Comparison of complex modeling strategies for prediction of a binary outcome based on a few, highly correlated predictors

Marco Chiabudini1,2 | Martin Schumacher1 | Erika Graf1

1Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany
2iOMEDICO AG, Freiburg, Germany

Correspondence
Erika Graf, Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Stefan-Meier-Straße 26, 79104 Freiburg, Germany.
Email: egr@imbi.uni-freiburg.de

Abstract
Motivated by a clinical prediction problem, a simulation study was performed to compare different approaches for building risk prediction models. Robust prediction models for hospital survival in patients with acute heart failure were to be derived from three highly correlated blood parameters measured up to four times, with predictive ability having explicit priority over interpretability. Methods that relied only on the original predictors were compared with methods using an expanded predictor space including transformations and interactions. Predictors were simulated as transformations and combinations of multivariate normal variables which were fitted to the partly skewed and bimodally distributed original data in such a way that the simulated data mimicked the original covariate structure. Different penalized versions of logistic regression as well as random forests and generalized additive models were investigated using classical logistic regression as a benchmark. Their performance was assessed based on measures of predictive accuracy, model discrimination, and model calibration. Three different scenarios using different subsets of the original data with different numbers of observations and events per variable were investigated. In the investigated setting, where a risk prediction model should be based on a small set of highly correlated and interconnected predictors, Elastic Net and also Ridge logistic regression showed good performance compared to their competitors, while other methods did not lead to substantial improvements or even performed worse than standard logistic regression. Our work demonstrates how simulation studies that mimic relevant features of a specific data set can support the choice of a good modeling strategy.

KEYWORDS
correlated predictors, elastic net, random forests, ridge regression, risk prediction models

1 | INTRODUCTION

A common aim in the analysis of clinical data sets is the development of models that allow prediction of an outcome of interest based on a set of given covariates. In the case of a binary outcome, the method of choice has been for a long time logistic
regression, sometimes combined with some variable selection strategy. A regression model is fitted to the available data (training data) and the regression coefficients are estimated. In new data these regression coefficients can then be used to predict the outcome. However, such models often perform well in the training data but may display a poor performance when applied to new data. Obviously, for a prediction model predictive performance in new data is what matters when it comes to application of the model in clinical practice. Predictive performance of a risk prediction model is often measured in terms of predictive accuracy, discrimination, and calibration (Steyerberg et al., 2010). The assessment of a model’s predictive performance is called model validation and can be done either internally or externally or both (Steyerberg, 2008). For binary outcomes, common measures to quantify the predictive performance are the Brier score (BS, Brier, 1950), the area under the receiver-operator characteristic curve (AUC, or C-Index), and the calibration slope (CS). BS measures predictive accuracy in terms of a mean squared error, AUC reflects the model’s discriminative ability, and CS is a measure for calibration.

One possible reason for poor performance in independent data is overfitting which means that the model fits the training data too well. An overfitted model not only captures the underlying process in the data but also some random noise in the training data (e.g., Hastie, Tibshirani, & Friedman, 2009; Steyerberg, 2008) and thus often has poor predictive performance in independent data. A simple way to address the problem of overfitting is shrinkage of the regression coefficients by a factor. The shrinkage factor can be interpreted as an estimate of the amount of overfitting (Harrell, 2015). Thus, shrinkage of coefficients is a way to estimate and counteract overfitting. The probably most widely used way of regression coefficient shrinkage is different versions of penalized regression where selection of the shrinkage factor is part of the modeling process (Harrell, 2015; Hastie et al., 2009). Usually methods like cross-validation or the bootstrap are used to determine an optimal penalty factor (i.e. the amount of shrinkage needed).

As stated above, a standard method to build a prediction model for a binary outcome is logistic regression. In logistic regression \( \logit(p_i) \) is modeled using a linear predictor. Therefore, a logistic model can be written as follows

\[
\logit(p_i) = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} \quad i = 1, \ldots, n \text{ and } k = 1, \ldots, p
\]

where \( p_i = P(Y_i = 1 \mid X_i = x_i). \) \( X_i \) is a vector of covariates drawn from an independent identically distributed sample of size \( n (i = 1, \ldots, n), Y_i \) is the outcome encoded with 0 and 1, \( \beta_0 \) is the intercept, \( \beta_k \) are the regression coefficients, and \( p \) is the number of predictors. Fitting a logistic model involves maximizing the logistic log-likelihood

\[
l(\beta_0, \beta) = \sum_{i=1}^{n} y_i \left( \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} \right) - \log \left( 1 + \exp \left( \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} \right) \right).
\]

In penalized regression, the likelihood is augmented by a penalty term, thus maximum likelihood estimation of the regression coefficients \( \beta_k \) involves maximization of the penalized likelihood

\[
l_{\lambda}(\beta_0, \beta) = l(\beta_0, \beta) - \lambda \Omega(\beta)
\]

where \( \lambda \) is the penalty parameter and \( \Omega(\beta) \) some penalty term. With \( \Omega(\beta) = \sum_{k=1}^{p} \beta_k^2 \) this becomes the Ridge penalty (L2 norm, Hoerl et al., 1970), with \( \Omega(\beta) = \sum_{k=1}^{p} |\beta_k| \) it is the Lasso penalty (L1 norm, Tibshirani, 1996), and with \( \Omega(\beta) = \sum_{k=1}^{p} ((1 - \alpha)\beta_k^2 + \alpha |\beta_k|) \) the Elastic Net penalty (a combination of L1 and L2 norm where \( \alpha \) is an additional tuning parameter (Zou et al., 2005). One way to optimize the tuning parameters \( \lambda \) and \( \alpha \) is cross-validation. This is done by minimizing the cross-validated deviance over a grid of \( \lambda \) and \( \alpha \) values. The advantage of penalized logistic regression over standard logistic regression is less overfitting due to shrink estimated regression coefficients \( \hat{\beta}_k \) and optimization of the tuning parameters using cross-validation. Additionally, penalized regression methods can also be applied in situations where usual logistic regression models fail to converge (e.g., if the number of predictors approaches or exceeds the number of available observations, \( p \approx n \) or \( p > n \)). While Ridge regression always retains all possible predictors in the tuned model and just shrinks their regression coefficients toward zero (Hoerl et al., 1970; Le Cessie et al., 1992), Lasso and Elastic Net regression can set some of the regression coefficients to exactly zero and thus can perform automated selection of predictors (Tibshirani, 1996; Zou et al., 2005). Consequently, Lasso and Elastic Net models tend to produce less complex prediction models. In terms of predictive performance Ridge and Elastic Net models have been shown to perform well in the presence of correlated predictors, while Lasso models have advantages in the presence of many noise predictors (Pavlou, Ambler, Seaman, De Iorio, & Omar, 2016; Zou et al., 2005).

