Comparison of likelihood penalization and variance decomposition approaches for clinical prediction models: a simulation study

Lohmann, A.; Groenwold, R.H.H.; Smeden, M. van

Received: 31 March 2022 | Revised: 30 September 2022 | Accepted: 10 November 2022

DOI: 10.1002/bimj.202200108

Biometrical Journal

RESEARCH ARTICLE

Comparison of likelihood penalization and variance decomposition approaches for clinical prediction models: A simulation study

Anna Lohmann[1,2] | **Rolf H. H. Groenwold**[2,3] | **Maarten van Smeden**[4]

[1]Department of Welfare, EAH Jena University of Applied Sciences, Jena, Germany

[2]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, The Netherlands

[3]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

[4]Julius Center for Health Science and Primary Care, University Medical Center Utrecht, Utrecht, The Netherland

Correspondence
Anna Lohmann, Department of Clinical Epidemiology, Leiden University Medical Center, 2300 RC Leiden, The Netherlands.
Email: anna.lohmann@eah-jena.de

Funding information
ZonMw, Grant/Award Number: 917.16.430

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

Logistic regression is one of the most commonly used approaches to develop clinical risk prediction models. Developers of such models often rely on approaches that aim to minimize the risk of overfitting and improve predictive performance of the logistic model, such as through likelihood penalization and variance decomposition techniques. We present an extensive simulation study that compares the out-of-sample predictive performance of risk prediction models derived using the elastic net, with Lasso and ridge as special cases, and variance decomposition techniques, namely, incomplete principal component regression and incomplete partial least squares regression. We varied the expected events per variable, event fraction, number of candidate predictors, presence of noise predictors, and the presence of sparse predictors in a full-factorial design. Predictive performance was compared on measures of discrimination, calibration, and prediction error. Simulation metamodels were derived to explain the performance differences within model derivation approaches. Our results indicate that, on average, prediction models developed using penalization and variance decomposition approaches outperform models developed using ordinary maximum likelihood estimation, with penalization approaches being consistently superior over the variance decomposition approaches. Differences in performance were most pronounced on the calibration of the model. Performance differences regarding prediction error and concordance statistic outcomes were often small between approaches. The use of likelihood penalization and variance decomposition techniques methods was illustrated in the context of peripheral arterial disease.

KEYWORDS
likelihood penalization, logistic regression, out-of-sample performance, simulation, variance decomposition

1 | INTRODUCTION

Binary logistic regression modeling remains one of the most common approaches for the development of clinical risk prediction models (Bouwmeester et al., 2012; Moons et al., 2015; Wynants et al., 2020). These prediction models are frequently used in clinical practice to estimate probabilities (risks) regarding current presence or future occurrence of health conditions in individual patients and, thereby, play an important role in modern medicine. Healthcare professionals as well as patients rely on such risk information to gain insight and make informed treatment decisions.

It is well known that standard maximum-likelihood-based logistic regression, which guarantees an asymptotically unbiased estimation and by definition provides a model with the highest likelihood in the data at hand, is often not optimal for making future predictions, especially when derivation data are small, the predictors are large in number, sparse, noisy, or highly correlated (Van Smeden et al., 2019). In clinical prediction modeling contexts, sample sizes are often relatively small and typically low dimensional ($n >> p$), often come with a relatively low number of events relative to the number of predictors considered for inclusion in the model (events per variable, EPV) (Bouwmeester et al., 2012). In such settings, overfitting of the maximum-likelihood-based prediction models is to be expected and data reduction approaches are applied to reduce overfitting of the maximum-likelihood-based prediction models and improve out-of-sample predictive performance (Harrell, 2015).

Data reduction can, for example, be achieved by regression shrinkage. Of special interest are penalized likelihood methods that incorporate the regression shrinkage within the estimation, such as through the elastic net (Zou & Hastie, 2005). An alternative approach for data reduction is to use variance decomposition methods (Harrell et al., 1984). These work by restructuring the data and deriving the model on a subset of variance components while ignoring components that account for the least amount of predictor variance. Although shrinkage approaches are gaining popularity for the development of clinical risk prediction models (Collins et al., 2015; Pavlou et al., 2016; Puhr et al., 2017; Van Smeden et al., 2019), data reduction via variance decomposition has so far been more widely applied in high-dimensional prediction contexts ($p > n$) (Harrell, 2015; Hastie et al., 2009).

Studies of the utility of variance decomposition approaches for the derivation of clinical risk prediction models in low-dimensional contexts as well as a comparison to alternatives in the form of shrinkage-based approaches are currently lacking. The present study attempts to fill this gap by simulation-based comparisons of risk prediction models based on binary logistic regression and incorporating either shrinkage or variance decomposition in their derivation. To compare the predictive performance under certain properties of the derivation data, we systematically varied the expected EPV, number of candidate predictors, event fraction, sparsity of predictors, and the presence of noise predictors using a full-factorial simulation design. Shrinkage was based on elastic net regression with its special cases: least absolute shrinkage and selection operator regression (LASSO) (Tibshirani, 1996) and ridge regression (Hoerl & Kennard, 1970; Le Cessie & van Houwelingen, 1992). Incomplete principal component regression and incomplete partial least squares regression (Frank & Friedman, 1993) were the variance-decomposition-based methods we studied. Variations in predictive performance were modeled using simulation metamodels.

The present study is conducted as a neutral comparison study (Boulesteix et al., 2018); that is, it aims at providing evidence-based guidance for the choice of data analytical approaches by comparing statistical methods in a neutral way. None of the authors have developed any of the methods under investigation nor do they have any other vested interest depending on the performance of any given method. To fulfill this aim, the methods section elaborates on justification regarding the choices of simulation scenarios as well as the reasoning that informed analysis/comparison choices of model performance. In an attempt to minimize selective reporting, the results section and Supporting Information contain an extensive number of tables and figures. Furthermore, the availability of the simulation code allows the interested reader to more closely inspect model performance for specific constellations of simulation factors and models. The discussion session draws extra attention to limitations of the present study that have to be kept in mind when weighting the evidence.

This article is structured as follows. In Section 2, we present a short introduction to the above-mentioned shrinkage-based methods and variance-decomposition-based approaches for data reduction. Section 3 describes the simulation setup and data-generating mechanism as well as details regarding modeling implementation and the derivation of metamodels with results presented in Section 4. In Section 5, we provide an illustration of the different approaches by emulating the derivation of a risk prediction model for peripheral arterial disease from the literature (Zhang et al., 2016). Section 6 discusses the findings.

2 | MODELS AND ESTIMATION

We assume a logistic regression model for estimating the risk of an event in individual i ($i = 1, \dots, N$), denoted by $y_i = 1$ in case of an event and $y_i = 0$ in case of a nonevent. The logistic regression model is parameterized via regression coefficients vector $\boldsymbol{\beta} = \beta_1, \cdots, \beta_p$ and intercept (a scalar) β_0 such that the risk of an event can be expressed as $\Pr(y_i = 1|\beta_0, \boldsymbol{\beta}, \mathrm{X}_i^\star) = \pi_i = 1/(1 + \exp\{-\beta_0 + \boldsymbol{\beta}\mathrm{X}_i^\star\})$, where $\mathbf{X}_i$ denotes the vector of predictor values for individual i and $\mathbf{X} \in \mathbb{R}^{N\times p}$ the matrix representing all N observations. With $\mathbf{X}_i^\star$, we denote a subset comprising $p^\star \leq p$ of the observed predictors or a transformation of the same, for example, the $p^\star$ first principle components. Regressors are generally assumed to be centered. Default maximum-likelihood-based logistic regression proceeds by maximizing the log-likelihood function

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

The following section describes two classes of data-reduction approaches applicable in the derivation of binary-logistic-regression-based clinical risk prediction models. (We use log throughout the article to refer to the natural logarithm.)

2.1 | Likelihood penalization

Likelihood penalization techniques are gaining popularity for the derivation of binary-logistic-based risk prediction models to improve out-of-sample predictive performance (Pavlou et al., 2016). The elastic net is one of such approaches to likelihood penalization, which proceeds by maximizing

$$\ell_{\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \lambda_2\|\boldsymbol{\beta}\|_2^2 - \lambda_1\|\boldsymbol{\beta}\|_1 = \ell(\boldsymbol{\beta}) - \alpha\|\boldsymbol{\beta}\|_2^2 - (1 - \alpha)\|\boldsymbol{\beta}\|_2^2,$$

with $\alpha = \lambda_2/(\lambda_2 + \lambda_1)$ (Zou & Hastie, 2005). The optimal value for the tuning parameters λ_2 and λ_1 can be approximated using K-fold cross-validation (CV) optimized for a particular predictive performance criterion. In this article, we apply 10-fold CV using deviance as the performance criterion, that is, finding values for the tuning parameters that minimize the deviance at CV.

