56.
10. Steyerberg EW. *Clinical prediction models*. New York, NY: Springer, 2009.
11. Steyerberg EW, Eijkemans MJC, Harrell FE, 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.
12. Steyerberg EW, Bleeker SE, Moll HA, et al. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *J Clin Epidemiol* 2003; **56**: 441–447.
13. Ambler G, Brady AR and Royston P. Simplifying a prognostic model: a simulation study based on clinical data. *Stat Med* 2002; **21**: 3803–3822.
14. Moons KGM, de Groot JAH, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med* 2014; **11**: e1001744.

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. 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.
17. Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* 2012; **98**: 683–690.
18. 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.
19. 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**: 163.
20. 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.
21. 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.
22. Demidenko E. Sample size determination for logistic regression revisited. *Stat Med* 2006; **26**: 3385–3397.
23. Cessie SL and Houwelingen JC. Ridge estimators in logistic regression. *Appl Stat* 1992; 191–201.
24. Tibshirani R. Regression shrinkage and selection via the Lasso. *J Royal Stat Soc Ser B (Stat Methodol)* 1996; **58**: 267–288.
25. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; **80**: 27–38.
26. Van Houwelingen JC and Le Cessie S. Predictive value of statistical models. *Stat Med* 1990; **9**: 1303–1325.
27. Harwell MR, Rubinstein EN, Hayes WS, et al. Summarizing Monte Carlo results in methodological research: the one-and two-factor fixed effects ANOVA cases. *J Educ Stat* 1992; **17**: 315–339.
28. Kleijnen JP and Sargent RG. A methodology for fitting and validating metamodels in simulation. *Eur J Operation Res* 2000; **120**: 14–29.
29. Agresti A. *Categorical data analysis*. Vol. 2, Hoboken, NJ: John Wiley & Sons, Inc., 2002.
30. James W and Stein C. Estimation with quadratic loss. In: *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. Berkeley, CA: University of California Press, 1961, pp.361–379.
31. Efron B and Morris C. Stein's paradox in statistics. *Scientific Am* 1977; **236**: 119–127.
32. Gart J and Zweifel J. On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika* 1967; **54**: 181–187.
33. Jewell N. Small-sample bias of point estimators of the odds ratio from matched sets. *Biometrics* 1984; **40**: 421–435.
34. 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.
35. Vittinghoff E and McCulloch CE. Relaxing the rule of ten events per variable in logistic and cox regression. *Am J Epidemiol* 2007; **165**: 710–718.
36. Nemes S, Jonasson J, Genell A, et al. Bias in odds ratios by logistic regression modelling and sample size. *BMC Med Res Methodol* 2009; **9**: 56.
37. Albert A and Anderson J. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* 1984; **71**: 1–10.
38. Heinze G and Schemper M. A solution to the problem of separation in logistic regression. *Stat Med* 2002; **21**: 2409–2419.
39. Hoerl AE and Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970; **12**: 55–67.
40. Mantel N. Why stepdown procedures in variable selection. *Technometrics* 1970; **12**: 621–625.
41. Altman DG and Andersen PK. Bootstrap investigation of the stability of a Cox regression model. *Stat Med* 1989; **8**: 771–83.
42. Sauerbrei W and Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. *Stat Med* 1992; **11**: 2093–2109.
43. Van Houwelingen JC and Le Cessie S. Predictive value of statistical models. *Stat Med* 1990; **9**: 1303–1325.
44. Pajouheshnia R, Pestman WR, Teerenstra S, et al. A computational approach to compare regression modelling strategies in prediction research. *BMC Med Res Methodol* 2016; **16**: 107.
45. Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Stat Med* 2006; **25**: 4216–4226.
46. Hastie T, Tibshirani R and Friedman J. *The elements of statistical learning*. New York, NY: Springer, 2009.
47. Friedman J, Hastie T and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Software* 2010; **33**: 1–22.
48. Rahman MS and Sultana M. Performance of Firth-and logF-type penalized methods in risk prediction for small or sparse binary data. *BMC Med Res Methodol* 2017; **17**: 33.
49. Zou H and Hastie T. Regularization and variable selection via the elastic net. *J Royal Stat Soc Ser B (Stat Methodol)* 2005; **67**: 301–320.
50. Hanley JA and McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**: 29–36.

51. Collins GS, Ogundimu EO and Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med* 2016; **35**: 214–226.
52. Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958; **45**: 562–565.
53. Miller ME, Langefeld CD, Tierney WM, et al. Validation of probabilistic predictions. *Med Decis Making* 1993; **13**: 49–58.
54. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**: 128–38.
55. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly weather review* 1950; **78**: 1–3.
56. R Core Team. A language and environment for statistical computing, http://www.r-project.org/ (2014, accessed 24 April 2018).
57. Venables WN and Ripley BD. *Modern applied statistics with S*. New York, NY: Springer, 2002.
58. Gelman A, Jakulin A, Pittau MG, et al. A weakly informative default prior distribution for logistic and other regression models. *Ann Appl Stat* 2008; **2**: 1360–1383.
59. Greenland S and Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Stat Med* 2015; **34**: 3133–3143.
60. Breiman L. Better subset regression using the nonnegative garrote. *Technometrics* 1995; **37**: 373–384.
61. Breiman L. Random forests. *Mach Learn* 2001; **45**: 5–32.