One drawback of the penalized regression methods discussed so far is that they allow only for additive linear predictors. However, for the prediction of a clinical outcome solely additive predictor effects might not be sufficient since predictors might interact with each other. The term interaction refers to a situation in which the influence of a predictor depends on the value
of one or more other predictors. Usually such interaction effects are modeled by including interaction terms into the model. However, including possible interactions into a model can be challenging since already with a small number \( p \) of predictors there are \( \binom{p}{2} \) interactions of order 1, thus increasing the predictor space. Usually an interaction is included in a model only if all corresponding main effects are included (strong hierarchy) or at least one corresponding main effect (weak hierarchy). A Lasso model including interactions that fulfill the constraint of strong hierarchy or weak hierarchy has been presented by Bien, Taylor, and Tibshirani (2013). This Lasso model combines variable selection with automated selection of interactions under the constraint of a strong hierarchy or a weak hierarchy.

Another possible issue when using standard logistic regression or penalized logistic regression is the question of the functional form of a variable. One possibility to address this issue is to include a set of transformations of the original predictors into the new set of predictors and use the variable selection property of Lasso regression or the Elastic Net. However, the decision whereby transformations are included is rather arbitrary although it might be based on clinical considerations. Yet another way to reflect different functional forms of a variable is generalized additive models (GAM, Hastie et al., 1986, 1990). GAMs are an extension of generalized linear models in which the linear components \( \beta_k x_k \) are replaced with a non-linear function \( f_k(x_k) \). In case of a logistic GAM the model can be written as follows

\[
\logit(p_i) = \beta_0 + \sum_{k=1}^{p} f_k(x_{ik}).
\]

Usually, the \( f_k(x_k) \) are smooth functions with the degree of smoothness determined by either generalized cross-validation or restricted maximum likelihood. Thus, GAMs take into account that the influence of a predictor is not necessarily linear. Besides the regression-based methods presented so far, there are also classification methods that do not rely on an explicitly given model. Such methods have been discussed as alternatives to regression-based techniques for classification tasks (e.g., Kruppa et al., 2014; Van der Ploeg, Austin, & Steyerberg, 2014; Van der Ploeg, Smits, Dippel, Hunink, & Steyerberg, 2011). One example is tree-based classifications. In a classification tree, class labels are assigned based on a series of splits along the tree. The splits are based on the predictive covariates. Since such classification trees are very sensitive to small changes in the data and often perform poorly when applied to new data, improvements using more than one classification tree and aggregating the trees have been introduced. Such an improvement is so-called random forests (RFs, Breiman, 2001). Instead of growing just one tree, in an RF hundreds of trees are grown. Additionally, for each split only a bootstrapped subset of predictors of size \( \sqrt{p} \) is drawn. This leads to a decorrelation of the single trees. RFs are generally believed to have low tendency for overfitting (Breiman, 2001).

Motivated by a clinical prediction problem, the aim of our study was to compare various modeling strategies for the prediction of a binary outcome in the presence of only a few, highly correlated predictors which, however, could exhibit dynamic, non-linear, as well as non-additive effects. Thus, the history of changes of predictors over time, non-linear terms and interaction terms were to be considered in addition to the predictors themselves. Using simulated data sets based on a clinical data set of patients with acute heart failure, we compared several strategies to produce reliable probabilistic predictors of hospital survival under assumptions that were as realistic as possible. The predictive performance of the different modeling techniques was compared with respect to their predictive accuracy (BS), discrimination (AUC), and calibration (CS). We compared approaches that include transformations and interactions of the original blood parameters based on clinical considerations as additional predictors with approaches that use only the original measurements. To the approaches including additional predictors we applied Elastic Net logistic regression, different penalized versions of logistic regression and RFs, to those using only the original measurements we applied Lasso logistic regression with automated selection of interactions and GAMs. As a benchmark we used also standard logistic regression utilizing only the original blood parameters as predictors. Section 2 describes the simulation strategy and the clinical data set in more detail, Section 3 summarizes the results of the simulation, and we conclude with a discussion of the results in Section 4.

### 2 | SIMULATION STUDY

We performed a simulation study in order to compare different strategies for building a prediction model. The simulation comprised the following steps:

1. Investigation of covariate structure in clinical example data to be mimicked in the simulation
2. Generation of training and validation data sets containing simulated covariates and outcomes, for each out of a range of reference models (see Section 2.2) fitted to the original data as true models
3. Fit of models to the training data using the different modeling strategies
4. Examination of the predictive performance (BS, AUC, and CS) in the validation data

### 2.1 Motivating clinical prediction problem: the ECLS data set

Our simulation study was based on a clinical data set of patients who required treatment with an extra-corporal life support (ECLS, “heart-lung machin”) due to acute heart failure. In the analysis of this data set, referred to as “ECLS data,” the original aim was to build a prediction model which would reliably quantify a patient’s probability of hospital survival, defined as the probability of leaving the hospital alive. So patients who died afterward were not counted as deaths. This was judged to be a more clinically relevant outcome than survival time because patients were admitted in an acute emergency situation which they would either survive or not (Schoenfeld, 2006).

Since the main goal of ECLS is to establish an adequate organ perfusion, the model should base the prediction exclusively on a limited set of blood parameters (pH, lactate, and standard bicarbonate [SBC]) which were measured repeatedly during the use of the ECLS. So further covariates, even if available, were not considered as candidate predictors (Wengenmayer et al., 2018). The blood parameters were thought to interact in an unknown, complex manner and with a high overall impact on survival. The aim of the modeling procedure was reliable prediction in independent data, hence predictive ability had priority over interpretability of the final model.

The full data set contained 205 patients of whom 61 survived hospital (\(n_{\text{surv}} = 61\)). Measurements of the blood parameters were available for the 1, 6, 12, and 24 hr time points. This data structure allowed to base predictions on patients who survived different periods of time. Separate prediction models were developed using measurements after 1 and 6 hr in patients who had survived at least 6 hr (“6 hr survivors,” \(n = 187\)), after 1, 6, and 12 hr (“12 hr survivors,” \(n = 170\)), or after 1, 6, 12, and 24 hr (“24 hr survivors,” \(n = 145\)), respectively. To reflect the presumed unknown, complex impact on survival, besides the original blood parameters, transformations and interactions of the initial measurements were also included in the set of candidate predictors.

These were determined based on a combination of clinical and statistical considerations. For example, the clinicians thought that specific values of blood parameters might be relevant only if found to lie above their normal range, giving rise to inclusion of transformed predictor variables that kept the original values if these were above normal range but were set equal to the upper limit otherwise. An overview of all derived variables is given in Table 1. In total, there were \(p = 64\) (6 hr survivors), \(p = 100\) (12 hr survivors), and \(p = 136\) (24 hr survivors) candidate predictors, respectively. The original values were highly correlated. For example, only one out of a total of 15 absolute linear correlation coefficients was below 0.4 in the data of 6 hr survivors (Table S1 in Supporting Information). Even higher correlations occurred in the full set of predictors including transformations and interaction terms.