The popular ridge regression model, also known as L2 regularization, can be seen as a special case of the elastic net (Hoerl & Kennard, 1970; Le Cessie & Van Houwelingen, 1992) where $\alpha = 0$. The penalized log-likelihood function simplifies to

$$\ell_{\lambda_2}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \lambda_2\|\boldsymbol{\beta}\|_2^2.$$

The LASSO model (Tibshirani, 1996), also known as L1 regularization, is a special case of the elastic net where $\alpha = 1$. The penalized log-likelihood function is

$$\ell_{\lambda_1}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \lambda_1\|\boldsymbol{\beta}\|_1.$$

Ridge, LASSO, and elastic net maximize a penalized likelihood which constrains the size of the sum of the absolute values of the coefficients (LASSO), the sum of squared values (ridge), or a combination (elastic net). Since LASSO and elastic net can shrink coefficients to obtain a value of zero thereby perform variable selection, unlike the squared values penalty of ridge regression which does not lead to shrunken coefficients being absolutely zero.

Besides the standard LASSO, we also consider the relaxed LASSO (Meinshausen, 2007), in which the LASSO is applied once to obtain candidate predictor subsets and a second time to obtain optimal cross-validated solutions on all candidate subsets,

$$\max_{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^{\mathcal{M}_\lambda+1}} \ell_{\lambda,\phi}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \phi\lambda\|\boldsymbol{\beta}\|_1 \quad \text{with } \mathcal{M}_\lambda = \{1 \leq k \leq p^\star | \widehat{\beta}_k^{\lambda_1} \neq 0\}.$$

In terms of predictor selection for prediction modeling, the relaxed-LASSO has shown to be promising in high-dimensional settings (Hastie et al., 2017) and is worth investigating further.

TABLE 1 Overview of simulation factors ($7 \times 5 \times 5 \times 3 \times 2 = 1050$ scenarios).

Simulation factor	Factor levels	Implementation details
Expected events per variable	3, 5, 10, 15, 20, 50, 100	
Expected event fraction	1/32, 1/16, 1/8, 1/4, 1/2	
Number of candidate predictors	4, 8, 16, 32, 64	
Fraction of noise predictors	0, 1/4, 1/2	Regression coefficient set to zero
Presence of sparse predictors	Yes, No	1/4 of predictors sparse
Pairwise predictor correlations	Sampled	Sampled from Beta (1, 3) distribution
Predictor effects	Sampled	Sampled from N(0,1) divided by the standard deviation of resulting linear predictor

2.2 | Variance decomposition

Variance decomposition approaches extract uncorrelated components that are linear combinations of the p (if only the predictor data are used) or $p + 1$ (if also the outcome is used) columns of the dataset. Each component maximizes the (remaining) amount of variance in the predictor and/or outcome data. Applications in prediction modeling often rely on incomplete component analyses, where a subset of most variance explaining components is chosen (dimension reduction). The selected $p^\star$ components are then used as predictor variables in a penalized or unpenalized logistic regression model.

Principal component analysis (PCA) restructures only the predictor data. The kth principal component $\mathbf{t}_k$ is given by

$$\mathbf{t}_k = \mathbf{X}\mathbf{w}_k \quad \text{with} \quad k = 1, \dots, p,$$

where $\mathbf{w}_k$ is the k-th eigenvector of $\mathbf{X}^T\mathbf{X}$.

Partial least squares (PLS) maximize the explained variance in the predictor space and the relationship to the outcome. Unlike PCA, PLS can be viewed as a supervised variance decomposition approach (Frank & Friedman, 1993). The $(h + 1)$-th partial least squares component is given by

$$\mathbf{t}_{h+1} = \frac{1}{\sum_{j=1}^{p} a_{h+1,j}^2} \sum_{j=1}^{p} \mathbf{a}_{h+1,j}\tilde{\mathbf{x}}_{hj} \quad \text{with} \quad h = 1, \dots, p-1,$$

where $\tilde{\mathbf{x}}_{hj}$ denotes the residuals obtained by ordinary least squares (OLS) linear regression of $\mathbf{X}_j$ on the previously found h components. $a_{h+1,j}$ represents the regression coefficients of $\tilde{\mathbf{x}}_{hj}$ in the binary logistic regression of $\mathbf{y}$ on $\mathbf{t}_1 \dots, \mathbf{t}_h$ and $\tilde{\mathbf{x}}_{hj}$ (Meyer et al., 2010).

There are various approaches to select $p^\star$ components for incomplete PCA and incomplete PLS, such as by the 90% explained variance rule (Cook, 2007). PCA and PLS can both be performed with the intention of dimensionality reduction, they both do not contribute directly to a more sparse solution in terms of a reduced number of predictors in the final model.

3 | METHODS

This simulation study was set up to evaluate and compare penalization, variance decomposition, and combinations of them in the context of developing clinical prediction models. Our primary interest was in the relative performance of different modeling strategies, and how they varied with characteristics of the data. The data generation is described in Section 3.1, and the modeling strategies are detailed in Section 3.2. Predictive performance measures are described in Section 3.3 and simulation metamodels in Section 3.3.2. Detection and handling of estimation errors are discussed in Section 3.4.

3.1 | Data simulation design

We conducted a full-factorial simulation study examining five design factors (Table 1). These five factors were the expected EPV, ranging from 3 to 100, expected events fraction ($\Pr(y = 1)$), ranging from 3% to 50%, number of candidate predictors

(p), ranging from 4 to 64, and the fraction of noise predictors from 0% to 50%. In total $n_{sim} = 1050$ unique constellations of simulation factors were investigated. Each of these scenarios was implemented $n_{iter} = 20$ times with a unique seed obtained by a hash function combining scenario ID and iteration ID. One iteration of a single scenario included the following steps:

1. Generation of one dataset for derivation and one independently generated validation dataset of size $n_{val} = 20 \times p \times \frac{1}{event\ fraction}$ (Table 1). The derivation and validation data were generated under the same data-generating mechanism, described in more detail in Section A of the Supporting Information.
2. Various risk prediction models were developed using likelihood penalization, variance decomposition, and combinations, as outlined in Section 3.2. Model development was carried out on the derivation set generated in step (1).
3. Application of the prediction models derived in step (2) on the validation data generated in step (1). The out-of-sample performance of each model was obtained with the predictive performance measures detailed in Section 3.3.

Predictor data were simulated by sampling from a multivariate standard normal distribution with a given variance-covariance matrix, $\mathbf{\Sigma} = \text{var}(\mathbf{X})$. For each simulation iteration, the pairwise predictor correlations $\sigma_{i,j}$ were sampled from a $Beta(1, 3)$ distribution. Regression coefficients for the data-generating model were sampled from a standard normal distribution for each simulation iteration and were scaled by the variance of the linear predictor to ensure realistic area under curve (AUC) (ranging from 0.303 to 0.982). The intercept was numerically approximated to correspond to the desired event fraction. Outcome data were simulated by draws from a Bernoulli distribution with success probability corresponding to the linear predictor imposing a logit link.

3.2 | Prediction models

We studied several of the risk prediction modeling strategies, described in Section 2, also applied in combination. As a reference, binary logistic regression based on maximum likelihood estimation (MLE) was also implemented using the `glm()` function in R.

Elastic net regression was implemented using the `cv.glmnet()` R-function from the `glmnet` R-package (version 2.0-16) (Simon et al., 2011). The hyperparameter α was chosen by 10-fold CV from the following set $\alpha \in \{0, 0.125, 0.25, 0.5, 0.625, 0.75, 0.875, 1\}$. The tuning parameter λ was obtained by minimizing the deviance in 10-fold CV with a grid of 100 possible λ values ranging from the smallest value necessary to shrink all coefficients to zero down to 10^{-3} (see Supporting Information B for exceptions).

Ridge and LASSO models were obtained as the corresponding $\alpha = 0$ and $\alpha = 1$ elastic net solutions, respectively. Relaxed LASSO models were derived by obtaining predictor subsets (supports) across all α's from the original LASSO solution. Ordinary LASSO regression models were then fit on each of these subsets individually by application of the `cv.glmnet()` R-function. All `cv.glmnet()`-based models were estimated with a random `foldid` parameter to ensure identical folds across approaches based on the same data.

PCA was implemented with the `pcrcomp()` R-function. Incomplete principal component regression was carried out by using principal component (sub)sets as predictors in MLE, ridge, and LASSO logistic regression models. Four commonly applied subsets selection procedures were incorporated: (1) all principal components corresponding to eigenvalues above one, (2) the principal components necessary to explain at least 90% of predictor variance, (3) the principal components corresponding to the lowest akaike information criterion (AIC) in the MLE model, and (4) the principal components corresponding to minimal deviance obtained by 10-fold CV.

PLS was implemented with the `plsRglm()` function from the `plsRglm` package (version 1.2.5) (Meyer et al., 2010), and the `cv.glmnet()` function was applied on the component scores. Further models were fit using a lower number of components obtained by the (`sparseStop=TRUE`) (Bastien et al., 2005) option, which retains additional components, if given the first components the remaining explanatory variables of the outcome among the original predictors reach statistical significance (alpha = 0.05). Due to practical computational time limitations and incidental extreme run-times, the number of PLS components was limited to a maximum of 30. The subsetted PLS component scores were used as predictors in MLE, ridge, and LASSO logistic regression models.

TABLE 2 Overview of performance measures.

