Statistics in Medicine

Research Article

Received 20 May 2016, Accepted 14 February 2017 Published online 12 March 2017 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7273

# Firth's logistic regression with rare events: accurate effect estimates and predictions?

Rainer Puhr,[a] Georg Heinze,[b] Mariana Nold,[c] Lara Lusa[d] and Angelika Geroldinger[b*†]

**Firth's logistic regression has become a standard approach for the analysis of binary outcomes with small samples. Whereas it reduces the bias in maximum likelihood estimates of coefficients, bias towards one-half is introduced in the predicted probabilities. The stronger the imbalance of the outcome, the more severe is the bias in the predicted probabilities. We propose two simple modifications of Firth's logistic regression resulting in unbiased predicted probabilities. The first corrects the predicted probabilities by a post hoc adjustment of the intercept. The other is based on an alternative formulation of Firth's penalization as an iterative data augmentation procedure. Our suggested modification consists in introducing an indicator variable that distinguishes between original and pseudo-observations in the augmented data. In a comprehensive simulation study, these approaches are compared with other attempts to improve predictions based on Firth's penalization and to other published penalization strategies intended for routine use. For instance, we consider a recently suggested compromise between maximum likelihood and Firth's logistic regression. Simulation results are scrutinized with regard to prediction and effect estimation. We find that both our suggested methods do not only give unbiased predicted probabilities but also improve the accuracy conditional on explanatory variables compared with Firth's penalization. While one method results in effect estimates identical to those of Firth's penalization, the other introduces some bias, but this is compensated by a decrease in the mean squared error. Finally, all methods considered are illustrated and compared for a study on arterial closure devices in minimally invasive cardiac surgery. Copyright © 2017 John Wiley & Sons, Ltd.**

**Keywords:** bias reduction; data augmentation; Jeffreys prior; penalized likelihood; sparse data

## 1. Introduction

In logistic regression, Firth's penalization [1] has gained increasing popularity as a method to reduce the small-sample bias of maximum likelihood (ML) coefficients. Penalizing the likelihood function utilizing the Jeffreys invariant prior does not only remove the first-order term in the asymptotic bias expansion of ML estimates. It also allows computation of reliable, finite estimates of coefficients if separation in a data set is observed, that is, when some linear combination of explanatory variables perfectly discriminates the two outcome states, causing ML estimation to fail [2]. Under separation, ML will yield some predicted probabilities equal to 0 or 1, which is problematic in medical consulting situations because they suggest unrealistic definiteness. Paradoxically, these probabilities are often accompanied with large inconclusive confidence intervals approaching $(0, 1)$. Though, while FL reduces the bias in the estimates of coefficients, it introduces bias in the predicted probabilities as will be illustrated later, whereas ML gives an average predicted probability equal to the observed proportion of events in each sample (also under separation), Firth's penalization biases the average predicted probability towards one-half. This bias in predicted probabilities may be non-negligible if events are very rare or very common.

[a] *The Kirby Institute, University of New South Wales, Sydney, Australia*
[b] *Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria*
[c] *Institute of Medical Statistics, Computer Sciences and Documentation, University Hospital Jena, Jena, Germany*
[d] *Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia*
**Correspondence to: Angelika Geroldinger, Medical University of Vienna, Vienna, Austria.*
† *E-mail: angelika.geroldinger@meduniwien.ac.at*

Whenever the number of observations is critically low, perhaps as indicated by the occurrence of separation, it is tempting to replace the ML estimates by penalized estimates. This might however be detrimental if one is not only interested in effect estimates but also in predicted probabilities. Thus, the present paper has two main objectives. First, we want to clarify how relevant the bias in predicted probabilities based on Firth's penalization is in practice. We investigate the origin of the bias by simple theoretical considerations and empirically quantify it for realistic situations using simulations. Second, we suggest two simple modifications of Firth's penalization to overcome the bias in predicted probabilities and compare them with alternative methods that were proposed for situations of rare events.

The bias in predicted probabilities based on Firth's penalization already becomes apparent in the simple example of logistic regression with a single binary predictor. Assume that we want to investigate the association between a binary outcome $y$ and some binary risk factor $x$ and observe the following counts:

| | | x | |
|---|---|---|---|
| | | 0 | 1 |
| y | 0 | 95 | 4 |
| | 1 | 5 | 1 |

For 2×2 tables, predicted probabilities obtained by ML estimation are equal to the proportion of events in the two groups. Thus, we obtain ML-predicted probabilities of 5% and 20% for $x = 0$ and $x = 1$, respectively, corresponding to an overall average predicted probability of $5.71\% = (5\% \cdot 100 + 20\% \cdot 5)/105$. However, Firth's penalization results in predicted probabilities of 5.45% and 25%, respectively, and an average predicted probability of $6.38\% = (5.45\% \cdot 100 + 25\% \cdot 5)/105$, that is, in an overestimation of 11.6%. This results from the implicit augmentation of cell counts by applying Firth's penalization to this simple case. Here, Firth's penalization is equivalent to ML estimation after adding a constant of 0.5 to each cell (cf. [2]). (Note that a $2 \times 2$ table constitutes the simplest case of a saturated model.) The four cell count modifications can be interpreted as four pseudo-observations, each with weight 0.5, which are added to the original data set. Because the pseudo data have an event rate of 0.5, Firth's penalization leads to overestimation of predicted probabilities in case of rare events.

The present paper proposes two simple modifications of Firth's multivariable logistic regression in order to obtain unbiased average predicted probabilities. First, we consider a simple post hoc adjustment of the intercept. This Firth's logistic regression with intercept-correction (FLIC) does not alter the bias-corrected effect estimates. By excluding the intercept from the penalization, we do not have to trade accuracy of effect estimates for better predictions.

The other approach achieves unbiased predicted probabilities by adjusting for an artificial covariate discriminating between original and pseudo-observations in the iterative weighting procedure mentioned earlier. In this way, Firth's logistic regression with added covariate (FLAC) recalibrates the average predicted probability to the proportion of events in the original data.

In Section 3, we present a comprehensive simulation study comparing the performance of FLIC and FLAC to the performance of other published methods based on logistic regression, which were proposed for both prediction and effect estimation in the situation of rare events. In particular, we would like to clarify whether the proposed modifications improve the accuracy of the predicted probabilities conditional on explanatory variables. Furthermore, in the case of FLAC, we investigate whether the unbiasedness of the average predicted probability is paid for by an inflation of bias in effect estimates. As comparator methods, we consider a compromise between ML and Firth's logistic regression recently proposed by Elgmati *et al.* [3], penalization by log-$F(1, 1)$ [4] or Cauchy priors [5], King and Zeng's 'approximate Bayesian', and 'approximate unbiased' method [6] and ridge regression. These methods are introduced in Section 2. In Section 4, the performance of these methods is illustrated in a study comparing the use of arterial closure devices to conventional surgical access in minimally invasive cardiac surgery.

## 2. Methods

The logistic regression model $P(Y = 1|x_i) = (1 + \exp(-x_i\beta))^{-1}$ with $i = 1, \dots N$ associates a binary outcome $Y$ with observed values $y_i \in \{0, 1\}$ to a vector of covariate values $x_i = (1, x_{i1}, \dots, x_{ip})$ using a $(p + 1)$-dimensional vector of regression parameters $\beta = (\beta_0, \beta_1, \dots, \beta_p)'$. The ML estimate $\hat{\beta}_{\mathrm{ML}}$ is given by the parameter vector maximizing the log-likelihood function $l(\beta)$ and is usually derived by solving the score equations $\partial l/\partial\beta_r = 0$ with $r = 0, \dots, p$. For ML estimates, the proportion of observed events is equal to the average predicted probability. This can be seen easily from the explicit form of the ML score function for the intercept $\partial l/\partial\beta_0=\sum_i y_i-\pi_i$, where $\pi_i=(1+\exp(-x_i\beta))^{-1}$ denotes the predicted probability for the $i$-th observation.