These characteristics motivated the use of penalized logistic regression. In particular, an Elastic Net approach was chosen due to its alleged predictive power in the presence of many weak, correlated predictors and variable selection properties (Zou et al., 2005). Selection of the tuning parameters for the Elastic Net was performed using repeated tenfold cross-validation. The predictive performance was pre-assessed internally by naive estimation from the entire development data set and by repeated nested tenfold cross-validation on different subsets of the original data with ten repetitions, that is, from 100 models in total. Two prediction models based either on the 6 or 12 hr data were developed and subsequently subjected to external validation (Wengenmayer et al., 2018; in contrast with the work presented here, Wengenmayer et al. did not use the 1 hr data for 6 hr survivors).

For the 6 hr data, model discrimination was assessed as AUC = 0.823 from the entire derivation cohort, as median AUC = 0.811 based on internal cross-validation in the derivation cohort, and as AUC = 0.718 from an external validation cohort. Predictive accuracy was assessed as BS = 0.161 from the entire development cohort, as median BS = 0.169 based on internal cross-validation and as BS = 0.222 by external validation (AUC = 0.839, 0.800, 0.735; BS = 0.158, 0.176, 0.209 for the 12 hr data, respectively). These results based on the Elastic Net were considered satisfactory by our clinical partners.

In the simulation study, we strived to investigate the performance of the Elastic Net and the competitive modeling strategies described in the introduction in situations as exemplified by our clinical data. Therefore, the simulated data used to assess the respective performance measures should be generated from a range of true models that would mimic essential features of the motivating clinical data set.

### 2.2 Building the reference models

Some modeling techniques to be investigated used only the basic blood parameters (i.e., original, untransformed measurements of pH, lactate, and SBC; referred to as base ECLS data sets), others used the full predictor set (basic blood parameters and
### Table 1: Overview of transformed and derived variables

| Variable name | Description |
|---------------|-------------|
| pH_2          | Squared pH  |
| pH_3          | Cubic pH    |
| pH_log        | Logarithm of pH |
| lact_log      | Logarithm of lactate |
| lact_loglog   | Double logarithm of lactate |
| SBC_log       | Logarithm of SBC |
| SBC_loglog    | Double logarithm of SBC |
| pH_belonorm   | pH dichotomized at lower normal limit |
| pH_abovnorm   | pH dichotomized at upper normal limit |
| lact_abovnorm | Lactate dichotomized at upper normal limit |
| SBC_belonorm  | SBC dichotomized at lower normal limit |
| SBC_abovnorm  | SBC dichotomized at upper normal limit |
| pH_contabovnorm | pH cut at upper norm limit; values below upper limit set to 7.450 |
| pH_contbelonorm | pH cut at lower norm limit; values above lower limit set to 7.350 |
| lact_mmolplabovnorm | Lactate cut at upper norm limit; values below upper limit set to 2.2 mmol/L. |
| SBC_mmolplabovnorm | SBC cut at upper norm limit; values below upper limit set to 27.0 mmol/L. |
| SBC_mmolplbelonorm | SBC cut at lower norm limit; values above lower limit set to 23.0 mmol/L. |
| pHabov_lactabov_cont | Interaction of pH_contabovnorm and lact_mmolplabovnorm |
| pHbelo_lactabov_cont | Interaction of pH_contbelonorm and lact_mmolplabovnorm |
| pHabov_lactabov_yrn | Interaction (binary) of pH_abovnorm and lact_abovnorm |
| pHbelo_lactabov_yrn | Interaction (binary) of pH_belonorm and lact_abovnorm |
| pHabov_SBC_abov_cont | Interaction of pH_contabovnorm and SBC_mmolplabovnorm |
| pHbelo_SBC_abov_cont | Interaction of pH_contbelonorm and SBC_mmolplabovnorm |
| pHabov_SBC_belo_cont | Interaction of pH_contabovnorm and SBC_mmolplbelonorm |
| pHbelo_SBC_belo_cont | Interaction of pH_contbelonorm and SBC_mmolplbelonorm |
| pH_1to6h_contabovnorm | Change from first to second time point of pH_contabovnorm |
| pH_1to6h_contbelonorm | Change from first to second time point of pH_contbelonorm |
| pH_1to6h_abovnorm | Change from first to second time point of pH_abovnorm |
| pH_1to6h_belonorm | Change from first to second time point of pH_belonorm |
| lact_1to6h_mmolplabovnorm | Change from first to second time point of lact_mmolplabovnorm |
| lact_1to6h_abovnorm | Change from first to second time point of lact_abovnorm |
| SBC_1to6h_mmolplelonorm | Change from first to second time point of SBC_mmolplbelonorm |
| SBC_1to6h_belonorm | Change from first to second time point of SBC_belonorm |

*For brevity, information on the time point (e.g., 1 hr, 6 hr) is omitted from the variable name.

*Changes from second to third and from third to fourth time point are included accordingly.

derived predictors; referred to as full ECLS data sets). Therefore, reference models were fitted to six different data sets in total:

- 6 hr survivors: \( n = 187 \) observations, hospital survival: \( 61/187 = 33\% \), \( p_{full} = 64 \) and \( p_{base} = 6 \) predictors (full/base ECLS data set, “6 hr data”)
- 12 hr survivors: \( n = 170 \) observations, hospital survival: \( 61/170 = 36\% \), \( p_{full} = 100 \) and \( p_{base} = 9 \) predictors (full/base ECLS data set, “12 hr data”)
- 24 hr survivors: \( n = 145 \) observations, hospital survival: \( 61/145 = 42\% \), \( p_{full} = 136 \) and \( p_{base} = 12 \) predictors (full/base ECLS data set, “24 hr data”)

Of note, due to deaths between the three time points, the data sets referring to later time points contained fewer observations but more predictor variables, thus the number of observations per variable \( OPV = n/p \) and the number of events (hospital survivors) per variable \( EPV = 61/p \) differed between data sets.
Since the true model cannot be known, we decided to use each of the modeling techniques to be investigated also to build a reference model. By this means we tried to avoid biased results in the sense that some techniques might overperform when applied to data in which the outcomes were generated with a reference model relying on the same modeling technique. Using either the full ECLS data set or the base ECLS data set which contained only the untransformed predictor variables, the following modeling techniques were applied to the original clinical data:

1. Full ECLS data set:
   a. Ridge logistic regression (RLR)
   b. Lasso logistic regression – full (LLR_f)
   c. Elastic Net logistic regression – full (ENLR_f)
   d. Random forests (RF)
2. Base ECLS data set:
   a. Standard logistic regression (SLR)
   b. Elastic Net logistic regression – base (ENLR_b)
   c. Lasso logistic regression with automated selection of interactions – base (LLR_ib)
   d. Logistic regression using generalized additive models (GAM)

For models employing a penalized version of logistic regression (i.e., RLR, LLR, ENLR), the tuning parameters were optimized using ten times repeated ten-fold cross-validation. Briefly, data sets were repeatedly split into different cross-validation folds and in each split the cross-validated deviance was calculated over the cross-validation folds. Next, the cross-validated deviance was averaged over the ten repetitions (for each \( \lambda \) or \( a/\lambda \) combination, respectively), and the minimum of the mean cross-validated deviance was used to select the optimal \( \lambda \) (for RLR and LLR) or the optimal \( a/\lambda \) combination (for ENLR models).