Performance measure	Definition
Discrimination	
Concordance (c-statistic)	Area under the ROC curve for $\hat{y}_i$
Calibration	
Re-calibration model	$\log(\frac{p_i}{1-p_i}) = a + b\log(\frac{\hat{y}_i}{1-\hat{y}_i})$, with $p_i = Pr(y_i = 1\|X_i)$
Calibration slope	b in recalibration model
Calibration in the large	a in a recalibration model with $b = 1$
Prediction error	
Brier score	$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
Root mean squared prediction error	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\pi_i - \hat{y}_i)^2}$
Mean absolute prediction error	$\frac{1}{N}\sum_{i=1}^{N}\|\pi_i - \hat{y}_i\|$

Note: N, sample size; y_i, observed outcome in validation data; $\hat{y}_i$, predicted event probability; π_i, true probability based on data-generating model.

3.3 | Predictive performance metrics

Predictive performance was evaluated with respect to discrimination, calibration, and total prediction error (Steyerberg et al., 2010) in the independently generated validation datasets. Table 2 provides a detailed overview of all performance measures assessed in the present study as well as their computation.

Discrimination was evaluated using the *concordance statistic* (c-statistic) which, for binary logistic regression models, is equivalent to the area under the ROC curve (Steyerberg et al., 2010). This rank order statistic can range from 0.5 (indicating no discrimination) to 1 (indicating perfect discrimination).

Calibration was assessed by the *calibration slope* and *calibration in the large* (CIL) (Miller et al., 1993; Steyerberg et al., 2010). Perfect (weak) calibration corresponds to a calibration slope of one, with the predicted risk matching the observed frequency and CIL of zero (van Calster et al., 2019). Calibration slopes < 1 indicate predictions that are too extreme, indicating overfitting; calibration slopes of > 1 indicate underfitting. The CIL > 0 indicates systematically too low predicted risk; CIL < 0 indicates systematically too high predicted risks (Steyerberg et al., 2010).

Prediction error was evaluated using the *Brier score* (also *average prediction error* (Steyerberg et al., 2004)), the *root mean squared prediction error* (rMSPE), and the *mean absolute prediction error* (MAPE). The rMSPE and MAPE pertain to the distance of predicted and true probabilities and can thus only be applied when the true probabilities are known (as in simulation studies). Perfect models have prediction errors of zero.

3.3.1 | Ranking

The predictive performance of methods was ranked per simulation iteration. Hence performance of modeling approaches was compared based on identical derivation as well as validation data. The performance was rounded to three decimals for the c-statistic and Brier score and two decimals for calibration slope. In the case of the calibration slope, slopes closer to one received a higher ranking. Ties received the minimum of shared ranks, while not affecting the other ranks.

3.3.2 | Derivation of metamodels

Variations in predictive performance across simulation parameters were modeled via simulation metamodels (Harwell et al., 1992; Kleijnen & Sargent, 2000), which quantify the impact of features of the derivation dataset on a given predictive performance measure per model derivation approach. A second set of metamodels was fit containing parameters that are easily conceivable in practical applications. We will refer to the former set of metamodels as *full* and the latter as *simplified*.

Metamodels were developed for the following performance indicators. Discrimination was evaluated in terms of loss in c-statistic (ΔAUC of the c-statistic obtained from the development dataset and the c-statistic from the validation set).

For the calibration, we modeled the absolute value of 1 minus the calibration slope. Higher values indicated a calibration slope further away from the ideal of one. Due to the lack of variance in CIL, this indicator of predictive performance was not modeled. For the prediction error outcomes, metamodels were developed for log-transformed Brier score, rMSPE, and MAPE.

All metamodels were developed on the simulation outcome data (each row is a simulation run) using OLS-based ridge regression with the performance indicators as outcomes and the following covariates for the full metamodels: p (natural log transformed), EPV (natural log transformed), estimated c-statistic in the derivation dataset, event fraction in the derivation dataset, percentage of noise predictors (ordinal with three levels), presence of sparse binary predictors (dummy coded), upper quartile of empirical bivariate absolute correlation in derivation dataset, square root of median variance inflation factor of candidate predictors, upper quartile of absolute predictor effects.

3.4 | Computation and error handling

Simulations were carried out on high-performance clusters. Details regarding the computational framework can be obtained from Supporting Information Section A. Overall, the simulation ran the equivalent of approximately 2400 CPU core hours and was computed on up to 200 servers in parallel in the AWS compute cloud. The statistical software R (version 3.5.3) (R Core Team, 2019) was used for data simulation, prediction model derivation, and all consecutive analyses. Data generation and prediction model derivation was carried out in docker containers based on identical images, ensuring identical software setup between simulation iterations. Estimation errors were closely monitored (see Table D.3 in the Supporting Information for descriptive statistics).

We followed the recommendations of Morris et al. (2019) to mimic real-life strategies in light of modeling difficulties by using substitutions when a certain method was not estimable. If component selection resulted in less than two components, no consecutive glmnet-based analysis could be performed as glmnet requires a minimum of two predictors. In such a case, the result was replaced by the corresponding MLE result. When LASSO regression resulted in a model without any predictors, the result was substituted with the MLE logistic regression solution. Extreme calibration slopes were winsorized at 10 and 0.01. For derivation of metamodels, these values were treated as missing.

The complete simulation and analysis code can be found on the Open Science Framework (OSF) https://osf.io/gcjn6/. Additional measures taken to ensure reproducibility can be found in Supporting Information A.

3.5 | Justification of design decisions to promote neutral comparison

Simulation parameters were derived from the literature. In order to not give methods an advantage that frequently "break" under less favorable simulation scenarios, we provide detailed error descriptives. These runs were furthermore substituted with the unpenalized maximum likelihood result which we consider the fallback option for most methods in practice. The ranking of methods was performed after rounding to three decimals. Hereby, we tried to emulate a realistic assessment of equally good performance As some models can in rare cases result in extreme calibration slopes thereby highly distorting average performance we winsorized calibration slopes at 10 and 0.01. To provide a more complete picture of the between model comparison in terms of ranking, we complemented the average rank of each method with the full ranking distribution, thereby allowing to potentially uncover whether some methods might show a U-shaped performance. In the same vein, we provide an insight into performance variability using boxplots.

4 | RESULTS

Figure 1 summarizes the average predictive performance of the likelihood penalization and variance decomposition risk prediction modeling approaches. It is shown that on average, with the exception of cross-validated PCA, all modeling approaches outperformed the default MLE on the c-statistic, calibration slope, rMSPE, and MAPE. In addition to superior average performance, the c-statistic and calibration slope of most MLE alternatives varied less over simulation iterations as indicated by the boxplot whiskers in Figure 1.

Figure 2 displays the average performance ranking regarding c-statistic, Brier score, and calibration slope. The likelihood-penalization-based methods persistently occupy top ranks compared to the variance decomposition approaches

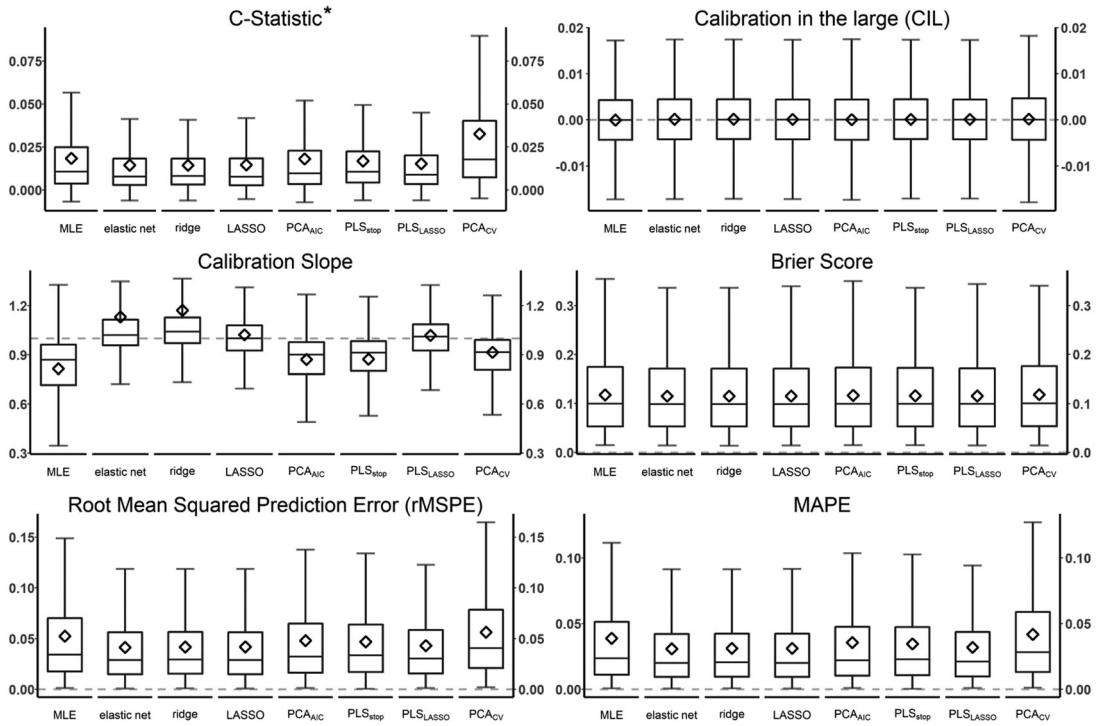

FIGURE 1 Boxplots of performance measures per modeling approach. Diamonds indicate mean performance. Whiskers extend to the third quartile plus $1.5 \times IQR$ as well as the first quartile minus $1.5 \times IQR$. *The panel displays the distance to an optimal c-statistic with a value of 1. MLE, binary logistic regression with maximum likelihood estimation; LASSO, least absolute shrinkage and selection operator; PCA_{AIC}, incomplete principal component regression with components that correspond to minimal AIC; PLS_{stop}, incomplete partial least squares regression with stopping criterion; PLS_{LASSO}, LASSO regression on PLS components; PCA_{CV}, incomplete principal component regression with components that correspond to minimal CV error; MAPE, mean absolute prediction error.