Firth showed that for exponential family models with canonical parametrization penalizing the likelihood function by the Jeffreys invariant prior, that is, by the square root of the determinant of the Fisher information matrix $|I(\beta)|^{1/2}$, removes the first-order term in the asymptotic bias expansion of ML coefficient estimates (cf. [1]). Because the penalty term is asymptotically negligible, penalized and unpenalized coefficient estimates will virtually coincide in large data sets. In logistic regression, the Fisher information matrix $I(\beta)$ is equal to $\mathbf{X}'\mathbf{W}\mathbf{X}$ with $\mathbf{X}$ the design matrix and $\mathbf{W}$ the diagonal matrix diag$(\pi_i(1-\pi_i))$, that is, Jeffreys invariant prior is given by $|\mathbf{X}'\mathbf{W}\mathbf{X}|^{1/2}$. Estimates of coefficients $\hat{\beta}_{\text{FL}}$ for Firth's logistic regression (FL) can be found by solving the corresponding modified score equations

$$\sum_{i=1}^{N}(y_i-\pi_i+h_i(\frac{1}{2}-\pi_i))x_{ir}=0, \quad r=0,\dots p, \qquad (1)$$

where $h_i$ is the $i$-th diagonal element of the hat matrix $\mathbf{W}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{\frac{1}{2}}$ [2]. Equation (1) for $r=0$ reveals that in general, the average predicted probability in FL regression is not equal to the observed proportion of events. Because the determinant of the Fisher information matrix is maximized for $\pi_i=1/2$, it is concluded that Firth's penalization tends to push the predicted probabilities towards one-half compared with ML estimation. Thus, in the situation of rare events, Firth's penalization is prone to overestimate predictions. We can gain more insight into the behavior of this bias by interpreting the modified score equations (1) as score equations for ML estimates for an augmented data set. This data set can be created by complementing each original observation $i$ with two pseudo-observations weighted by $h_i/2$ with unchanged covariate values and with response values set to $y=0$ and $y=1$, respectively. Thus, FL estimates could be obtained by iteratively applying ML estimation to the augmented data. Given that the trace of the hat matrix is always equal to $p+1$ with $p$ the number of explanatory variables, we see that the augmented data contain $(p+1)/2$ more events than the original one. Consequently, the predicted probability by FL, averaged over the observations in the augmented data, is equal to $(k+(p+1)/2)/(N+p+1)$ with $k$ the number of events. This gives a very rough approximation of the predicted probability averaged over the original observations, being the closer to the true value, the more homogeneous the diagonal elements of the hat matrix are. Unsurprisingly, the relative bias in the average predicted probability is larger for smaller numbers of events $k$ and for larger numbers of parameters $p$ – exactly in the same situations where the application of FL is indicated to reduce small-sample bias.

In the following, we suggest two simple modifications of FL that provide average predicted probabilities equal to the observed proportions of events, while preserving the ability to deal with separation.

First, we consider altering only the intercept of Firth's estimates such that the predicted probabilities become unbiased while keeping all other coefficients constant. In practice, estimates for this FLIC can be derived as follows:

(1) Determine coefficient estimates $\hat{\beta}_{\text{FL}}$ by Firth's penalization.
(2) Calculate the linear predictors $\hat{\eta}_i=\hat{\beta}_{\text{FL},1}x_{i1}+\cdots+\hat{\beta}_{\text{FL},p}x_{ip}$, omitting the intercept.
(3) Determine the ML estimate $\hat{\gamma}_0$ of the intercept in the logistic model $P(y_i=1)=(1+\exp(-\gamma_0-\hat{\eta}_i))^{-1}$, containing only a single predictor $\hat{\eta}_i$ with regression coefficient equal to one. This can be achieved by including an offset in a standard routine or by direct calculation.
(4) The FLIC estimate $\hat{\beta}_{\text{FLIC}}$ is then given by Firth's estimate $\hat{\beta}_{\text{FL}}$ with the intercept replaced by $\hat{\gamma}_0$, so $\hat{\beta}_{\text{FLIC}}=(\hat{\gamma}_0,\hat{\beta}_{\text{FL},1},\dots,\hat{\beta}_{\text{FL},p})$.

See the first section of the Supporting Information for a description of an R implementation of FLIC. By definition, for FLIC, the average predicted probability is equal to the proportion of observed events. For the example of the $2\times2$ table in the Introduction, that is, 100 subjects with five events for $x=0$ and five subjects with one event for $x=1$, FLIC yields predicted probabilities of 4.86% and 22.82%, respectively. Similarly as for Firth's logistic regression, we suggest the use of profile (penalized) likelihood confidence intervals (CIs) for the coefficients estimated by FLIC except for the intercept [2]. We suggest to approximate the standard error for the intercept as the square root of the upper left entry of the matrix $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$, where similarly as given earlier, $\mathbf{W}$ denotes the $(N\times N)$-diagonal matrix diag$(\pi_{\text{FLIC},i}(1-\pi_{\text{FLIC},i})$. Wald-type CIs based on this approximation were evaluated in the simulation study (Section 3).

Second, we introduce FLAC. The basic idea is to discriminate between original and pseudo-observations in the alternative formulation of Firth's estimation as an iterative data augmentation procedure, which was described previously. For instance, in the case of $2\times 2$ tables, where FL amounts to ML estimation of an augmented table with each cell count increased by 0.5, FLAC estimates are obtained by a stratified analysis of the original $2 \times 2$ table and the pseudo data, given by a $2 \times 2$ table with each cell count equal to 0.5. In the general case, FLAC estimates $\beta_{\text{FLAC}}$ can be obtained as follows:

(1) Apply Firth's logistic regression and calculate the diagonal elements $h_i$ of the hat matrix.
(2) Construct an augmented data set by stacking

  (i) the original observations weighted by 1,
  (ii) the original observations weighted by $h_i/2$ and,
  (iii) the original observations weighted by $h_i/2$ but with reversed values of the binary outcome variable ($y_i$ replaced by $1 - y_i$).

(3) Define an indicator variable $g$ on this augmented data set, where for (i) $g = 0$ and for (ii) and (iii) $g = 1$.
(4) The FLAC estimates $\hat{\beta}_{\text{FLAC}}$ are then obtained by ML estimation on the augmented data set adding $g$ as covariate.

See the first section of the Supporting Information for a description of an `R` implementation of FLAC. If one does not include the indicator variable $g$ as additional covariate in the last step, this algorithm will give identical results to Firth's original penalization. For the example of the $2\times 2$ table in the Introduction, that is, 100 subjects with five events for $x = 0$ and five subjects with one event for $x = 1$, FLAC yields predicted probabilities of 5.16% and 16.83%, respectively. For data augmentation in $2\times 2$ tables, the idea of discriminating between original and pseudo-observations was already explored in [7]. For $2\times 2$ tables, odds ratios estimated with FLAC are closer to one, that is, less extreme than odds ratios estimated with ML, as shown in the Appendix. Moreover, in the Appendix, it is shown that the average predicted probability by FLAC is exactly equal to the proportion of observed events for any not necessarily univariate model. Confidence intervals for coefficients estimated by FLAC can be deduced from the ML estimation on the augmented data in the last step. Again, we prefer profile likelihood over Wald CIs.

Because of the asymptotically vanishing influence of the pseudo-observations, coefficient estimates by FLIC and FLAC will approach FL (and ML) estimates in large samples. Furthermore, FLIC and FLAC estimates are invariant in the same sense as ML estimates: Transforming the covariates only changes the estimates in the expected algebraic way, for example, replacing age in years with age in decades, the corresponding coefficient estimate will multiply tenfold. Both FLIC and FLAC give finite coefficient estimates in the case of separation.