The \( \lambda \) grids consisted of 100 points (equidistant on log scale) with data-dependent range determined by the \texttt{cv.glmnet} function from the \texttt{glmnet} package (Friedman, Hastie, & Tibshirani, 2010; see help material of the \texttt{glmnet} package for details). ENLR models used a grid of \( \alpha \) values ranging from 0 to 1 in steps of size 0.1, thus allowing for the selection of a Ridge model \((\alpha = 0\), only L_2 norm penalty; model retains all possible predictors), or a Lasso model \((\alpha = 1\), only L_1 norm penalty; sparse model with maximal predictor selection), or a combination of the L_1 and L_2 norm penalty \((0 < \alpha < 1\), sparse model with different grades of sparsity).

For models employing RFs, 1000 trees were grown. For SLR, maximum likelihood estimation of the regression coefficients was applied. GAMs were fitted using cubic splines as smooth function with the degree of smoothness determined by restricted maximum likelihood.

Each of these models was subsequently used as reference (“true”) model for the generation of outcomes in the simulated data sets.

2.3 | Using the ECLS data for data simulation

Generation of the simulated data sets (both training and validation data) was performed in two main steps: covariate simulation and outcome generation. First, preliminary covariates were created as combinations of multivariate normal variables in a complex procedure. Then covariates mimicking the covariate distribution in the original data as closely as possible were derived from these preliminary covariates. Second, the reference models (compare Section 2.2) were used to generate the outcome data. For each time point, 500 training data sets of size \( n \) and a validation data set of size \( 50 \times n \) were simulated (6 hr: \( n = 187 \), 12 hr: \( n = 170 \), 24 hr: \( n = 145 \)). Briefly, data simulation comprised the following steps (a more detailed description of the simulation steps is given in Section S3 in Supporting Information):

1. After log-transformation of lactate and an exponential transformation of pH followed by a square transformation, means, variances, and the correlation matrix were determined in each of the three base ECLS data sets.
2. In order to reflect the bimodal distribution of lactate (supplemental figure S1, left panels), a mixture of two normal distributions was fitted separately for each of the \( l \) time points (6 hr: \( l = 2 \), 12 hr: \( l = 3 \), 24 hr: \( l = 4 \)).
3. Using the $2^l$ means and variances from the components of the mix distributions for log-lactate, and the means, variances and correlation structure of step 1 otherwise, a set of $2^l$ multivariate normal distributions (each of dimension $p_{\text{base}}$) was determined.

4. A decision rule was implemented stating how to select one out of the $2^l$ multivariate normal distributions from which a vector of covariates was subsequently simulated. This step realized the mixture distributions needed to reflect bimodality of lactate.

5. After back-transformation of variables lactate and pH, the full predictor sets were calculated (compare Table 1).

Finally, after completion of covariate simulation, binary outcomes were generated. The outcomes $Y_i$ were simulated based on the simulated covariates using the different reference models fitted to the original data: $p_i$ were calculated using the reference models, and the $0/1$-outcomes $Y_i$ were generated using a Bernoulli distribution ($Y_i \sim \text{Bern}(p_i)$). In order to check how well the simulated data mimicked the features of the original data, histograms of the different blood parameters, descriptive statistics, and the correlation matrices were generated. Examples for the 6 hr data are shown in Tables S1 and S2, Figures S2–S4 in Supporting Information. Histograms of the base parameters in 6 hr data sets (original data set and three randomly selected simulated data sets) show good concordance of the simulated data with the original data, thus justifying our simulation approach. This is further supported by comparing the correlation structures and the descriptive statistics in the simulated data and in the original data (Tables S1 and S2 in Supporting Information).

### 2.4 Assessing the predictive performance of the different modeling techniques

Each modeling technique used to generate a reference model (described in Section 2.2) was also applied to build a prediction model in the training data sets. Using the outcomes of one reference model at a time, all eight modeling techniques were applied to model the respective outcomes in the training data. Model fitting was performed as described in Section 2.2, except that for RLR, LLR, and ENLR simple tenfold cross-validation instead of repeated cross-validation was performed to limit computation times. For each modeling technique/reference model combination, a prediction model was fitted to each of the 500 training data sets, that is, 500 prediction models in total for each scenario. Then, each of these 500 prediction models was applied to the covariates in the respective validation data set, predictions $\hat{p}_i$ were calculated and compared to the outcomes $Y_i$ in the validation data. For each modeling strategy predictive accuracy (BS), model discrimination (AUC), and model calibration (CS) were determined. BS measures predictive accuracy in terms of a mean squared error and was calculated as follows (Brier, 1950):

$$BS = \frac{1}{50 \times n} \sum_{i=1}^{50 \times n} (\hat{p}_i - Y_i)^2.$$ 

Therefore, BS indicates the predictive performance of a model in terms of how well the predicted probabilities actually match the real outcome and becomes smaller the better the predictions are. Additionally, BS is a strictly proper scoring function, i.e., one that is optimized only by the true model (Hilden, Habbema, & Bjerregaard, 1978). Two important benchmarks exist for BS which can be easily derived by replacing $\hat{p}_i$ either with the true $p_i$ or with the marginal survival rate $\frac{n_{\text{surv}}}{n}$. The prediction model given by the true $p_i$ results in the lowest (best) possible BS, while the constant predictor given by the marginal survival rate results in the lowest possible BS among non-informative models. Of note, the BS associated with this best non-informative model depends on the marginal survival rate of the specific data set.

AUC indicates how well a model is able to discriminate between cases and non-cases (here: survivors and non-survivors), that is, it tells whether higher scores are more likely to be associated with being a case and vice versa, irrespective of the absolute score. Therefore, AUC is a proper, but not a strictly proper scoring function. AUC is maximized by the $p_i$ (or any transformation of the $p_i$ that preserves their ranking), with a maximum value of 1.0. In contrast to BS, the AUC of the best non-informative model is always 0.5, independent of the marginal survival rate.

CS was obtained by regressing the outcomes $Y_i$ against the predictions $\hat{p}_i$. CS indicates whether a model is over- or underfitted with slopes larger than 1.0 indicating underfitting and slopes smaller than 1.0 indicating overfitting (Steyerberg et al., 2010).