and penalization–decomposition combinations regarding the validated c-statistic and Brier score. Patterns are less clear regarding calibration slope. With the elastic net and LASSO also occupying the top tiers regarding calibration slope, some variance decomposition approaches and penalization–decomposition combinations obtain a high ranking despite their low rankings on the other performance measures. Additionally, the low ranking of ridge regression is noteworthy. Differences between methods regarding Brier score and rMSPE and MAPE were generally small, leading to a large number of shared ranks.

Below we discuss the simulation results for discrimination, calibration, and prediction error separately.

4.1 | Discrimination

Figures 1 and 3 and Table 3 show the loss in the area under the curve expressed as the difference of the c-statistic obtained in the validation dataset and the c-statistic of the data-generating model.

For all methods, the loss in the c-statistic was greater when EPV and the true AUC were lower and the event fraction was closer to 0.5 (Table 3).

The presence of sparse predictors, negatively affected the discrimination of all modeling approaches whereby the effect on MLE was the most pronounced, and discrimination of penalization approaches was affected the least. Up to 15 EPV, penalization as well as PLS-based modeling approaches clearly exhibited a lower loss in the c-statistic compared to MLE both in terms of average performance as well as in a lower interquartile range (IQR).

While the regression penalization approaches systematically outranked the variance decomposition approaches (see Figure 2), the differences in average discriminative performance were generally small (see Figure 3), with the exception of the poorer performing cross-validated PCA. This observation is further supported by the results of the metamodels for loss in c-statistics compared to the data-generating model (Table 3), with similar results across penalization and variance decomposition approaches.

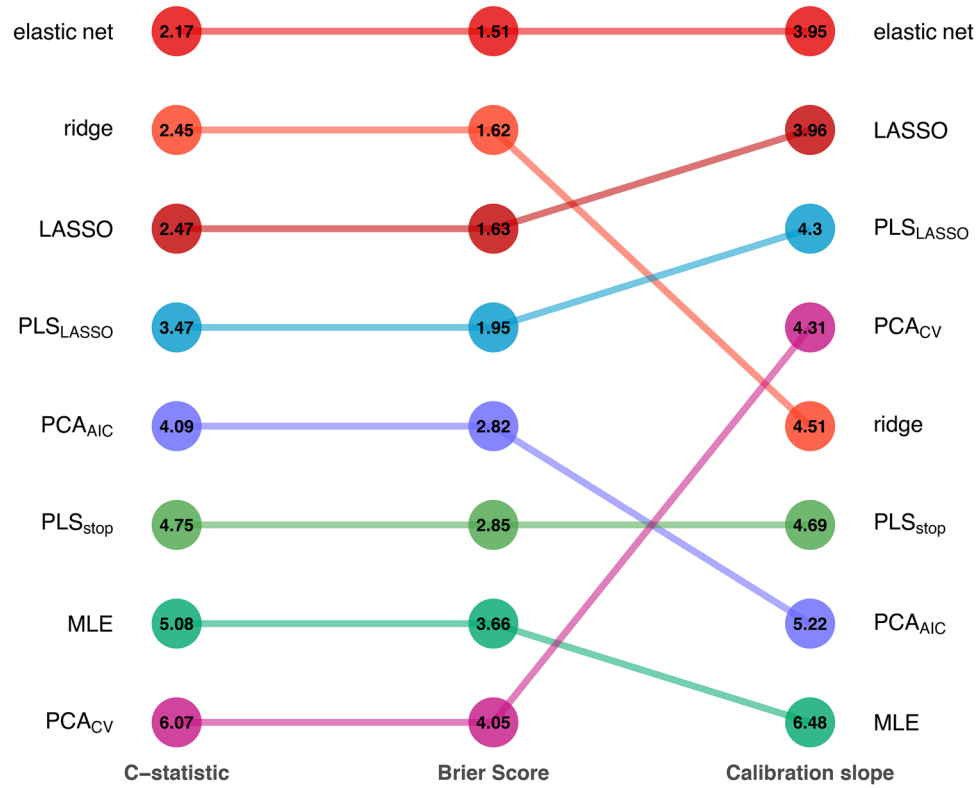

FIGURE 2 Average ranking of modeling approaches. Ranking of calibration slope based on the absolute distance to one. Ranking was performed on performance rounded to two decimals. Sport ranking was applied with equal performances receiving a minimum of shared rank. Numbers in circles refer to average ranking. MLE, binary logistic regression with maximum likelihood estimation; LASSO, least absolute shrinkage and selection operator; PCA_{AIC}, incomplete principal component regression with components that correspond to minimal AIC; PLS_{stop}, incomplete partial least squares regression with stopping criterion; PLS_{LASSO}, LASSO regression on PLS components; PCA_{CV}, incomplete principal component regression with components that correspond to minimal CV error.

4.2 | Calibration

Calibration in the large did not differ noticeably between penalization and variance decomposition approaches. All methods produced models with near-perfect CIL with small variation of similar magnitude across modeling approaches (Figure 1).

Conversely, the calibration slope varied strongly between different modeling approaches. On average, most penalization and variance decomposition approaches outperformed the MLE. The impact of individual simulation parameters is illustrated in Figure 4. As expected, EPV noticeably influenced calibration slopes across modeling approaches, with calibration slopes coming closer to the ideal value with increasing EPV. As shown in Figure 4, the variance-decomposition-based approaches exhibited a tendency to produce calibration slopes under 1, consistent with model overfitting and in the same direction as MLE, while penalization-based approaches as well as the hybrid PLS_{LASSO} are better calibrated with a slight tendency to produce calibration slopes above 1, consistent with model underfitting. The calibration slopes across models were largely unaffected by the presence of noise predictors or sparse predictors.

Figure 5 further shows the variation of the calibration slope from the ideal value (value of 1), following the work in Van Calster et al. (2020) expressed in the root mean squared difference (RMSD) in the logarithm of the calibration slope. The general trend of improved performance is comparable across approaches with differences between modeling approaches largely disappeared at EPV above 25, with the exception of the consistently poorer performing cross-validated PCA.

Compared to the other performance measures, within-model variation regarding calibration slope cannot be well explained by the covariates of the metamodels, as evidenced by low R^2 values (Table 4). Especially among variance decomposition approaches, explained variance was low indicating limited influence of the data characteristics under investigation beyond EPV in this study.

TABLE 3 Metamodel for c-statistic, modeled outcome: $\log(\Delta AUC) = \log(|c_{dgm} - c_{val}|)$.

Model	Method	interc	log(*EPV*)	log(*event frac*)	log(*p*)	log(*AUC*)	Sparse	Noise	$\frac{VIF}{p(p-1)/2}$	ES	R^2	*n*
Full	MLE	−14.121	−1.154	0.555	−0.216	25.652	0.766	−0.012	6.60e-06	−0.576	0.429	19727
Simplified	MLE	−13.396	−1.140	0.546	0.031	24.682	0.735	0.036	2.85e-06	–	0.418	19727
Full	enet	−13.035	−0.918	0.440	−0.303	23.410	0.776	−0.017	−1.67e-05	−0.214	0.339	19727
Simplified	enet	−11.882	−0.864	0.406	−0.193	21.658	0.697	0.000	−1.72e-05	–	0.334	19727
Full	ridge	−13.474	−0.894	0.432	−0.338	23.789	0.803	0.030	−1.60e-05	−0.28	0.347	19727
Simplified	ridge	−12.495	−0.856	0.407	−0.205	22.347	0.740	0.051	−1.72e-05	–	0.341	19727
Full	LASSO	−13.836	−0.990	0.483	−0.334	24.949	0.802	−0.096	1.99e-06	−0.291	0.366	19348
Simplified	LASSO	−12.543	−0.934	0.444	−0.189	22.995	0.716	−0.067	−1.35e-07	–	0.361	19348
Full	PCA_{AIC}	−10.540	−1.123	0.494	−0.343	21.350	0.846	0.004	−9.86e-06	−0.23	0.364	19727
Simplified	PCA_{AIC}	−9.853	−1.086	0.473	−0.236	20.304	0.798	0.021	−1.12e-05	–	0.358	19727
Full	PLS_{stop}	−13.485	−0.924	0.434	−0.259	24.282	0.782	−0.003	−1.71e-05	−0.15	0.364	19727
Simplified	PLS_{stop}	−13.357	−0.925	0.434	−0.197	24.135	0.779	0.009	−1.81e-05	–	0.360	19727
Full	PLS_{LASSO}	−14.750	−0.966	0.464	−0.310	25.809	0.782	−0.010	−3.80e-06	−0.389	0.384	19596
Simplified	PLS_{LASSO}	−14.181	−0.956	0.457	−0.140	25.036	0.757	0.023	−6.28e-06	–	0.376	19596
Full	PCA_{CV}	−4.754	−1.529	0.692	−0.945	21.453	0.765	0.037	−6.19e-05	0.778	0.335	29024
Simplified	PCA_{CV}	−5.934	−1.599	0.734	−1.328	23.421	0.836	−0.025	−6.19e-05	–	0.336	29024