One of the main aims of the simulation study in Section 3 is to investigate whether these adjustments of the average predicted probability are also reflected in improved-predicted probabilities at the subject level. Besides ML and classical Firth's penalization, we compared the performance of FLIC and FLAC to the performance of the following methods:

- A 'weakened' Firth's penalization (WF) is proposed by Elgmati *et al.*, where the likelihood is penalized by $|I(\beta)|^{\tau}$ with weight $\tau$ between 0 (corresponding to ML estimation) and one-half (corresponding to Firth's penalization) (cf. [3]). This gives a compromise between accuracy of predictions, where Firth's method is outperformed by ML estimation and the handling of stability and separation issues, which is a strength of Firth's penalization. We chose the weight $\tau$ equal to 0.1 as recommended by Elgmati *et al.* Similarly as Firth's original penalization, WF estimates could be obtained by iteratively applying ML estimation to an augmented data set, in this case, with an event rate of $(k + (p + 1)/10)/(N + p + 1)$.
- Penalizing by log-$F(1, 1)$ priors (LF) amounts to multiplying the likelihood by $\prod e^{\beta_j/2}/(1 + e^{\beta_j})$, where the product ranges over $j \in \{1, \dots, p\}$. This type of penalization is regarded a better choice of a 'default prior' by Greenland and Mansournia [4] compared with methods such as Firth's penalization or Cauchy priors. We followed their suggestion to omit the intercept from the penalization, resulting in an average predicted probability equal to the proportion of observed events. Quantitative explanatory variables are advised to be scaled in units that are contextually meaningful (cf. [4]). Unlike Jeffreys prior, the LF does not incorporate the correlation between explanatory variables.
- Penalization by Cauchy priors (CP) in logistic regression is suggested as 'default choice for routine applied use' by Gelman *et al.* [5]. Unlike Greenland and Mansournia, they give an explicit

recommendation for data preprocessing, which is also implemented in the corresponding `R` function `bayesglm` in the package `arm`. All explanatory variables are shifted to have a mean of 0. Binary variables are coded to have a range of 1, and all others are scaled to have a standard deviation of 0.5. Then, all explanatory variables are penalized by CP with center 0 and scale 2.5. The intercept is assigned a weaker CP, with center 0 and scale 10, which implies that the average predicted probability is in general not equal, but very close to the observed proportion of events.

- King and Zeng's approximate Bayesian method (KAB) (cf. [6]) takes as starting point small-sample bias-corrected logistic regression coefficients, such as coefficients estimated by FL, $\hat{\beta}_{\mathrm{FL}}$. Predicted probabilities by KAB, $\hat{\pi}_{\mathrm{KAB},i}$, are then obtained by averaging the predicted probabilities by FL over the posterior distribution of the coefficients,

$$\hat{\pi}_{\mathrm{KAB},i} = \int (1 + \exp(-x_i\beta^*))^{-1} f(\beta^*) d\beta^*, \tag{2}$$

  where $f(\beta^*)$ denotes the posterior density of the parameter $\hat{\beta}_{\mathrm{FL}}$, approximated by $\mathcal{N}(\hat{\beta}_{\mathrm{FL}}, -(\frac{\partial^2 l}{\partial \beta^2}(\hat{\beta}_{\mathrm{FL}}))^{-1})$. Instead of deriving $\hat{\pi}_{\mathrm{KAB},i}$ by numerical integration, we make use of the approximation $\hat{\pi}_{\mathrm{KAB},i} = \hat{\pi}_{\mathrm{FL},i} + C_i$, where the correction factor $C_i$ is given by $(0.5 - \hat{\pi}_{\mathrm{FL},i})h_i$ (cf. [6]). In general, equality of average predicted probability and observed event rate does not hold for this method; the bias of the average predicted probability by KAB is exactly twice as large as by Firth's penalization. This follows from the fact that the sum over $C_i$ is equal to $\sum_i \hat{\pi}_{\mathrm{FL},i} - y_i$ according to the modified score equation for the intercept in FL estimation. King and Zeng are aware that their estimator introduces bias into the predicted probabilities but claim that this is compensated by a reduction of the root mean squared error (RMSE).
- King and Zeng's approximate unbiased estimator (KAU) (cf. [6]) can be understood as a counterpart to the approximate Bayesian method; using the notation introduced earlier, predicted probabilities by the KAU are now defined as $\hat{\pi}_{\mathrm{KAU},i} = \hat{\pi}_{\mathrm{FL},i} - C_i$. Similarly, as we have seen that the bias in the average predicted probability by KAB is twice as large as by FL, one can easily verify that the average predicted probability by KAU is equal to the observed proportion of events, that is, unbiased. Nevertheless, King and Zeng consider KAB preferable to KAU 'in the vast majority of applications'. One should be aware that the definition of KAU does not ensure that predicted probabilities fall inside the range of 0 to 1.
- In ridge regression (RR), the log-likelihood is penalized by the square of the Euclidean norm of the regression parameters, $\beta_1^2 + \ldots + \beta_p^2$, multiplied by a tuning parameter $\lambda$ [8]. Because the intercept is omitted from the penalty, the average predicted probability is equal to the proportion of observed events. Following Verweij and Van Houwelingen [9], in our simulation study, the tuning parameter $\lambda$ was chosen by minimizing the penalized version of the Akaike's Information Criterion $\mathrm{AIC} = -2l(\hat{\beta}) + 2df_e$, with effective degrees of freedom $df_e = \mathrm{trace}(\frac{\partial^2 l}{\partial \beta^2}(\hat{\beta})(\frac{\partial^2 l^*}{\partial \beta^2}(\hat{\beta}))^{-1})$ and $l^*$ denoting the penalized log-likelihood. Wald-type CIs were deduced from the penalized variance–covariance matrix with fixed tuning parameter. RR was always performed on scaled explanatory variables with standard deviation equal to one, but results are reported on the original scale.

The following types of CIs were chosen for the different methods: Wald-type CIs were used for ML, CP, and RR and profile likelihood CIs for WF, FL, FLAC and LF. For FLIC, CIs were constructed as explained earlier.

## 3. Simulation study

The empirical performance of the methods introduced in Section 2 was evaluated in a comprehensive simulation study. Integral parts were the comparison of bias and RMSE of predicted probabilities and estimates of coefficients. Furthermore, we assessed discrimination, quantified by the c-statistic, and the calibration slope of the models [10]. CIs were evaluated with regard to power, length, and coverage. Results on predictions are presented for all methods introduced earlier, ML, WF, FL, FLIC, FLAC, LF, CP, KAU, KAB, and RR, whereas results on coefficient estimates only concern ML, WF, FL, FLIC, FLAC, LF, CP, and RR. Whenever we are solely interested in coefficient estimates except for the intercept, results for FL and FLIC agree and are presented jointly.

### 3.1. Data generation

Binary outcomes $y_i$ were generated from the logistic model $P(Y|x_{i1}, \dots, x_{i10}) = (1 + \exp(-\beta_0 - \beta_1 x_{i1} - \dots \beta_{10} x_{i10}))^{-1}$, $i = 1 \dots N$, with 10 explanatory variables $x_{i1}, \dots, x_{i10}$. The joint distribution of the explanatory variables follows Binder, Sauerbrei and Royston [11] and includes continuous as well as categorical variables. Because one focus of our paper is on predictions, we reduced the number of explanatory variables from 17 [11] to 10 and considered only linear effects.

First, we generated 10 standard normal random variables $z_{ij} \sim \mathcal{N}(0,1)$, $j = 1, \dots, 10$, $i = 1, \dots N$, with correlation structure as listed in Table A.1. By applying the transformations described in Table A.1 to the variables $z_{ij}$, four continuous, four binary, and two ordinal variables $x_{ij}$ were derived. The continuous variables $x_{i1}, x_{i4}, x_{i5}$, and $x_{i8}$ were truncated at the third quartile plus five times the interquartile distance in each simulated data set.