Stability of the simulation results was checked as follows. To ensure that 500 training data sets were sufficient, the three performance measures were calculated from the validation data set for each of the 500 models derived from training data sets, and the resulting means (BS, AUC) and medians (CS) of the first $l$ training data sets were plotted against $l$ for $l \to 500$. To ensure that $50 \times n$ observations were sufficient for validation, the three performance measures were calculated from data sets containing the first $l$ observations of the validation data for each of the 500 models, and the resulting means (BS, AUC) and medians (CS) over the first $l$ observations of the validation data set were plotted against $l$ for $l \to 50 \times n$. Results were found to be sufficiently stable toward the ends of the respective $x$-axes (Figures S12–S75 in Supporting Information).
2.5 | Software

All calculations were performed using R, version 3.2.3 (R Core Team, 2016). RLR, LLR, and ENLR models were fitted using the package glmnet, version 2.0.5 (Friedman et al., 2010). For RFs the package randomForest, version 4.6.12 was used (Liaw et al., 2002). LLR was performed with the package hierNet, version 1.6 (Bien et al., 2014). For GAM models the package mgcv, version 1.8.9 was used (Wood, 2006, 2011). Further R packages that were used include: mixtools, version 1.0.4 (Benaglia, Chauveau, Hunter, & Young, 2009) for mixed distributions, ModelGood, version 1.0.9 (Gerds, 2014) and pROC, version 1.8 (Robin et al., 2011) for calculations of AUC and BS, and MASS, version 7.3.45 (Venables et al., 2002) for data simulation from multivariate normal distributions.

3 | RESULTS OF THE SIMULATION

The characteristics of the six different full/base ECLS data sets in terms of sample size, survival rate, number of candidate predictors, and numbers of observations and events per variable can be summarized as follows:

- 6 hr : n = 187, survival : 33%; p\text{full} = 64, OPV\text{full} = 2.9, EPV\text{full} = 1.0, p\text{base} = 6, OPV\text{base} = 31.2, EPV\text{base} = 10.2
- 12 hr : n = 170, survival : 36%; p\text{full} = 100, OPV\text{full} = 1.7, EPV\text{full} = 0.6, p\text{base} = 9, OPV\text{base} = 18.9, EPV\text{base} = 6.8
- 24 hr : n = 145, survival : 42%; p\text{full} = 136, OPV\text{full} = 1.1, EPV\text{full} = 0.4, p\text{base} = 12, OPV\text{base} = 12.1, EPV\text{base} = 5.1

Of note, while the time point considered and, along with it, the sample size, survival rates, and base predictors stemmed from the clinical prediction problem, the question to be addressed by the simulation study was how to make best use of the corresponding base predictors—both with and without inclusion of the additional derived variables in the full predictor set.

Since the true generating mechanism of the clinical data is unknown, a modeling strategy that would perform well under a wide range of reference models in the simulation study would be preferable. Among the models considered here, all reference models using the full predictor set are characterized by a complex structure regarding the impact of base predictors on the outcome. Among the models using only base predictors, SLR and ENLR\text{b} include only simple main effects. The complexity of LLR\text{b} and GAM models lies in between.

So, the results of the simulation study should be considered per time point under these aspects. It turned out that the 6 and 12 hr data gave roughly similar results that differed from those found in the 24 hr data.

3.1 | Data sets of 6 and 12 hr survivors

For the purpose of comparison across true models, performance measures of the eight modeling strategies were ranked within each reference model (6 hr data: within columns of Tables 2–4), with lower ranks indicating better performance of a modeling strategy for the respective data generating mechanism. The sum of ranks across reference models was calculated for each strategy (6 hr data: sum within rows of Tables 2–4). Ranks of the sum of ranks are shown in the last column of the respective table, with lower ranks indicating better performance across all reference models.

Judging from this, the best performance with respect to BS and AUC irrespective of the underlying reference model was obtained with ENLR\text{b} (rank 1 for both BS and AUC in 6 and 12 hr data), followed by GAM, RLR, and ENLR\text{f} (all ranks between 2 and 4; Tables 2 and 3; Tables S3 and S4 in Supporting Information; see also Figure 1A,B left and middle panels and Figures S4–S10 in Supporting Information). LLR\text{b}, SLR, LLR\text{ib}, and RF followed on ranks 5 to 8 (identical for BS and AUC in 6 and 12 hr data; Tables 2 and 3; Tables S3 and S4 in Supporting Information).

Slight differences were observed when the nature of the reference model was also taken into account. Reference models using only the base predictor set favored two modeling techniques that also used only the base predictor sets (ENLR\text{b} and GAM). If the true model was a more complex one built with the full predictor set, strategies also using the full predictor set (RLR and ENLR\text{f}), but also the base modeling strategy ENLR\text{b} were among the best performers (Tables 2 and 3; Tables S3 and S4 in Supporting Information; ranks within subgroups of true models not shown). As regards calibration measured in terms of CS, LLR, and ENLR approaches shared ranks 1 to 4 in the comparison across all reference models, followed by GAM, RLR, SLR, and RF on ranks 5 to 8 (identical in 6 and 12 hr data; Table 4; Table S5 in Supporting Information; see also left and middle panels in Figure 1C and Figures S5–S11 in Supporting Information).
In summary, penalized logistic regression, especially ENLR<sub>b</sub>, RLR, and also ENLR<sub>f</sub> improved predictive performance compared to the benchmark standard SLR. These improvements came along with reduced overfitting as indicated by CS values closer to 1.0 compared to SLR (Table 4; Table S5 in Supporting Information). GAM was also a good competitor. RF clearly was the worst choice in the 6 and 12 hr data.

### 3.2 Data set of 24 hr survivors

The data set of the 24 hr survivors is characterized by a lower sample size and consequently fewer observations and events per variable compared to the 6 and 12 hr data. In this scenario, modeling techniques using the full predictor set (RLR, ENLR<sub>f</sub>) had better predictive performance in terms of BS and AUC across the entire range of data generating mechanisms than the next best methods ENLR<sub>b</sub>, and GAM, which used only the base predictor set (Tables S6 and S7 in Supporting Information; right panels in Figure 1; Figures S5–S11 in Supporting Information).

Within reference models using only the base predictor set, results were more heterogeneous in the 24 hr data than for the 6 and 12 hr data: SLR and ENLR<sub>b</sub>, the least complex strategies in terms of predictors included, performed particularly badly in data from the GAM reference model, whereas for data generated from SLR and ENLR<sub>b</sub> reference models, the contrary was true—here, ENLR<sub>b</sub> was best in terms of BS and AUC among the eight modeling strategies. ENLR<sub>b</sub> was also best for each of the four base reference models with regard to calibration (Tables S6–S8, Figures S8, S9, and S11 in Supporting Information).