Abbreviations: Full, includes all predictors; Simplified, does not contain ES (upper quartile of absolute predictor effect); interc, intercept; EPV, events per variable; event frac, empirical event fraction; *p*, number of candidate predictors; AUC, empirical area under the ROC curve; sparse, sparse predictors present (yes/no); VIF, variance inflation factor; ES, upper quartile of predictor effects; R^2 cross-validated R^2; n = number of complete cases that were used for the derivation of the metamodel; MLE, binary logistic regression with maximum likelihood estimation; LASSO, least absolute shrinkage and selection operator; PCA_{AIC}, incomplete principal component regression with components that correspond to minimal AIC; PLS_{stop}, incomplete partial least squares regression with stopping criterion; PLS_{LASSO}, LASSO regression on PLS components; PCA_{CV}, incomplete principal component regression with components that correspond to minimal CV error.

TABLE 4 Metamodel for calibration slope, modeled outcome: $|1 - calibrationslope|$.

Model	Method	interc	log(*EPV*)	log(*event frac*)	log(*p*)	log(*AUC*)	Sparse	Noise	$\frac{VIF}{p(p-1)/2}$	ES	R^2	*n*
Full	MLE	0.729	−0.104	0.040	−0.015	−0.246	0.034	0.001	−1.83e-07	0.008	0.536	19727
Simplified	MLE	0.725	−0.102	0.039	−0.016	−0.242	0.033	0.001	−1.58e-07	–	0.536	19727
Full	enet	1.311	−0.000	0.000	−0.000	−0.000	0.000	−0.000	−9.53e-41	0	0.041	19646
Simplified	enet	1.311	−0.000	0.000	−0.000	−0.000	0.000	−0.000	−9.53e-41	–	0.040	19646
Full	ridge	1.818	−0.000	0.000	−0.000	−0.000	0.000	−0.000	−1.25e-40	0	0.050	19642
Simplified	ridge	1.818	−0.000	0.000	−0.000	−0.000	0.000	−0.000	−1.25e-40	–	0.048	19642
Full	LASSO	0.144	−0.000	0.000	−0.000	−0.000	0.000	0.000	5.93e-42	0	0.206	19348
Simplified	LASSO	0.244	−0.006	0.001	−0.005	−0.096	0.002	0.000	4.74e-07	–	0.206	19348
Full	PCA_{AIC}	0.243	−0.000	−0.000	−0.000	−0.000	0.000	−0.000	−6.43e-42	0	0.004	19727
Simplified	PCA_{AIC}	0.243	−0.000	−0.000	−0.000	−0.000	0.000	−0.000	−6.43e-42	–	0.003	19727
Full	PLS_{stop}	0.521	−0.059	0.020	−0.023	−0.230	0.029	0.000	−1.37e-07	0.061	0.400	19727
Simplified	PLS_{stop}	0.555	−0.061	0.020	−0.030	−0.215	0.031	−0.001	3.11e-08	–	0.400	19727
Full	PLS_{LASSO}	0.538	−0.023	0.007	−0.017	−0.407	0.003	0.001	1.81e-06	0.066	0.242	19596
Simplified	PLS_{LASSO}	0.421	−0.015	0.004	−0.014	−0.267	0.003	0.000	1.40e-06	–	0.240	19596
Full	PCA_{CV}	0.340	−0.000	0.000	−0.000	−0.000	−0.000	−0.000	−1.86e-41	0	0.002	29024
Simplified	PCA_{CV}	0.340	−0.000	0.000	−0.000	−0.000	−0.000	−0.000	−1.86e-41	–	0.002	29024

Abbreviations: Full, includes all predictors; Simplified, does not contain ES (upper quartile of absolute predictor effect); interc, intercept; EPV, events per variable; event frac, empirical event fraction; *p*, number of candidate predictors; AUC, empirical area under the ROC curvel sparse, sparse predictors present (yes/no); VIF, variance inflation factor; ES, upper quartile of predictor effects; R^2, cross validated R^2; *n*, number of complete cases that were used for the derivation of the metamodel; MLE, binary logistic regression with maximum likelihood estimation; LASSO, least absolute shrinkage and selection operator; PCA_{AIC}, incomplete principal component regression with components that correspond to minimal AIC; PLS_{stop}, incomplete partial least squares regression with stopping criterion; PLS_{LASSO}, LASSO regression on PLS components; PCA_{CV}, incomplete principal component regression with components that correspond to minimal CV error.

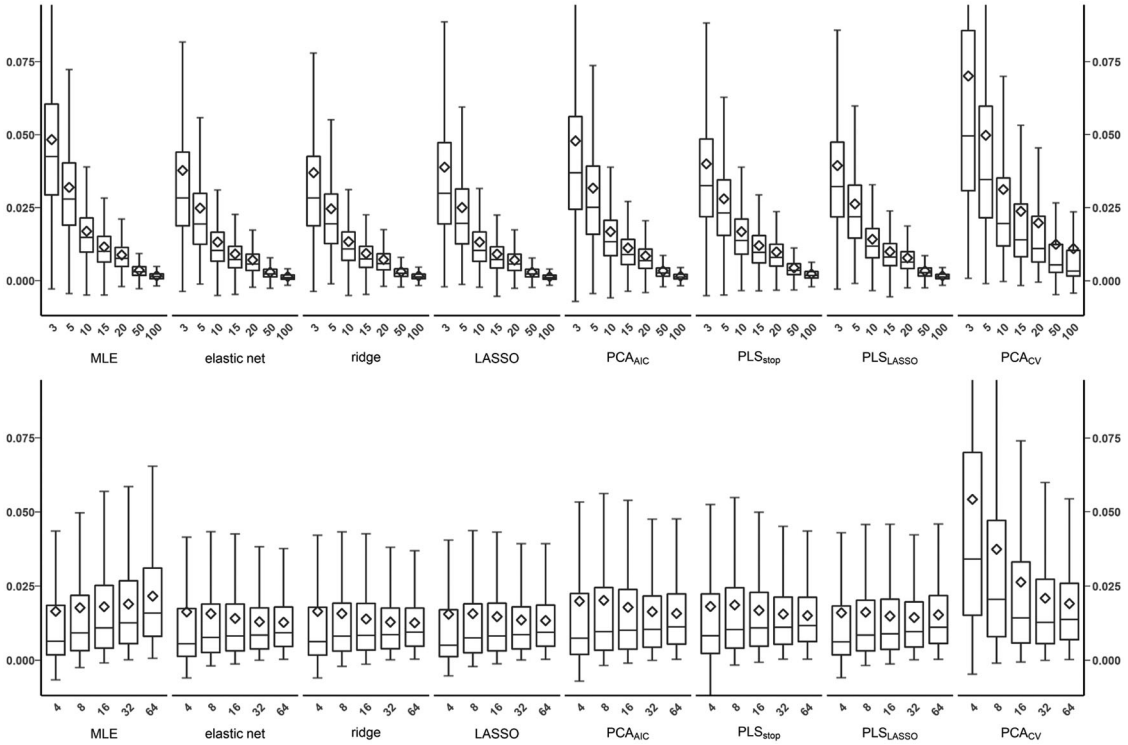

FIGURE 3 Boxplots of distance to optimal c-statistic of 1 per modeling approach and simulation factor. (Top: expected EPV. Bottom: number of candidate predictors.) Diamonds indicate mean performance. Whiskers extend to the third quartile plus $1.5 \times IQR$ as well as the first quartile minus $1.5 \times IQR$. Outliers are omitted due to their large number. MLE, binary logistic regression with maximum likelihood estimation; LASSO, least absolute shrinkage and selection operator; PCA_{AIC}, incomplete principal component regression with components that correspond to minimal AIC; PLS_{stop}, incomplete partial least squares regression with stopping criterion; PLS_{LASSO}, LASSO regression on PLS components; PCA_{CV}, incomplete principal component regression with components that correspond to minimal CV error.

4.3 | Prediction error

The Brier score did not differ noticeably between penalization and variance decomposition approaches (Figure 1). Variation in the Brier score could be well explained by metamodels across modeling approaches (Table 5), showing similar results across penalization and variance decomposition approaches.