Effect strengths $\theta_2$, $\theta_6$, $\theta_9$, and $\theta_{10}$ of the binary variables were set to 0.69, and $\theta_3$ and $\theta_7$, corresponding to ordinal variables, to 0.345. This relates to an odds ratio of 2 for binary variables and to an odds ratio of $\sqrt{2}$ for consecutive categories of ordinal variables. Effect strengths $\theta_1$, $\theta_4$, $\theta_5$, and $\theta_8$ of the continuous variables were chosen such that the difference between the first and the fifth sextile of the empirical distribution function corresponds to an odds ratio of 2. To obtain scenarios with no effects, small effects, and large effects, we considered a global effect size parameter $a$ with values 0, 0.5, and 1, respectively, and defined the true model coefficients as $\beta_j = a\theta_j$, $j = 1, \dots, 10$. The intercept $\beta_0$ was chosen such that a certain desired proportion of events was obtained. We considered all combinations of sample sizes $N \in \{500, 1400, 3000\}$ and population event rates $\pi \in \{1\%, 2\%, 5\%, 10\%\}$, which were such that the expected number of events was greater than 20. Finally, we also considered cases where effects have different directions by multiplying $\beta_j, j = 6, \dots, 10$ by $-1$ ('mixed effect directions' scenarios). For each of these 45 scenarios, 1000 data sets were simulated and analyzed by the methods described in Section 2. The sample sizes were chosen such that in approximately 50%, 80%, and 95% of 1000 data sets with event rate of 10% and positive, large effects, coefficients by Firth's penalization were significantly different from zero.

Details on the software used for data generation and on the implementation of the different methods can be found in the Supporting Information.

### 3.2. Results

To improve readability, figures and tables of results in this section and in the Supporting Information are restricted to some selected scenarios with mixed effect directions, that is, $\text{sgn}(\beta_j) = -1, j = 6, \dots, 10$. Results for scenarios with only positive coefficients $\beta$ were similar and are available from the authors upon request. In the text, we refer to all scenarios if not stated otherwise. Separation was encountered at most in 4 out of 1000 simulated data sets per scenario. Although some methods can handle separation in data sets, we excluded these cases from analyses to retain comparability.

We begin by analyzing the discrepancy between the average predicted probability and the observed event rate for the methods WF, FL, CP, and KAB (Table I). All other methods (ML, FLIC, FLAC, LF, KAU, and RR) give average predicted probabilities equal to the observed event rate by construction. Among all considered scenarios, the scenario with a sample size of 500, an expected event rate of 0.05 and with all explanatory variables being unrelated to the outcome (effect size of zero) was associated with the largest relative difference between average predicted probability and the observed event rate for WF (4%), FL (19.4%), and for KAB (38.8%). The difference was about five times larger for FL than for WF and exactly half the size of the difference for KAB. For CP, the difference between average predicted probability and observed event rate can be considered negligible (relative difference smaller than 0.3% for all scenarios). These numbers show that in cases of rare events, the systematic error in the average predicted probability by FL cannot be ignored.

Table II shows the bias and RMSE of the individual predicted probabilities. ML, FLIC, FLAC, LF, KAU, and RR give unbiased predicted probabilities – the fact that the respective numbers are not exactly equal to zero is due to sampling variability in the simulation. RR was associated with considerably lower RMSE than other methods in all but one scenarios, with an RMSE up to 52.4% lower than the second best performing method, which was either FLAC or, in some situations with large effect size, FLIC. Both FLAC and FLIC did not only correct the bias in predicted probabilities but also reduced the variance and in particular the RMSE (by up to 23.4% and 16.2%, respectively) in comparison to FL. CP resulted in smaller RMSE than LF, which still performed better than WF in all scenarios. King and Zeng's approximate Bayesian method KAB performed worst with respect to bias and RMSE in almost all settings.

**Table I.** Mean relative difference (×100) between the average predicted probability $\sum_i \hat{\pi}_i/N$ and the observed event rate $\sum_i y_i/N$ in 1000 data sets simulated from the scenarios with small effect size ($a = 0.5$).

| | | Event rate ($\pi$) | | | |
|---|---|---|---|---|---|
| Sample size ($N$) | Method | 0.01 | 0.02 | 0.05 | 0.1 |
| 500 | WF | | | 3.7 | 1.6 |
| | FL | | | 18.2 | 7.8 |
| | CP | | | 0.2 | 0.1 |
| | KAB | | | 36.4 | 15.5 |
| 1400 | WF | | 3.7 | 1.3 | 0.6 |
| | FL | | 18.5 | 6.6 | 2.8 |
| | CP | | 0.2 | 0.1 | 0.0 |
| | KAB | | 37.1 | 13.2 | 5.6 |
| 3000 | WF | 3.6 | 1.7 | 0.6 | 0.3 |
| | FL | 17.9 | 8.6 | 3.1 | 1.3 |
| | CP | 0.3 | 0.1 | 0.0 | 0.0 |
| | KAB | 35.9 | 17.1 | 6.2 | 2.6 |

The relative difference was computed as $\frac{1}{1000}\sum_{s=1}^{1000}\left(\sum_{i=1}^{N}\hat{\pi}_{s,i}/N - \sum_{i=1}^{N} y_{s,i}/N\right)/(\sum_{i=1}^{N} y_{s,i}/N)$, where $\hat{\pi}_{s,i}$ denotes the predicted probability and $y_{s,i}$ the binary outcome for the $i$-th observation in the $s$-th simulated data set.
The methods not considered in the table (ML, FLIC, FLAC, LF, KAU, and RR) have zero difference by construction.
WF, weakened Firth's logistic regression; FL, Firth's logistic regression; CP, penalization by Cauchy priors; KAB, King and Zeng's approximate Bayesian method.

For 17 simulation scenarios with smaller number of events, King and Zeng's approximate unbiased method KAU yielded predicted probabilities outside of $[0, 1]$, with a minimum value of $-0.02$ and a maximum value of 1.0008. With increasing sample size, event rate and effect size, differences between methods diminished.

All methods except for RR yielded mean calibration slopes smaller than 1 throughout all scenarios, indicating underestimation of small event probabilities and overestimation of larger ones (cf. Table II). The method achieving the calibration slope closest to the optimal value of 1 was either FLAC or RR depending on the scenario. Both FLIC and FLAC outperformed ordinary Firth's penalization with respect to calibration. With increasing sample size and expected event rate, differences between methods and the distance to the optimal value of 1 decreased.

Figure 1 investigates the bias and RMSE of predicted probabilities in relation to the size of the true linear predictor exemplarily for one simulation scenario. In line with the results on the calibration slope in Table II, we find that RR strongly overestimates small event probabilities and underestimates large ones. A similar, but less pronounced, pattern is shown by FLAC. Predicted probabilities by KAU were not only close to unbiased over the whole range of predictions but were also associated with considerable variance, in particular for larger probabilities. For better discriminability of methods, Figure 1 shows the bias and RMSE scaled by the standard error of proportions corresponding to the respective true linear predictors. Figure S2 is the unscaled version of Figure 1.

The discriminative power of the models (in terms of c-indices) was evaluated with regard to a new, independently generated outcome drawn from the logistic model with the same covariate values as in the training data. Highest discrimination was achieved by either RR or KAB depending on the scenario (cf. Table S2). Though, RR also performed worst in more than one fifth of scenarios. The variation across methods can be considered relatively low with a range of c-indices smaller than 0.013 for all scenarios. By construction, FL and FLIC give the same c-indices because adjusting the intercept does not change the order of predicted probabilities. The 'optimal value' in Table S2 was obtained by calculating the c-index based on event probabilities from the true model.