For all reference models built with the full predictor set, RLR was clearly the best choice (rank 1 for BS and AUC) despite the fact that LLR<sub>f</sub> showed somewhat better calibration (Tables S6–S8 in Supporting Information, right panels in Figure 1 and in Figures S5–S7 in Supporting Information). However, RLR displayed poor model discrimination (AUC) and the worst predictive accuracy and model calibration (BS, CS) out of all strategies when the data were generated from the simplest base SLR reference model (Tables S6–S8 in Supporting Information, right panels in Figure S8 in Supporting Information).
### Table 3
Mean AUCs in 6 hr data set

| Modeling strategy | True model | Base | Rank |
|-------------------|------------|------|------|
|                   | Full       | SLR  | ENLR<sub>f</sub> | LLRib |
|                   | RLR        | ENLR<sub>f</sub> | ENLR<sub>b</sub> | GAM    |        |
|       |             |       |       |       |        |        |
| Full            | 0.724 (0.0005) | 0.715 (0.0004) | 0.723 (0.0004) | 0.740 (0.0004) | 0.777 (0.0005) | 0.735 (0.0005) | 0.758 (0.0004) | 0.760 (0.0004) | 2 |
| LLR<sub>f</sub> | 0.712 (0.0006) | 0.708 (0.0006) | 0.712 (0.0006) | 0.733 (0.0006) | 0.778 (0.0006) | 0.738 (0.0006) | 0.751 (0.0006) | 0.759 (0.0006) | 5 |
| ENLR<sub>f</sub> | 0.720 (0.0006) | 0.712 (0.0005) | 0.719 (0.0006) | 0.737 (0.0005) | 0.778 (0.0005) | 0.738 (0.0005) | 0.755 (0.0005) | 0.760 (0.0005) | 4 |
| RF              | 0.679 (0.0008) | 0.673 (0.0008) | 0.679 (0.0008) | 0.727 (0.0007) | 0.747 (0.0006) | 0.698 (0.0007) | 0.729 (0.0007) | 0.728 (0.0007) | 8 |
| Base            | 0.710 (0.0006) | 0.705 (0.0005) | 0.711 (0.0006) | 0.733 (0.0006) | 0.789 (0.0004) | 0.741 (0.0005) | 0.749 (0.0005) | 0.757 (0.0005) | 6 |
| ENLR<sub>b</sub> | 0.718 (0.0004) | 0.712 (0.0004) | 0.719 (0.0004) | 0.740 (0.0004) | 0.790 (0.0004) | 0.746 (0.0004) | 0.755 (0.0004) | 0.762 (0.0004) | 1 |
| LLRib           | 0.704 (0.0009) | 0.703 (0.0008) | 0.706 (0.0008) | 0.728 (0.0009) | 0.780 (0.0007) | 0.738 (0.0007) | 0.755 (0.0007) | 0.760 (0.0007) | 7 |
| GAM             | 0.711 (0.0006) | 0.710 (0.0005) | 0.712 (0.0006) | 0.734 (0.0006) | 0.787 (0.0005) | 0.745 (0.0005) | 0.756 (0.0005) | 0.765 (0.0005) | 3 |
| True mod.       | 0.742 | 0.737 | 0.742 | 0.797 | 0.803 | 0.758 | 0.790 | 0.783 | |

True mod., performance of the respective reference (“true”) model in the validation data.
Standard errors in parentheses. Highest AUCs for each reference model in bold.
Abbreviations: RLR, ridge logistic regression; LLR<sub>f</sub>, lasso logistic regression in full predictor set; ENLR<sub>f</sub>, elastic net logistic regression in full predictor set; RF, random forest; SLR, standard logistic regression; ENLR<sub>b</sub>, elastic net logistic regression in base predictor set; LLRib, lasso logistic regression with interactions in base predictor set; GAM, generalized additive model.

Full, full predictor set; Base, base predictor set.

### Table 4
Median calibration slopes in 6 hr data set

| Modeling strategy | True model | Base | Rank |
|-------------------|------------|------|------|
|                   | Full       | SLR  | ENLR<sub>f</sub> | LLRib |
|                   | RLR        | ENLR<sub>f</sub> | ENLR<sub>b</sub> | GAM    |        |
|                   |             |       |       |       |        |        |        |
| Full            | 1.149 (0.0119) | 1.075 (0.0116) | 1.111 (0.0124) | 1.108 (0.0117) | 1.063 (0.0093) | 1.099 (0.0114) | 1.022 (0.0100) | 1.083 (0.0097) | 6 |
| LLR<sub>f</sub> | 1.061 (0.0121) | 1.036 (0.0115) | 1.030 (0.0122) | 1.056 (0.0113) | 1.064 (0.0092) | 1.095 (0.0122) | 0.997 (0.0106) | 1.074 (0.0107) | 1 |
| ENLR<sub>f</sub> | 1.113 (0.0118) | 1.063 (0.0106) | 1.069 (0.0116) | 1.071 (0.0104) | 1.057 (0.0091) | 1.093 (0.0119) | 0.997 (0.0094) | 1.071 (0.0100) | 3 |
| RF              | 0.647 (0.0026) | 0.611 (0.0027) | 0.634 (0.0025) | 0.742 (0.0028) | 0.756 (0.0030) | 0.673 (0.0025) | 0.716 (0.0031) | 0.721 (0.0023) | 8 |
| Base            | 0.849 (0.0057) | 0.806 (0.0055) | 0.833 (0.0056) | 0.854 (0.0059) | 0.901 (0.0044) | 0.872 (0.0056) | 0.831 (0.0047) | 0.866 (0.0051) | 7 |
| ENLR<sub>b</sub> | 1.089 (0.0099) | 1.045 (0.0100) | 1.074 (0.0099) | 1.082 (0.0095) | 1.049 (0.0068) | 1.083 (0.0090) | 1.016 (0.0075) | 1.045 (0.0075) | 2 |
| LLRib           | 0.928 (0.0118) | 0.922 (0.0116) | 0.924 (0.0114) | 0.896 (0.0113) | 1.015 (0.0106) | 0.999 (0.0125) | 0.914 (0.0105) | 0.981 (0.0104) | 3 |
| GAM             | 0.905 (0.0073) | 0.886 (0.0070) | 0.901 (0.0073) | 0.897 (0.0067) | 0.955 (0.0059) | 0.963 (0.0073) | 0.886 (0.0057) | 0.923 (0.0058) | 5 |
| True mod.       | 1.002 | 0.998 | 1.000 | 1.014 | 1.008 | 1.007 | 0.982 | 1.001 | |
FIGURE 1 Predictive performance of different modeling techniques based on a Ridge reference model. Model building and evaluation of its predictive performance was based on outcomes generated with an RLR reference model. (A) Predictive accuracy. Dashed line represents lowest mean BS. (B) Model discrimination. Dashed line represents highest mean AUC. (C) Model calibration. Dashed line represents optimal CS. For clarity, outliers have been excluded in panel C. For abbreviations see Table 2

In summary, the best modeling strategy in the 24 hr data depended more heavily on the unknown true data generating mechanism than in the 6 and 12 hr data. RLR proved to be a good choice for the more complex true models, but performed poorly in data generated from the SLR model. ENLR$_b$ performed fairly well in data generated from the full set of predictors. Among base reference models, it was the best choice for data generated from the SLR and ENLR$_b$ models, but it performed badly for the GAM reference model. As in the 6 and 12 hr data, RF and LLR$_b$ were still no good competitors. However, in the 24 hr they outperformed the benchmark SLR, which suffered from strong overfitting in a number of instances. Again, penalized logistic regression approaches reduced overfitting compared to SLR, resulting in underfitting models in some cases, in particular for RLR.