Small but noticeable differences are found in the rMSPE and MAPE outcomes, in a similar pattern to that found for the c-statistic (Figure 1). While differences are small, on average, the regression penalization approaches systematically perform better than the variance decomposition approaches (see Table 6).

4.4 | Additional results: Combinations between penalization and decomposition approaches

Supplementary Figure D.4 displays additional results where penalization–decomposition is combined.

(See Supporting Information Section D for the full distribution of ranking for each method and performance measure.)

5 | APPLICATION TO THE PREDICTION OF PERIPHERAL ARTERIAL DISEASE

To illustrate the likelihood penalization and variance decomposition methods in a real example, we imitated a study by Zhang et al. (2016) who derived a risk prediction model for peripheral arterial disease using data from the National Health and Nutrition Examination Survey (NHANES). The identified candidate predictors consisted of a mixture of demography, lifestyle, co-morbidity, and physiology. The predictors were low-to-moderately correlated with bivariate correlations

TABLE 5 Metamodel for the Brier score, modeled outcome: log(*brier*)).

Model	Method	interc	log(*EPV*)	log(*eventfrac*)	log(*p*)	log(*AUC*)	Sparse	Noise	$\frac{VIF}{p(p-1)/2}$	ES	R^2	*n*
Full	MLE	−0.479	−0.026	0.673	−0.005	−0.601	−0.003	−0.001	5.72e-07	−0.012	0.961	19727
Simplified	MLE	−0.469	−0.026	0.673	0.000	−0.614	−0.004	−0.000	5.12e-07	–	0.961	19727
Full	enet	−0.470	−0.016	0.663	−0.006	−0.683	−0.005	−0.001	2.97e-07	−0.007	0.960	19727
Simplified	enet	−0.464	−0.016	0.663	−0.002	−0.690	−0.005	−0.001	2.62e-07	–	0.960	19727
Full	ridge	−0.469	−0.016	0.663	−0.006	−0.684	−0.005	−0.001	2.11e-07	−0.007	0.960	19727
Simplified	ridge	−0.463	−0.016	0.663	−0.003	−0.692	−0.005	−0.000	1.73e-07	–	0.960	19727
Full	LASSO	−0.499	−0.015	0.664	−0.005	−0.653	−0.004	−0.001	3.01e-07	−0.008	0.961	19348
Simplified	LASSO	−0.491	−0.015	0.664	−0.001	−0.663	−0.004	−0.001	2.55e-07	–	0.961	19348
Full	PCA_{AIC}	−0.494	−0.022	0.669	−0.008	−0.601	−0.002	−0.001	4.38e-07	−0.009	0.960	19727
Simplified	PCA_{AIC}	−0.487	−0.022	0.669	−0.004	−0.611	−0.002	−0.000	3.93e-07	–	0.960	19727
Full	PLS_{stop}	−0.498	−0.019	0.667	−0.006	−0.619	−0.002	−0.001	1.38e-07	−0.007	0.961	19727
Simplified	PLS_{stop}	−0.492	−0.019	0.667	−0.003	−0.627	−0.003	−0.000	1.03e-07	–	0.961	19727
Full	PLS_{LASSO}	−0.510	−0.015	0.664	−0.004	−0.635	−0.004	−0.001	2.00e-07	−0.009	0.961	19596
Simplified	PLS_{LASSO}	−0.502	−0.015	0.664	−0.000	−0.645	−0.004	−0.000	1.50e-07	–	0.961	19596
Full	PCA_{CV}	−0.481	−0.021	0.671	−0.010	−0.586	−0.004	−0.000	1.60e-07	−0.002	0.960	29024
Simplified	PCA_{CV}	−0.479	−0.021	0.671	−0.009	−0.588	−0.004	−0.000	1.51e-07	–	0.960	29024

Abbreviations: Full, includes all predictors; Simplified, does not contain ES (upper quartile of absolute predictor effect); interc, intercept; EPV, events per variable; event frac, empirical event fraction; *p*, number of candidate predictors; AUC, empirical area under the ROC curve; sparse, sparse predictors present (yes/no); VIF, variance inflation factor; ES, upper quartile of predictor effects; R^2, cross-validated R^2; *n*, number of complete cases that were used for the derivation of the metamodel; MLE, binary logistic regression with maximum likelihood estimation; LASSO, least absolute shrinkage and selection operator; PCA_{AIC}, incomplete principal component regression with components that correspond to minimal AIC; PLS_{stop}, incomplete partial least squares regression with stopping criterion; PLS_{LASSO}, LASSO regression on PLS components; PCA_{CV}, incomplete principal component regression with components that correspond to minimal CV error.

TABLE 6 Metamodel for MAPE, modeled outcome: log(*MAPE*).

Model	Method	interc	log(*EPV*)	log(*eventfrac*)	log(*p*)	log(*AUC*)	Sparse	Noise	$\sqrt{VIF}$	ES	R^2	*n*
Full	MLE	−0.548	−0.481	0.809	−0.003	−0.349	−0.012	−0.002	2.72e-06	−0.009	0.942	19727
Simplified	MLE	−0.541	−0.481	0.809	0.001	−0.358	−0.012	−0.001	2.68e-06	–	0.942	19727
Full	enet	0.266	−0.424	0.779	−0.042	−1.616	−0.011	−0.015	−3.02e-07	0.11	0.928	19727
Simplified	enet	0.176	−0.424	0.779	−0.092	−1.494	−0.008	−0.025	2.15e-07	–	0.927	19727
Full	ridge	0.028	−0.392	0.771	−0.051	−1.410	−0.000	0.006	−9.18e-07	0.108	0.929	19727
Simplified	ridge	−0.060	−0.392	0.771	−0.100	−1.291	0.003	−0.004	−4.06e-07	–	0.928	19727
Full	LASSO	0.088	−0.424	0.775	−0.038	−1.423	−0.006	−0.024	−3.56e-07	0.096	0.926	19348
Simplified	LASSO	0.024	−0.427	0.780	−0.082	−1.305	−0.004	−0.033	1.15e-08	–	0.925	19348
Full	PCA_{AIC}	−0.445	−0.464	0.799	−0.064	−0.479	0.040	0.000	5.27e-07	0.036	0.933	19727
Simplified	PCA_{AIC}	−0.474	−0.464	0.799	−0.080	−0.439	0.041	−0.003	6.95e-07	–	0.933	19727
Full	PLS_{stop}	−0.520	−0.403	0.779	−0.030	−0.598	0.017	−0.004	−2.87e-06	0.091	0.926	19727
Simplified	PLS_{stop}	−0.594	−0.403	0.779	−0.071	−0.497	0.019	−0.013	−2.44e-06	–	0.925	19727
Full	PLS_{LASSO}	0.002	−0.409	0.775	−0.026	−1.334	−0.034	0.001	6.16e-07	0.085	0.933	19596
Simplified	PLS_{LASSO}	−0.072	−0.409	0.775	−0.065	−1.236	−0.032	−0.007	1.05e-06	–	0.932	19596
Full	PCA_{CV}	−0.231	−0.357	0.770	−0.151	−0.371	0.001	0.010	−5.08e-06	0.147	0.889	29024
Simplified	PCA_{CV}	−0.351	−0.357	0.770	−0.218	−0.209	0.004	−0.003	−4.39e-06	–	0.887	29024

Abbreviations: Full, includes all predictors; Simplified, does not contain ES (upper quartile of absolute predictor effect); interc, intercept; EPV, events per variable; event frac, empirical event fraction; *p*, number of candidate predictors; AUC, empirical area under the ROC curve; sparse, sparse predictors present (yes/no); VIF, variance inflation factor; ES, upper quartile of predictor effects; R^2, cross-validated R^2; *n*, number of complete cases that were used for the derivation of the metamodel; MLE, binary logistic regression with maximum likelihood estimation; LASSO, least absolute shrinkage and selection operator; PCA_{AIC}, incomplete principal component regression with components that correspond to minimal AIC; PLS_{stop}, incomplete partial least squares regression with stopping criterion; PLS_{LASSO}, LASSO regression on PLS components; PCA_{CV}, incomplete principal component regression with components that correspond to minimal CV error.