Because the relation between predicted probabilities and linear predictors is nonlinear of nonvanishing curvature for linear predictors smaller than 0, a reduced bias of predictions often comes at the cost of

**Table II.** Bias and RMSE (×10000) of predicted probabilities $\hat{\pi}_i$, mean, and standard deviation (×100) of calibration slopes, for selected simulation scenarios. (See Table S1 for further scenarios and Figure S1 for a graphical illustration.)

| | | | Predicted probabilities | | | | | | Calibration slope | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Bias (×10000) | | | RMSE (×10000) | | | Mean (×100) | | SD (×100) | |
| | | | Effect size ($a$) | | | Effect size ($a$) | | | Effect size ($a$) | | Effect size ($a$) | |
| Sample size ($N$) | Event rate ($\pi$) | Method | 0 | 0.5 | 1 | 0 | 0.5 | 1 | 0.5 | 1 | 0.5 | 1 |
| 500 | 0.05 | ML | −1 | 0 | −1 | 351 | 403 | 469 | 43 | 80 | 16 | 18 |
| | | WF | 18 | 18 | 14 | 359 | 408 | 469 | 43 | 80 | 16 | 17 |
| | | FL | 91 | 87 | 74 | 392 | 430 | 472 | 41 | 78 | 14 | 16 |
| | | FLIC | −1 | 0 | −1 | 332 | 375 | 437 | 48 | 87 | 17 | 20 |
| | | FLAC | −1 | 0 | −1 | 312 | 360 | 435 | 50 | 91 | 19 | 22 |
| | | LF | −1 | 0 | −1 | 340 | 391 | 453 | 45 | 83 | 17 | 19 |
| | | CP | 0 | 1 | 0 | 326 | 377 | 440 | 47 | 86 | 18 | 20 |
| | | KAU | −1 | 0 | −1 | 351 | 407 | 473 | 43 | 80 | 16 | 18 |
| | | KAB | 184 | 174 | 150 | 457 | 477 | 495 | 39 | 76 | 12 | 14 |
| | | RR | −1 | 0 | −1 | 153 | 282 | 424 | 128 | 117 | 85 | 66 |
| | 0.10 | ML | −1 | −4 | −2 | 463 | 503 | 533 | 61 | 87 | 16 | 13 |
| | | WF | 16 | 11 | 11 | 466 | 504 | 531 | 60 | 87 | 15 | 13 |
| | | FL | 82 | 71 | 63 | 481 | 509 | 529 | 60 | 88 | 15 | 13 |
| | | FLIC | −1 | −4 | −2 | 447 | 481 | 512 | 64 | 91 | 16 | 14 |
| | | FLAC | −1 | −4 | −2 | 434 | 476 | 512 | 65 | 93 | 17 | 14 |
| | | LF | -1 | −4 | −2 | 456 | 495 | 523 | 62 | 88 | 16 | 13 |
| | | CP | 0 | −3 | −1 | 446 | 486 | 514 | 63 | 90 | 16 | 14 |
| | | KAU | -1 | −4 | −2 | 463 | 506 | 535 | 60 | 87 | 16 | 13 |
| | | KAB | 164 | 147 | 127 | 516 | 526 | 536 | 59 | 88 | 14 | 12 |
| | | RR | -1 | -4 | -2 | 235 | 406 | 506 | 116 | 102 | 53 | 23 |
| 3000 | 0.01 | ML | 0 | 0 | 0 | 66 | 84 | 137 | 51 | 85 | 17 | 20 |
| | | WF | 4 | 4 | 4 | 68 | 86 | 138 | 49 | 83 | 17 | 19 |
| | | FL | 18 | 18 | 16 | 78 | 97 | 144 | 45 | 78 | 14 | 16 |
| | | FLIC | 0 | 0 | 0 | 65 | 82 | 130 | 52 | 88 | 17 | 21 |
| | | FLAC | 0 | 0 | 0 | 60 | 75 | 127 | 58 | 97 | 20 | 25 |
| | | LF | 0 | 0 | 0 | 65 | 82 | 134 | 52 | 86 | 18 | 21 |
| | | CP | 0 | 0 | 1 | 62 | 79 | 130 | 54 | 89 | 19 | 22 |
| | | KAU | 0 | 0 | 0 | 66 | 85 | 139 | 50 | 84 | 17 | 20 |
| | | KAB | 36 | 35 | 32 | 94 | 114 | 156 | 40 | 73 | 12 | 14 |
| | | RR | 0 | 0 | 0 | 29 | 60 | 125 | 135 | 111 | 81 | 40 |

The bias of predicted probabilities was calculated as $\frac{1}{1000\cdot N}\sum_{s=1}^{1000}\sum_{i=1}^{N}\hat{\pi}_{s,i}-\pi_{s,i}$, where $\hat{\pi}_{s,i}$ and $\pi_{s,i}$ denote the estimated and true predicted probability for the $i$-th observation in the $s$-th simulated data set, respectively. The root mean squared error (RMSE) was computed as $\left(\frac{1}{1000\cdot N}\sum_{s=1}^{1000}\sum_{i=1}^{N}(\hat{\pi}_{s,i}-\pi_{s,i})^2\right)^{1/2}$.

Effect sizes $a \in \{0, 0.5, 1\}$ refer to scenarios with no, small, and large effects, respectively, are global multipliers of the log odds ratios as described in Section 3.1.

ML, maximum likelihood; WF, weakened Firth's logistic regression; FL, Firth's logistic regression; FLIC, Firth's logistic regression with intercept-correction; FLAC, Firth's logistic regression with added covariate; LF, penalization by log-$F(1,1)$ priors; CP, penalization by Cauchy priors; KAU, King and Zeng's approximate unbiased method; KAB, King and Zeng's approximate Bayesian method; RR, ridge regression.

an increased bias of linear predictors and vice versa. This effect is less pronounced if the variability of the estimator is small. For instance, RR, having the smallest RMSE of linear predictors among the investigated methods, performed well with regard to both, linear predictors and predicted probabilities (cf. Table S3). Bias and RMSE of linear predictors were of largest absolute size for ML across all 45 scenarios. FL was least biased in almost all scenarios but with considerably larger RMSE than RR. Again, with increasing number of expected events and effect size, differences between methods decreased.

Table III shows the absolute bias and RMSE of the standardized coefficients averaged over all explanatory variables, omitting the intercept. In 28 out of 36 simulation scenarios with nonzero covariate effects, FL outperformed the other methods with respect to absolute bias. Even in unfavorable scenarios, its average standardized bias did not exceed 1%. Concerning the RMSE, FL ranked at least fourth place,

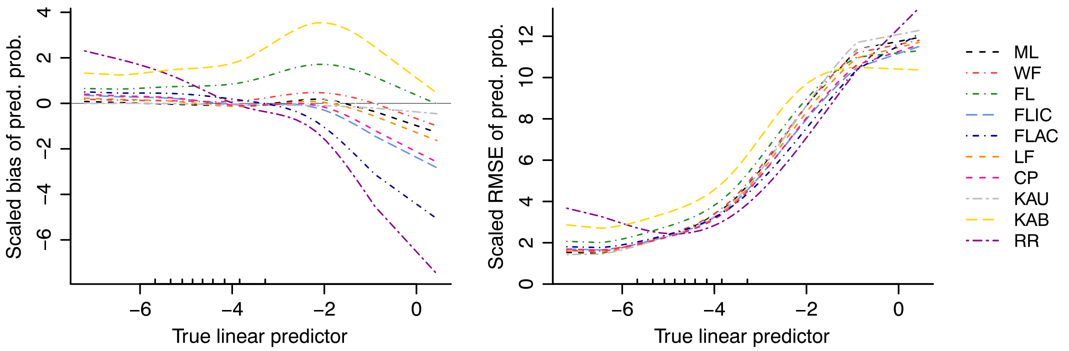

**Figure 1.** Bias and root mean squared error (RMSE) of predicted probabilities (scaled by the standard error of proportions $\sqrt{p(1-p)/N}$ with $p$ the probability corresponding to the true linear predictor) by true linear predictor, exemplarily for the scenario $N = 1400$, $\pi = 0.02$, large effect size ($a = 1$), and mixed effect directions. For the calculation of bias and RMSE, the predicted probabilities were splitted into 30 groups using adequate quantiles of the true linear predictor. Cubic smoothing splines were then fitted to the derived bias and RMSE values in the 30 groups. Upward directed ticks on the $x$-axis mark the deciles of the true linear predictor. (See Figure S2 for an unscaled version of this plot). ML, maximum likelihood; WF, weakened Firth's logistic regression; FL, Firth's logistic regression; FLIC, Firth's logistic regression with intercept-correction; FLAC, Firth's logistic regression with added covariate; LF, penalization by log-$F(1, 1)$ priors; CP, penalization by Cauchy priors; KAU, King and Zeng's approximate unbiased method; KAB, King and Zeng's approximate Bayesian method; RR, ridge regression.

almost always clearly outmatched by RR and throughout all scenarios slightly worse than CP and, interestingly, FLAC, but never worse than LF. ML and WF were associated with the largest and second largest RMSE throughout all simulation scenarios. Unsurprisingly, RR gave the smallest absolute bias in simulation scenarios with zero covariate effects but closely followed by FL.