3.3 | Sparsity of ENLR models

We were also interested in the behavior of the ENLR approaches in a more detailed manner, that is, we wanted to have a closer look at the sparsity of the selected models. Sparsity of the selected models is reflected by the corresponding $\alpha$ value, with $\alpha = 0$ reflecting a full model that retained all possible predictors (a Ridge model) and with $\alpha = 1$ reflecting a maximally sparse model.
FIGURE 2 Selected values of $\alpha$. Histograms of the selected $\alpha$ values for both ENLR-based approaches and the different data sets. Shown are absolute frequencies aggregated over the different reference models. The absolute frequencies of each $\alpha$ are given on top of the bars in which only a small number of predictors were retained in the model (a Lasso model). Consequently, increasing intermediate values of $\alpha$ are associated with increasing sparsity.

Therefore, we had a closer look on which values for $\alpha$ had been selected during cross-validation irrespective of the underlying true model. Consistent with the observation that in many cases RLR had shown the best performance, the most frequently selected value across all reference models was $\alpha = 0$ for both ENLR approaches at all time points (Figure 2). This minimum value was selected in a third to half of all instances. Interestingly, in nearly all scenarios the second most frequently selected value was $\alpha = 1$. Altogether, an extreme value of either $\alpha = 0$ or 1 was selected in approximately half to two thirds of all cases. Except for a single instance (the SLR reference model and ENLR modeling strategy in the 24 hr data) where the second smallest value $\alpha = 0.1$ was the one most frequently selected, this was also true if the selected $\alpha$ values were considered for each reference model individually (data not shown). Hence, either models including all or almost all possible predictors (lowest possible sparsity) or models including a highly selected subset of predictors were preferred (highest possible sparsity). Highly sparse Lasso models ($\alpha = 1$) were slightly more often selected if only the base predictors were used for ENLR model building (Figure 2, lower panels).

4 | DISCUSSION

Penalized regression methods which are widely used in the field of high-dimensional data ($n < p$) have been suggested to be advantageous also when applied to data of lower dimension (e.g., Göbl et al., 2015; Pavlou et al., 2016; Porzelius, Schumacher, & Binder, 2010). Penalization of the likelihood is a way to correct overfitting, a problem that occurs frequently when prediction models are fitted using classical methods like SLR.

Here, we compared different model building strategies involving penalized regression methods to build a model with a focus on predictive performance of the resulting probabilistic predictions. Due to the clinical problem that motivated this research, prediction should be based on only a few, strongly correlated continuous clinical parameters which were thought to interact in an unknown, complex manner, with a high overall impact on the binary outcome. This led us to consider two main approaches: one using only the basic clinical parameters for prediction, the other including also an additional set of predictors derived from the original measurements based on clinical subject-matter and statistical considerations. In both approaches, different penalized regression methods were compared with other model building techniques, using SLR with base predictors as a benchmark.
In the particular setting of our motivating application, the use of ENLR and to some degree RLR turned out to be preferable across the range of the data generating mechanisms considered in the simulation study. Both methods corrected for overfitting occurring with SLR and displayed the best predictive accuracy and discriminative ability among the methods investigated. In the data sets with smaller numbers of predictors (i.e., the 6 and the 12 hr data sets), simple ENLR models (base predictors) performed comparably to complex RLR models (full predictors). The predictive performance of ENLR$_{f}$ was slightly inferior, although still fairly close to that of ENLR$_{p}$. Only in the 24 hr data set (OPV$_{full} \approx 1$, EPV$_{full} = 0.4$), complex RLR models displayed a better predictive accuracy and model discrimination, at the cost of some more underfitting. Additionally, simple reference models that generated data only from the base predictor set favored the use of also simpler modeling strategies, while more complex reference models favored also modeling strategies relying on the full predictor set.

Among the penalizations techniques considered, those employing a Lasso-type penalty (LLR$_{f}$ and LLR$_{p}$) clearly fell behind the other two (RLR and ENRL) with respect to predictive accuracy (BS) and discrimination (AUC), though displaying good performances with respect to calibration (CS). This notion is further supported by the observation that in the simulation, all Elastic Net approaches had a clear tendency toward complex, non-sparse models, that is, they selected most frequently the minimum value $\alpha = 0$. This implies that for ENLR, the most frequently selected model was actually a Ridge model. The overall tendency to select complex models retaining all or most predictors came as an unexpected finding since we had anticipated that modeling strategies allowing for predictor selection (ENLR and especially LLR) would select only the most important predictors and that such sparse models would have led to more stable predictions in independent data. The fact that the two Lasso methods were not well suited for our purpose is probably due to the high degree of sparsity in Lasso models in combination with the strong correlations of the predictors in our data. As noted by Zou et al. (2005): "If there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected." So, superior predictive performance of RLR and ENLR may have resulted from the fact that they tend to keep the additional bits of predictive information contained in correlated predictors, while Lasso models have more advantages in the presence of many genuine noise predictors (Pavlou et al., 2016; Zou et al., 2005).

Besides the aforementioned penalized regression methods, RFs and GAMs were also included as candidate techniques—the former serving as an example for a technique from the field of genuine machine learning techniques, the latter as a flexible method that allows for functional forms of the predictors different from being solely linear, thus providing an alternative to the inclusion of prespecified transformations of variables into the set of predictors. However, RFs provided an improvement over the benchmark unpenalized SLR only for the data set of 24 hr survivors, while for the data sets of the 6 and 12 hr survivors the performance was worse, probably due to overfitting. This finding confirms observations by Couronné, Probst, and Boulesteix (2018) who noted that RFs tend to outperform benchmark unpenalized SLR more clearly for lower sample sizes and hence lower numbers of observations per variable, as seen in our 24 hr data setting. Although RFs are thought not to overfit (Breiman, 2001), a tendency toward overfitting in smaller data sets has also been noted by others (Van der Ploeg et al., 2014).

As a strength of our approach rarely seen in other comparative studies of modeling strategies, all modeling approaches considered were also used to generate the underlying reference ("true") models, thus allowing a comparison of the different modeling strategies with regard to the complexity of the true model. This design made it possible to identify an issue worth noticing: The best modeling strategy for a given true reference model and performance measure in our study often turned out to be different from the true data generating mechanism, that is, outside the diagonal in Tables 2–4 and Tables S3–S8 in Supporting Information. Although the true model with the true regression coefficients will always yield the best predictive performance, the true regression coefficients are unknown in practice. For the purpose of prediction rather than effect estimation for the respective predictors, we have shown that modeling strategies reflecting a model structure different from the true data generating model can outperform those that are built according to the structure of the true model. Indeed, this finding allowed us to identify strategies that performed better than other competitors across a certain spectrum of true data generating mechanisms.