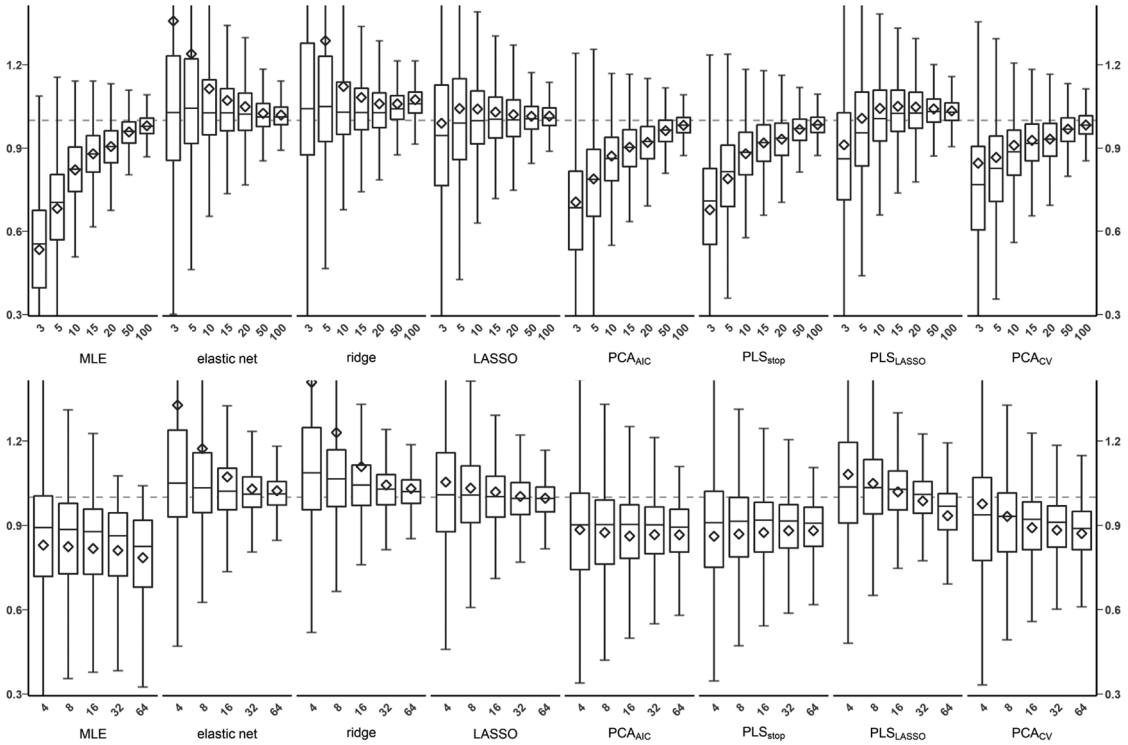

FIGURE 4 Boxplots of calibration slope per modeling approach and simulation factor (top: expected EPV, bottom: number of candidate predictors). Diamonds indicate mean performance. Whiskers extend to the third quartile plus $1.5 \times IQR$ as well as the first quartile minus $1.5 \times IQR$. Outliers are omitted due to their large number. MLE, binary logistic regression with maximum likelihood estimation; LASSO, least absolute shrinkage and selection operator; PCA_{AIC}, incomplete principal component regression with components that correspond to minimal AIC; PLS_{stop}, incomplete partial least squares regression with stopping criterion; PLS_{LASSO}, LASSO regression on PLS components; PCA_{CV}, incomplete principal component regression with components that correspond to minimal CV error.

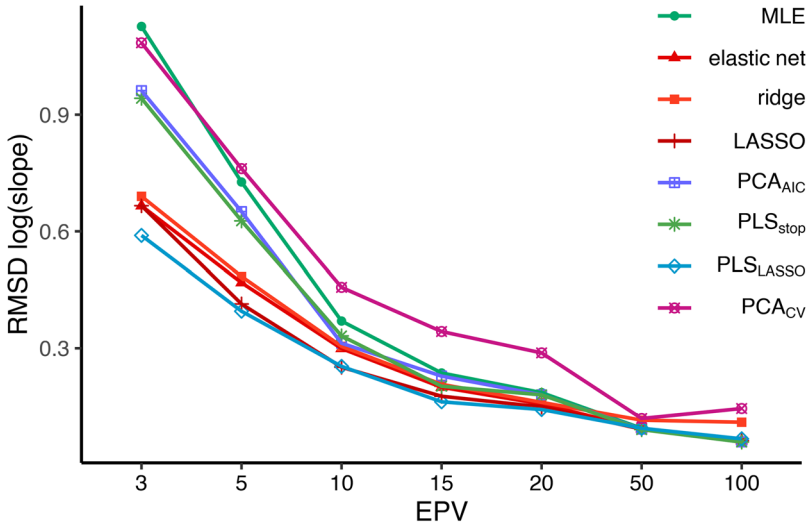

FIGURE 5 Root mean squared distance (RMSD) of the logarithm of the calibration slope. EPV, events per variable; MLE, binary logistic regression with maximum likelihood estimation; LASSO, least absolute shrinkage and selection operator; PCA_{AIC}, incomplete principal component regression with components that correspond to minimal AIC; PLS_{stop}, incomplete partial least squares regression with stopping criterion; PLS_{LASSO}, LASSO regression on PLS components; PCA_{CV}, incomplete principal component regression with components that correspond to minimal CV error.

up to $r = 0.56$. The final model reported by Zhang and colleagues included eight of the 12 candidate predictors (race, age, sex, body mass index, smoking status, total cholesterol (TC), diabetes, hypertension, pulse pressure, TC/high density lipoprotein (HDL) ratio, HbA1c, and HDL) and reported a c-statistic of 0.82 (95% confidence interval (CI) 0.82, 0.83) in the derivation set and 0.76 (95% CI 0.72, 0.79) in the independent validation set. The original derivation sample size was $N = 6059$ (NHANES cohorts 1999–2000 and 2001–2002, 491 cases of peripheral arterial disease), the validation sample size was $N = 3086$ (NHANES cohorts 2003–2004, 322 cases of peripheral arterial disease).

In addition to the original derivation dataset, we repeated model derivation with as a smaller set comprising a random sample of $N = 600$ (50 cases of peripheral arterial disease) from the derivation dataset. With the smaller sample, we mimic a setting with a lower EPV of 5.6. Derivation of prediction models was carried out in the same way as with the simulated data.

The general pattern of performance across modeling approaches in this example resembles the result from the simulations. Penalization methods outperformed variance decomposition approaches as well as MLE regarding c-statistic and calibration slope in the original-sized sample. The noteworthy exception to this trend is the PLS–LASSO combination which achieved the highest average ranking in the full sample regarding calibration slope. None of the regression approaches evaluated in the present study was able to match the original c-statistic from the literature. The MAPE was slightly above the MLE performance for penalization methods and combined approaches, whereas variance decomposition methods exhibited a slight improvement compared to MLE. Cross-validated PCA exhibited lower average MAPE as well as lower variability of the same. The other performance outcomes did not vary noticeably across approaches for the full sample. The highest mean out-of-sample c-statistics for the original sample were obtained by LASSO (0.74, 95% CI 0.71, 0.76) with most other techniques showing overlapping intervals (Figure C in the Supporting Information). In line with findings from the simulation study, performance differences were more pronounced in the 10% random sample (Figure C.2 in the Supporting Information), especially regarding calibration slope. The MLE-derived model exhibited an average calibration slope of 0.13 (standard error = 0.0086) in 100 bootstrap samples of the validation set, which is a profoundly worse performance compared to the top-ranking approach (PCA_{CV}) with a mean calibration slope of 0.97 (standard error = 0.0098). Despite the favorable calibration slopes obtained using cross-validated PCA, the method scored worst or among the worst on all other performance measures illustrating the importance and difficulty of evaluating model performance across various indicators.

More details can be found in Supporting Information C.

6 | DISCUSSION

In this study, we considered various likelihood penalization and variance decomposition approaches for the derivation of clinical risk prediction models. Our simulation study showed that deriving prediction models in low dimensional settings with penalized likelihood regression, variance decomposition techniques, and combinations of them generally improves the average predictive performance as compared to regular MLE-based logistic regression, especially in settings with smaller sample size, and/or a larger number of candidate predictor variables.

Predictive performance differences between modeling strategies were most pronounced for the calibration slope outcome. This is in line with previous simulation studies that focused on likelihood penalization approaches only (Ambler et al., 2012; Steyerberg et al., 2001; Van Smeden et al., 2019). In particular, in small sample size settings, we found likelihood penalization approaches to be more effective than the variance decomposition approaches in producing models with an average calibration slope close to ideal, that is a value of 1. However, as also shown elsewhere (Van Calster et al., 2020), the variation in performance in small datasets shows that the out-of-sample predictive performance can be poor despite approaches that aim to circumvent model overfitting such as likelihood penalization and variance decomposition. This supports the notion that the development of usable risk prediction models requires an appropriate sample size (Ogundimu et al., 2016; Riley et al., 2019; Van Smeden et al., 2019).

Likelihood penalization approaches often occupied top-ranking positions and were generally more stable in average performance across various performance indicators. In particular, the elastic net approach showed the highest average ranking, albeit the performance differences between modeling approaches often being relatively small compared to the variation between simulation conditions. Variance decomposition methods seemed to benefit most from a combination with the LASSO especially with regard to the calibration slope whereas ridge regression generally did not yield favorable performance of calibration slopes. We further investigated the influence of data characteristics on the predictive performance using simulation metamodels. Low EPV, the presence of noise predictors, and the sparsity of predictors was shown to consistently negatively affect out-of-sample performance under all modeling approaches. MLE-based models were affected most by these properties of the data, with penalization methods showing better performance on average as well as less variation between the simulation iterations.