Finally, the approximate Wald-type standard error for the intercept in FLIC suggested in Section 2 was evaluated by comparing the corresponding CIs to intervals based on jackknife and bootstrap (200 repetitions) estimates of the standard error. Because of the computational burden, this comparison was restricted to the 25 simulation scenarios with sample size $N = 500$ and $N = 1400$. In 17 out of 20 simulation scenarios with nonzero covariate effects, the approximate approach yielded the CIs that most often excluded zero, but differences between methods were marginal (cf. Table S5). The approximate Wald-type CI was also the shortest. Especially in extreme situations, the coverage exceeded the nominal significance level for all three methods.

For all methods, results on 95% CIs averaged over all coefficients omitting the intercept can be found in Table IV. Coverage was reasonable for all methods except for RR, ranging between 93.9% and 96% across methods and scenarios. RR CIs were overly conservative in simulation scenarios without any covariate effects with coverage levels as high as 99.7% and were too optimistic for scenarios with moderate or large effects. They were clearly shorter with less power to exclude 0 compared with the other methods. In 29 out of 36 simulation scenarios, CP and FLAC were among the four methods with smallest power, often combined with a slight conservatism. Though, in general, there were little differences in the behavior of the CIs among all methods except for RR.

The likelihood of separation was very low among these scenarios (at most 4 out of 1000 simulated data sets). In order to investigate whether a higher proportion of data sets with separation would yield different results, we extended our simulations by setting the effect strength $\theta_2$ to 3.47 instead of 0.69, corresponding to an odds ratio of 32. This resulted in proportions of data sets with separation of up to 50% (Figure S4). With effect size $a = 0$, these separation-prone scenarios and the scenarios described earlier agree. Whenever separation occurs, at least one of the ML coefficients does not exist, that is, the estimation algorithm does not converge. In this case, we either used the results from the last iteration (referred to as ML) or plugged in Firth estimates (referred to as MLFL). RR was particularly vulnerable to the occurrence of separation and performed much worse in the separation-prone scenarios than before (Tables S7 and S8). This is because the *AIC* criterion was then often optimized at $\lambda = 0$, leading to and consequently inheriting the problems of ML estimation. In those scenarios, FLAC was often the best method, effectively removing bias from predicted probabilities and providing the smallest RMSEs, while still supplying very accurate regression coefficients (Tables S7 and S8).

**Table III.** Absolute bias and RMSE (×1000) of standardized coefficients averaged over all explanatory variables (omitting the intercept), for selected simulation scenarios. (See Table S4 for further scenarios and Figure S3 for a graphical illustration.)

| | | | Bias (×1000) Effect size ($a$) | | | RMSE (×1000) Effect size ($a$) | | |
|---|---|---|---|---|---|---|---|---|
| Sample size ($N$) | Event rate ($\pi$) | Method | 0 | 0.5 | 1 | 0 | 0.5 | 1 |
| 500 | 0.05 | ML | 23 | 17 | 29 | 277 | 266 | 288 |
| | | WF | 19 | 14 | 21 | 272 | 261 | 281 |
| | | FL/FLIC | 7 | 5 | 9 | 253 | 244 | 259 |
| | | FLAC | 17 | 16 | 16 | 239 | 235 | 252 |
| | | LF | 22 | 10 | 12 | 265 | 252 | 266 |
| | | CP | 18 | 14 | 24 | 245 | 238 | 251 |
| | | RR | 3 | 109 | 124 | 78 | 166 | 244 |
| | 0.10 | ML | 11 | 7 | 21 | 191 | 188 | 203 |
| | | WF | 10 | 6 | 17 | 189 | 186 | 201 |
| | | FL/FLIC | 4 | 3 | 3 | 181 | 178 | 191 |
| | | FLAC | 10 | 9 | 7 | 177 | 175 | 189 |
| | | LF | 11 | 5 | 9 | 187 | 183 | 196 |
| | | CP | 10 | 8 | 10 | 179 | 177 | 189 |
| | | RR | 2 | 92 | 75 | 68 | 143 | 190 |
| 3000 | 0.01 | ML | 22 | 15 | 18 | 234 | 223 | 231 |
| | | WF | 19 | 12 | 14 | 232 | 221 | 228 |
| | | FL/FLIC | 8 | 6 | 8 | 223 | 213 | 218 |
| | | FLAC | 18 | 20 | 17 | 208 | 203 | 212 |
| | | LF | 21 | 12 | 8 | 227 | 214 | 219 |
| | | CP | 19 | 16 | 18 | 214 | 206 | 212 |
| | | RR | 3 | 104 | 102 | 71 | 154 | 210 |

The absolute bias was calculated as $\frac{1}{10\cdot 1000}\sum_{j=1}^{10}|\sum_{s=1}^{1000}\hat{\beta}_{s,j}-\beta_j|$, where $\hat{\beta}_{s,j}$ and $\beta_j$ denote the standardized estimated and true coefficient of the $j$-th explanatory variable for the $s$-th simulated data set, respectively. The root mean squared error (RMSE) was computed as $\frac{1}{10}\sum_{j=1}^{10}\left(\frac{1}{1000}\sum_{s=1}^{1000}(\hat{\beta}_{s,j}-\beta_j)^2\right)^{\frac{1}{2}}$.

Effect sizes $a \in \{0, 0.5, 1\}$ refer to scenarios with no, small, and large effects, respectively, are global multipliers of the log odds ratios as described in Section 3.1.

ML, maximum likelihood; WF, weakened Firth's logistic regression; FL, Firth's logistic regression; FLIC, Firth's logistic regression with intercept-correction; FLAC, Firth's logistic regression with added covariate; LF, penalization by log-$F(1,1)$ priors; CP, penalization by Cauchy priors; RR, ridge regression.

## 4. Example: arterial closure devices in minimally invasive cardiac surgery

In a retrospective study at the Department of Cardiothoracic Surgery of the University Hospital of the Friedrich-Schiller University Jena, the use of arterial closure devices (ACDs) in minimally invasive cardiac surgery was compared with conventional surgical access with regard to the occurrence of cannulation-site complications. Of the 440 patients eligible for analysis, 16 (3.6%) encountered complications. About one fifth of surgeries (90 cases) were performed with conventional surgical access to the groin vessels. The complication rate was 8.9% (8 cases) for the conventional surgical access and 2.3% (8 cases) for the ACDs group. For the purpose of illustration, the analysis in the present paper was restricted to four adjustment variables selected by medical criteria, the *logistic EuroSCORE*, which estimates the risk of mortality in percent, the presence of *previous cardiac operations* (yes/no), the body mass index (BMI), for which five missing values were replaced by the mean BMI, and the presence of *diabetes* (yes/no). The aim of the study was twofold: first, to estimate the adjusted effect of the surgical access procedure on the occurrence of complications and, second, to illustrate this effect also on the level of predicted probabilities. Multivariable logistic regression models with the five explanatory variables *type of surgical access*, *logistic EuroSCORE*, *previous cardiac operations*, *BMI*, and *diabetes* were estimated by ML, WF, FL, FLAC, LF, CP, and RR. In addition, predicted probabilities were obtained by FLIC, KAB, and KAU. All methods gave significant, large effect estimates for *type of surgical access* ranging

**Table IV.** Coverage, power, and standardized length (×1000) of 95% confidence intervals (CIs) for effect estimates averaged over all explanatory variables (omitting the intercept), for selected simulation scenarios. (See Table S6 for further scenarios.)