This study adds to the already existing literature an investigation where modeling strategies for prediction of a binary outcome are compared in a situation with only a few, highly correlated predictors that might exhibit dynamic, non-linear as well as non-additive effects. For this situation which actually resembles a concrete clinical scenario, it is difficult to transfer the concept of EPV criteria in order to give advice for sample size requirements since it is not obvious what the number of variables actually is. In addition, in most studies (see Riley et al., 2019, and Van Smeden et al., 2019, for two recent examples), a larger number of independent or only moderately correlated predictors are usually considered. In the actual clinical application, ENLR$_{f}$ models using the full predictors in the 6 and 12 hr data were finally selected for external validation, due to overlapping time frames for the simulation study and the collection of external validation data. When the results of the simulation became available, ENLR$_{f}$ was found to perform slightly better—although, reassuringly, it was shown that ENLR$_{f}$ actually worked quite well, too. Possibly, further investigations into other ENLR or RLR approaches, for example, with a modified set of predictors, would also have followed, in order to provide further evidence supporting the choice of the most suitable modeling strategy for the specific
In addition, the 24 hr results made it clear that for the 24 hr time point, the best modeling approach depended strongly on the true mechanism that generated the data, known in the simulation but unknown in practice. As a consequence, for predictions after 24 hr, had they been relevant in the actual application, our simulation would have provided the basis to choose a modeling strategy among the reference models considered, selecting the best strategy for those reference models thought to be most realistic in the manner of an educated guess. This uncomfortable finding reflects a challenge due to features of the data at hand, but the simulation approach allowed to explore the available information as best possible.

These observations underline the usefulness of our simulation approach beyond the specific characteristics of its motivating application. We demonstrated in Sections 2.1 and 2.3 how to capture features of a specific clinical prediction problem to select a suitable prediction method under assumptions as realistic as possible for the problem at hand. Careful consideration of the structures in the available set of predictors and integration of clinical subject-matter knowledge can allow to generate simulated data from a range of “known” true models that mimic essential features of the real data and thus support an informed decision. Since the real data generating mechanism is always unknown, an important feature is inclusion of a range of plausible data generating models.

As a possible limitation of the study, the focus here was clearly on penalized regression methods, while other machine learning methods beyond RFs were not included. However, Van der Ploeg et al. (2014) showed that such methods can use their strengths mainly in larger data sets, which was not the case in our setting. Our finding that penalized logistic regression approaches, especially RLR and ENLR, are not inferior to RF machine learning methods is also in line with other studies; see, for example, the recent systematic review of Christodoulou et al. (2019). Due to the strict focus on the exploration of good modeling strategies for one particular clinical setting, no attempt was made here to vary relevant characteristics of the data. In particular, we did not increase the sample size in order to study the effect on the relative performance of the modeling strategies. Furthermore, the specific choice and number of non-linear terms included in the full set of predictors may have affected the results.

In summary, we have demonstrated that in a situation in which a risk prediction model should be based on a small set of highly correlated and interconnected predictors, penalized ENLR using these base predictors or an expanded set of predictor variables, but also RLR can show good performance compared to other competitors in a setting with small sample sizes. Simulation studies that mimic relevant features of a specific data set as exemplified in this research can support the choice of a good modeling strategy.

ACKNOWLEDGMENTS

We would like to thank the reviewers and editors for their thoughtful and valuable comments, which have vastly improved this manuscript.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID

Martin Schumacher https://orcid.org/0000-0002-5037-0099
Erika Graf https://orcid.org/0000-0002-6831-0756

REFERENCES

Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. (2009). Mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software, 32*(6), 1–29.
Bien, J., Taylor, J., & Tibshirani, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics, 41*(3), 1111–1141.
Bien, J., & Tibshirani, R. (2014). *hierNet: A Lasso for hierarchical interactions*. R package version 1.6.
Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3.
Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology, 110*, 12–22.
Couronné, R., Probst, P., & Boulesteix, A. L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics, 19*, 270.
Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1–22.
Gerds, T. A. (2014). *ModelGood: Validation of risk prediction models*. R package version 1.0.9.
Göblk, C. S., Bozkurt, L., Tura, A., Pacini, G., Kautzky-Willer, A., & Mittlböck, M. (2015). Application of penalized regression techniques in modelling insulin sensitivity by correlated metabolic parameters. *PLoS ONE, 10*(11), 1–19.

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

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models* (Vol. 43). Boca Raton, FL: CRC Press.

Hastie, T. J., Tibshirani, R. J., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2011 corr.), 2013 (7e corr.). Springer Series in New York: Springer.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science, 1*(3), 297–310.

Hilden, J., Habbema, J., & Bjerregaard, B. (1978). The measurement of performance in probabilistic diagnosis. III. Methods based on continuous functions of the diagnostic probabilities. *Methods of Information in Medicine, 17*(4), 238–246.

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

Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I. R., Mailey, J. D., & Ziegler, A. (2014). Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory. *Biometrical Journal, 56*(4), 534–563.

Le Cessie, S., & Van Houwelingen, J. C. (1992). Ridge estimators in logistic regression. *Applied Statistics, 41*(1), 191–201.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22.

Pavloiu, 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.

Porzelius, C., Schumacher, M., & Binder, H. (2010). Sparse regression techniques in low-dimensional survival data settings. *Statistics and Computing, 20*(2), 151–163.

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E. Jr, Moons, K. G. M., & 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.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*, 77.

Schoenfeld, D. (2006). Survival methods, including those using competing risks analysis, are not appropriate for intensive care unit outcome studies. *Critical Care, 10*, 103.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., … Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology, 21*(1), 128–138.

Steyerberg, E. (2008). *Clinical prediction models: A practical approach to development, validation, and updating*. Berlin, Germany: Springer Science & Business Media.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267–288.

Van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMCMedical Research Methodology, 14*, 137.

Van der Ploeg, T., Smits, M., Dippel, D. W., Hunink, M., & Steyerberg, E. W. (2011). Prediction of intracranial findings on CT-scans by alternative modelling techniques. *BMCMedical Research Methodology, 11*, 143.

Van Smeden, M., Moons, K. G. M., de Groot, J. A. H., 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*, 2455–2474.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B), 73*(1), 3–36.

Wood, S. (2006). *Generalized additive models: An introduction with R*. London: Chapman & Hall/CRC.

Wengenmayer, T., Duerschmied, D., Graf, E., Chiabudini, M., Benk, C., Mühlenschlegel, S., … Staudacher, D. L. (2018). Development and validation of a prognostic model for survival in patients treated with venoarterial extracorporeal membrane oxygenation: The PREDICT VA-ECMO score. *European Heart Journal: Acute Cardiovascular Care, 8*(4), 350–359.

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.

**SUPPORTING INFORMATION**

Additional supporting information including source code to reproduce the results may be found online in the Supporting Information section at the end of the article.

---

**How to cite this article:** Chiabudini M, Schumacher M, Graf E. Comparison of complex modeling strategies for prediction of a binary outcome based on a few, highly correlated predictors. *Biometrical Journal*. 2020;62:568–582.

https://doi.org/10.1002/bimj.201800243