In line with previous findings (Van Smeden et al., 2019), the application of ridge regression tended to produce calibration slopes above 1 on average, indicating a tendency to produce underfitted prediction models. Elastic net, which had not been the subject of investigation in the risk prediction literature as frequently, exhibited similar properties. The same holds for

the LASSO albeit to a much lesser degree. For ranking performance, we treated the distance from a perfect calibration slope (slope of 1) as symmetrical. We acknowledge that in practice this symmetry is usually not a given. The effects of overfitted risks (too extreme predictions, calibration slopes < 1) or underfitted risks (not extreme enough, calibration slopes > 1) may vary and, dependent on the context. Hence, different costs to the miscalibration may be involved.

Some limitations apply to our study. We investigated 1050 low-dimensional simulation scenarios where we aimed to mimic realistic risk prediction modeling settings. However, as with any simulation study, the number of simulation settings remains finite and generalization far beyond the simulation conditions that we have studied is not advised. Also, the present study hence focused on more frequently discussed regression penalization and variance decomposition approaches for which software implementations to derive binomial logistic regression models are available without too much burden on the part of the researcher. Several techniques were implemented with default settings because the computational complexity of the simulation did not allow for further hyperparameter tuning. In practice, we expect that this tuning may sometimes further improve the performance. The effect of hyperparameter tuning in low-dimensional small sample setting deserves further investigation.

In addition to limited hyperparameter tuning, the automated nature of simulation studies has additional shortcomings that make it difficult to assess performance. Deranged models with "extreme" properties such as no predictors or very extreme predictor weights would be easily spotted in practice. For an automated assessment, any cutoff for such extremeness is arbitrary. In order to provide a realistic and neutral comparison, we aimed at methods to neither suffer nor benefit from inherent modeling properties that would have been easily spotted in practice.

In conclusion, the likelihood penalization ridge, LASSO, and elastic net performed well across a wide range of low-dimensional prediction modeling settings, and in most cases occupied the top ranks in our simulations. The variance decomposition methods were mostly inferior to likelihood penalization approaches but in many cases performance difference was minor. However, despite the fact that most approaches investigated in the present study showed an improved average performance compared to MLE, all prediction modeling approaches exhibited a large performance variability. Hence, the application of any modeling technique for deriving risk prediction models from datasets with a relatively small sample size (n) and/or a large number of parameters (p) should be viewed with skepticism.

ACKNOWLEDGMENTS

RHHG was funded by the Netherlands Organization for Scientific Research (ZonMW-Vidi project 917.16.430) and a LUMC fellowship.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

None of the authors declare any conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on the Open Science Framework at https://osf.io/gcjn6. These data were partially derived from resources available in the public domain listed in the Supporting Information.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Anna Lohmann https://orcid.org/0000-0002-4004-4265
Rolf H. H. Groenwold https://orcid.org/0000-0001-9238-6999
Maarten van Smeden https://orcid.org/0000-0002-5529-1541

REFERENCES

Ambler, G., Seaman, S., & Omar, R. Z. (2012). An evaluation of penalised survival methods for developing prognostic models with rare events. *Statistics in Medicine*, *31*(11–12), 1150–1161. https://doi.org/10.1002/sim.4371

Bastien, P., Vinzi, V. E., & Tenenhaus, M. (2005). PLS generalised linear regression. *Computational Statistics & Data Analysis*, *48*(1), 17–46. https://doi.org/10.1016/j.csda.2004.02.005

Boulesteix, A.-L., Binder, H., Abrahamowicz, M., Sauerbrei, W., & for the Simulation Panel of the STRATOS Initiative. (2018). On the necessity and design of studies comparing statistical methods. *Biometrical Journal*, *60*(1), 216–218. https://doi.org/10.1002/bimj.201700129

Bouwmeester, W., Zuithoff, N. P. A., Mallett, S., Geerlings, M. I., Vergouwe, Y., Steyerberg, E. W., Altman, D. G., & Moons, K. G. M. (2012). Reporting and methods in clinical prediction research: A systematic review. *PLoS Medicine*, *9*(5), e1001221. https://doi.org/10.1371/journal.pmed.1001221

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ*, *350*, g7594–g7594. https://doi.org/10.1136/bmj.g7594

Cook, R. D. (2007). Fisher Lecture: Dimension reduction in regression. *Statistical Science*, *22*(1), 1–26. https://doi.org/10.1214/088342306000000682

Frank, I. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*(2), 109–135. https://doi.org/10.2307/1269656

Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). Springer.

Harrell, F. E., Jr, Lee, K. L., Califf, R. M., Pryor, D. B., & Rosati, R. A. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, *3*(2), 143–152.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, *17*(4), 315. https://doi.org/10.2307/1165127

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

Hastie, T., Tibshirani, R., & Tibshirani, R. J., 2017. *Extended comparisons of best subset selection, forward stepwise selection, and the Lasso*. arXiv. https://doi.org/10.48550/arXiv:1707.08692

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Kleijnen, J. P., & Sargent, R. G. (2000). A methodology for fitting and validating metamodels in simulation. *European Journal of Operational Research*, *120*(1), 14–29. https://doi.org/10.1016/S0377-2217(98)00392-0

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *41*(1), 191–201. https://doi.org/10.2307/2347628

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, *52*(1), 374–393. https://doi.org/10.1016/j.csda.2006.12.019

Meyer, N., Maumy-Bertrand, M., & Bertrand, F. (2010). Comparaison de de régressions logistiques PLS et de régression PLS sur variables qualitatives : Application aux donnees d'allelotypage [Comparing the linear and the logistic PLS regression with qualitative predictors: Application to allelotyping data]. *Journal de la Societe Francaise de Statistique*, *151*, 1–18.

Miller, M. E., Langefeld, C. D., Tierney, W. M., Hui, S. L., & McDonald, C. J. (1993). Validation of probabilistic predictions. *Medical Decision Making*, *13*(1), 49–57. https://doi.org/10.1177/0272989X9301300107

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., & Collins, G. S. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine*, *162*(1), W1–W72. https://doi.org/10.7326/M14-0698

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods: Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*, 2074–2102. https://doi.org/10.1002/sim.8086

Ogundimu, E. O., Altman, D. G., & Collins, G. S. (2016). Adequate sample size for developing prediction models is not simply related to events per variable. *Journal of Clinical Epidemiology*, *76*, 175–182. https://doi.org/10.1016/j.jclinepi.2016.02.031

Pavlou, M., Ambler, G., Seaman, S., De Iorio, M., & Omar, R. Z. (2016). Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in Medicine*, *35*(7), 1159–1177. https://doi.org/10.1002/sim.6782

Puhr, R., Heinze, G., Nold, M., Lusa, L., & Geroldinger, A. (2017). Firth's logistic regression with rare events: Accurate effect estimates and predictions? *Statistics in Medicine*, *36*(14), 2302–2317. https://doi.org/10.1002/sim.7273

R Core Team. 2019. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell, F. E., Jr, Moons, K. G., & Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: PART II - Binary and time-to-event outcomes. *Statistics in Medicine*, *38*(7), 1276–1296. https://doi.org/10.1002/sim.7992

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, *39*(5), 1–13.

Steyerberg, E. W., Borsboom, G. J. J. M., van Houwelingen, H. C., Eijkemans, M. J. C., & Habbema, J. D. F. (2004). Validation and updating of predictive logistic regression models: A study on sample size and shrinkage. *Statistics in Medicine*, *23*(16), 2567–2586. https://doi.org/10.1002/sim.1844

Steyerberg, E. W., Eijkemans, M. J. C., Harrell, F. E., & Habbema, J. D. F. (2001). Prognostic modeling with logistic regression analysis: In search of a sensible strategy in small datasets. *Medical Decision Making*, *21*(1), 45–56. https://doi.org/10.1177/0272989X0102100106

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, *21*(1), 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., Bossuyt, P., Collins, G. S., Macaskill, P., McLernon, D. J., Moons, K. G. M., Steyerberg, E. W., Van Calster, B., van Smeden, M., Vickers, A., & On behalf of topic group "Evaluating diagnostic tests and prediction models" of the STRATOS initiative. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, *17*(1), 230. https://doi.org/10.1186/s12916-019-1466-7

Van Calster, B., van Smeden, M., De Cock, B., & 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. https://doi.org/10.1177/0962280220921415

van Smeden, M., Moons, K. G., de Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J., & Reitsma, J. B. (2019). Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*, *28*(8), 2455–2474. https://doi.org/10.1177/0962280218784726

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M. J., Dahly, D. L., Damen, J. A., Debray, T. P. A., de Jong, V. M. T., De Vos, M., Dhiman, P., Haller, M. C., Harhay, M. O., Henckaerts, L., Heus, P., Kammer, M., Kreuzberger, N., … van Smeden, M. (2020). Prediction models for diagnosis and prognosis of Covid-19: Systematic review and critical appraisal. *BMJ*, *369*, m1328. https://doi.org/10.1136/bmj.m1328

Zhang, Y., Huang, J., & Wang, P. (2016). A prediction model for the peripheral arterial disease using NHANES data. *Medicine*, *95*(16), e3454. https://doi.org/10.1097/MD.0000000000003454

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Lohmann, A., Groenwold, R. H. H., & van Smeden, M. (2024). Comparison of likelihood penalization and variance decomposition approaches for clinical prediction models: A simulation study. *Biometrical Journal*, *66,* 2200108. https://doi.org/10.1002/bimj.202200108