| | | | Coverage (×1000)<br>Effect size ($a$) | | | Power (×1000)<br>Effect size ($a$) | | Length (×1000)<br>Effect size ($a$) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sample size ($N$) | Event rate ($\pi$) | Method | 0 | 0.5 | 1 | 0.5 | 1 | 0 | 0.5 | 1 |
| 500 | 0.05 | ML | 952 | 951 | 953 | 133 | 326 | 1016 | 987 | 1059 |
| | | WF | 945 | 944 | 944 | 136 | 341 | 1019 | 990 | 1061 |
| | | FL/FLIC | 952 | 950 | 954 | 128 | 321 | 986 | 957 | 1019 |
| | | FLAC | 956 | 953 | 950 | 114 | 323 | 945 | 926 | 979 |
| | | LF | 949 | 948 | 951 | 128 | 327 | 1004 | 971 | 1032 |
| | | CP | 959 | 958 | 959 | 116 | 295 | 954 | 930 | 980 |
| | | RR | 996 | 912 | 863 | 45 | 238 | 462 | 536 | 772 |
| | 0.10 | ML | 949 | 946 | 946 | 197 | 477 | 715 | 707 | 763 |
| | | WF | 945 | 943 | 943 | 199 | 488 | 716 | 708 | 763 |
| | | FL/FLIC | 949 | 948 | 948 | 192 | 472 | 702 | 694 | 745 |
| | | FLAC | 952 | 947 | 946 | 184 | 476 | 690 | 684 | 731 |
| | | LF | 946 | 944 | 946 | 195 | 478 | 711 | 702 | 753 |
| | | CP | 954 | 952 | 950 | 183 | 447 | 693 | 686 | 732 |
| | | RR | 995 | 896 | 900 | 91 | 378 | 401 | 462 | 645 |
| 3000 | 0.01 | ML | 951 | 953 | 951 | 153 | 401 | 877 | 839 | 866 |
| | | WF | 948 | 947 | 945 | 158 | 417 | 881 | 842 | 869 |
| | | FL/FLIC | 949 | 950 | 949 | 155 | 405 | 866 | 828 | 850 |
| | | FLAC | 958 | 956 | 951 | 138 | 402 | 831 | 803 | 826 |
| | | LF | 950 | 950 | 949 | 151 | 406 | 870 | 830 | 851 |
| | | CP | 958 | 958 | 955 | 140 | 373 | 838 | 804 | 823 |
| | | RR | 996 | 903 | 881 | 65 | 310 | 437 | 496 | 685 |

The coverage was determined as the proportion of simulated data sets, where the CI contains the true value, averaged over all explanatory variables (omitting the intercept). The power was calculated as the proportion of simulated data sets, where the CI excludes zero, again averaged over all explanatory variables (omitting the intercept). The last three columns of the table show the standardized length of the CI, averaged over all simulated data sets and all explanatory variables (omitting the intercept).

Effect sizes $a \in \{0, 0.5, 1\}$ refer to scenarios with no, small, and large effects, respectively, are global multipliers of the log odds ratios as described in Section 3.1.

ML, maximum likelihood; WF, weakened Firth's logistic regression; FL, Firth's logistic regression; FLIC, Firth's logistic regression with intercept-correction; FLAC, Firth's logistic regression with added covariate; LF, penalization by log-$F(1, 1)$ priors; CP, penalization by Cauchy priors; RR, ridge regression

For ML, CP, and RR, Wald-type CIs were used. For WF, FL, FLIC, FLAC, and LF, profile likelihood CIs were used.

between an odds ratio of 3.1 for RR and of 5.66 for ML, accompanied by wide CIs but with lower bounds always greater than 1 (Table V). Given the small events-per-variable ratio of 3.2 and the small-sample bias-reducing property of FL, the coefficient estimates by FL might be preferable to the ML estimates in this situation although the difference is not substantial. Coefficient estimates by WF, which can be regarded as a compromise between ML and FL, fell between the ones from ML and FL. None of the four adjustment variables showed a significant effect with the methods ML, WF, FL, FLAC, and LF, but some of them with CP and RR. This is somewhat in contrast to our simulation results, where CIs from RR were rather more likely to include 0 than CIs from other methods, independent of the effect size. We explain this discrepancy by one of the assumptions of our simulation study, assuming that in each scenario, all explanatory variables had similar effect sizes. Counter to intuition, RR may induce bias away from zero in the effect estimates of irrelevant variables that are correlated to strong predictors.

Figure 2(a) shows the distribution of predicted probabilities by method, separately for patients with and without complications. As expected, both FLIC and FLAC pushed the predicted probabilities towards zero, resulting in values clearly smaller than the FL counterparts. On the contrary, KAB even increased the FL predictions in magnitude and supplied the largest individual predicted probabilities both, for the event (24%) and non-event group (33.9%). RR had the smallest range of predicted probabilities. The observed proportion of events was overestimated most severely with KAB (by 30.1%), with FL by 15%,

**Table V.** Odds ratios with 95% confidence intervals (CIs).

| | ML | WF | FL/FLIC | FLAC | LF | CP | RR |
|---|---|---|---|---|---|---|---|
| *Type of surgical access* | 5.66 | 5.61 | 5.38 | 4.93 | 4.99 | 4.88 | 3.1 |
| (conv. vs. ACD) | (1.89,16.95) | (1.9,17.35) | (1.88,15.97) | (1.73,14.47) | (1.73,14.91) | (2.81,8.48) | (2.07,4.64) |
| *Logistic EuroSCORE* | 1.36 | 1.36 | 1.37 | 1.31 | 1.36 | 1.34 | 1.23 |
| (standardized) | (0.9,2.05) | (0.87,2) | (0.89,1.98) | (0.86,1.9) | (0.86,1.99) | (1.23,1.45) | (1.16,1.31) |
| *Previous cardiac operations* | 3.39 | 3.43 | 3.56 | 2.98 | 2.87 | 2.83 | 2.24 |
| (yes vs. no) | (0.79,14.61) | (0.69,13.37) | (0.79,13.02) | (0.64,11.12) | (0.61,10.93) | (1.06,7.54) | (1,5.02) |
| *BMI* | 0.7 | 0.71 | 0.73 | 0.73 | 0.72 | 0.74 | 0.84 |
| (standardized) | (0.39,1.27) | (0.37,1.23) | (0.39,1.25) | (0.41,1.23) | (0.39,1.24) | (0.63,0.86) | (0.77,0.92) |
| *Diabetes* | 1.79 | 1.8 | 1.81 | 1.7 | 1.7 | 1.68 | 1.45 |
| (yes vs. no) | (0.57,5.59) | (0.54,5.34) | (0.57,5.18) | (0.53,4.87) | (0.53,4.94) | (0.92,3.06) | (0.93,2.26) |

ML, maximum likelihood; WF, weakened Firth's logistic regression; FL, Firth's logistic regression; FLIC, Firth's logistic regression with intercept-correction; FLAC, Firth's logistic regression with added covariate; LF, penalization by log-*F*(1, 1) priors; CP, penalization by Cauchy priors; RR, ridge regression

For ML, CP, and RR, Wald-type CIs were used. For WF, FL, FLIC, FLAC, and LF, profile likelihood CIs were used.

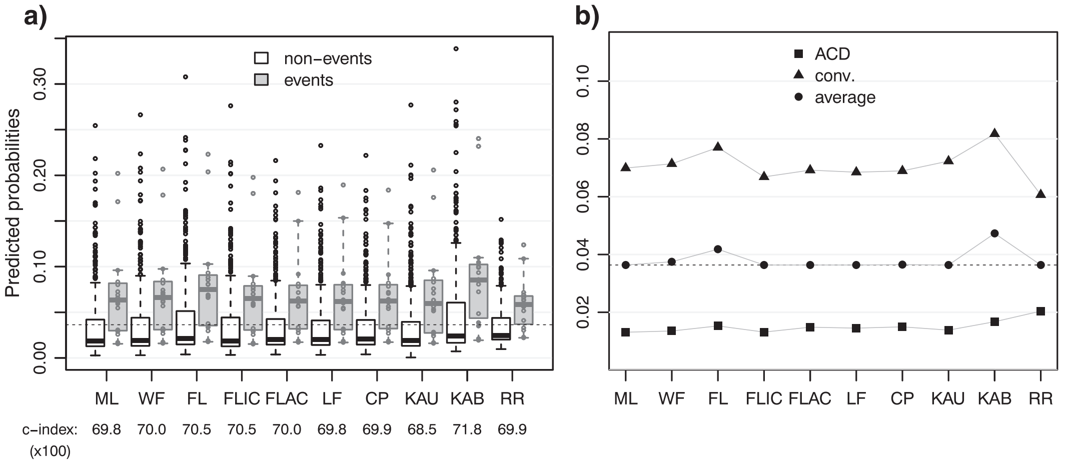

**Figure 2.** (a) Boxplots of predicted probabilities by method and occurrence of event. C-indices were estimated using .632+ bootstrap with 200 repetitions (cf. [12]). (b) Rectangles and triangles give the predicted probabilities for patients with typical (median) covariate values (logistic EuroSCORE of 5.82, without previous cardiac operation, BMI of 26.6, suffering from diabetes) and with either arterial closure device (ACD) or conventional surgical access, respectively. Circles mark the average predicted probabilities. In both plots, the horizontal dashed line marks the observed proportion of events (3.6%). ML, maximum likelihood; WF, weakened Firth's logistic regression; FL, Firth's logistic regression; FLIC, Firth's logistic regression with intercept-correction; FLAC, Firth's logistic regression with added covariate; LF, penalization by log-*F*(1, 1) priors; CP, penalization by Cauchy priors; KAU, King and Zeng's approximate unbiased method; KAB, King and Zeng's approximate Bayesian method; RR, ridge regression.

with WF by 3.1%, and with CP by only 0.3% (Figure 2 (b)). The estimated event probabilities for patients with average covariate values (median logistic EuroSCORE of 5.82, without previous cardiac operation, median BMI of 26.6, suffering from diabetes) and w7ith either ACD or conventional surgical access shown in Figure 2(b) exhibit a similar pattern as the boxplots in 2(a). Again, predicted probabilities by RR had the smallest variability. Predicted probabilities by WF fell between predicted probabilities by ML and FL, whereas for the conventional surgical access group, FLIC and FLAC resulted in predicted probabilities smaller than the ones from both, ML and FL. Discrimination in terms of cross-validated c-indices (Figure 2) was best for KAB and FL (FLIC) and was slightly worse for the other methods.

All considered methods gave rise to similar conclusions on the role of ACD use, being associated with a significantly lower complication rate than conventional surgical access. For ACD patients with average covariate values, the different methods predicted complication risks between 1.3% and 2%, while for comparable conventional access, patients predictions ranged between 6.1% and 8.2%.

## 5. Discussion

Our simulation study shows that both our suggested methods, Firth's logistic regression with intercept-correction and Firth's logistic regression with added covariate, efficiently improve on predictions from Firth's logistic regression. The complete removal of the bias in the average predicted probability, which amounted up to 19.4% in our simulations, was accompanied by more accurate individual predicted probabilities, as revealed by an RMSE of FLIC and FLAC predicted probabilities ranking at most fourth in 42 out of 45 simulation scenarios, only constantly outperformed by RR and never worse than FL. Fortunately, this improvement in predictions has not to be paid for by a lower performance of effect estimation: while for FLIC the effect estimates are identical to those of FL, FLAC introduces some bias, but this is compensated by a decreased RMSE. Based on our simulation results, we slightly prefer FLAC over FLIC because of the lower RMSE of predicted probabilities. CIs for coefficient estimates for both FLIC and FLAC can be easily derived and performed reasonably well in all considered scenarios. Though, in the case of FLIC, the covariance between the intercept and other coefficient estimates cannot be based on model output and if required, would have to be estimated by resampling methods. If CIs for predicted probabilities are needed, one could avoid the determination of a full covariance matrix by centering the explanatory variables at the values for which prediction is requested and re-estimating the model. The CI of the intercept can then be interpreted as a CI for the linear predictor. Finally, the two methods can be readily implemented using only some data manipulation steps, ML estimation and an implementation of FL, which is available in software packages such as SAS, R, Stata, and Statistica [13], [14], [15].

The 'weakened' Firth's penalization, a compromise between ML and FL estimation, indeed performed better than FL with regard to bias and RMSE of predicted probabilities but was outperformed by most of the other methods, by FLIC, FLAC LF, CP, and RR. Moreover, reducing the bias in predicted probabilities compared with FL comes at the cost of enlarging the RMSE of effect estimates, as shown by our simulation results in line with intuition. Of course, WF depends essentially on the choice of the weight parameter $\tau$, which was set to 0.1 as suggested by Elgmati *et al.* (cf. [3]). Future investigations should clarify whether tuning the weight parameter by optimizing for instance the cross-validated modified $AIC_{\text{mod}}$, as performed in ridge regression in our study can make the WF a more attractive option.

King and Zeng's approximate Bayesian method could not make up for the introduction of bias in predicted probabilities, which is exactly twice as large as in FL estimation, but also performed rather poorly with respect to RMSE. Our results even suggest its inferiority with respect to the approximate unbiased method, giving average predicted probabilities equal to the proportions of events with often smaller RMSE than KAB. Thus, we could not confirm King and Zeng's recommendation to prefer KAB over KAU 'in the vast majority of applications'. Though, the discrepancy between their and our simulation results advise caution: it seems that none of the two methods is superior to the other in most situations, but that the behavior strongly depends on the setting. This was also emphasized by a spot-check simulation with balanced outcome, where KAB showed lower RMSE than KAU (results not shown). One disadvantage of KAU is that predicted probabilities can fall outside the plausible range of 0 to 1. However, the question of deciding between KAB and KAU might not be a relevant one, because both methods were clearly outperformed by FLIC, FLAC, CP, LF, and, as long as there is no separation, RR. It should also be taken into account, that with KAB and KAU, the analytical relation between linear predictors and predicted probabilities is lost.

The three methods based on weakly informative priors, penalization by LF, CP, and RR, do not only differ in the choice of the prior distributions but also in the strategy of data preprocessing. While Greenland and Mansournia [4] advocate the use of reasonable, study-independent multiples of SI-units for LF, Gelman *et al.*, [5] suggest to scale continuous variables to have a standard deviation of 0.5 for CP. In the case of RR, we followed the widespread practice to standardize variables to unit variance and to choose the tuning parameter by the penalized version of the AIC (see, for instance, [9]). Of course, it is a legitimate question to ask whether other combinations of strategies, for instance, penalization by LF after stringent scaling of variables or even tuning of prior distribution parameters might yield better performances, but this would go beyond the scope of this study. Instead, we focused on readily available methods, aimed at routine use in statistical applications.

Whenever one is willing to accept introduction of bias towards zero for the sake of a small RMSE, RR turned out to be the method of choice as long as there is no separation. It outperformed all other methods with respect to RMSE of coefficients in 38 and of predicted probabilities in 44 out of 45 simulation scenarios. However, CIs do not reach their nominal coverage levels because of the combination of bias and reduced variance. The naive approach of deducing Wald CIs from the penalized covariance matrix, which was applied in our simulation study, cannot be recommended. In order to avoid these issues in the construction of CIs, CP, or FLAC, providing substantially less biased effect estimates than RR with a reasonable RMSE might be preferable. Moreover, in the presence of separation, CP and FLAC were usually better choices than RR, which often resulted in nonconvergent ML estimation. LF showed a similar performance pattern as CP but was slightly outperformed with regard to the RMSE of coefficients as well as predicted probabilities by CP throughout all simulation scenarios.

When comparing CP to FLAC, the main difference is the application of independent, univariate priors by CP while a multivariate prior is employed by FLAC. In all methods that use independent priors (including LF and RR), linear transformations of explanatory variables, such as scaling, or interaction coding, or to achieve orthogonality, will affect predictions from the model. Therefore, `bayesglm` [16] includes an automatic preprocessing of the explanatory variables following [5]. This preprocessing leads to more stringent penalization of interaction effects than of main effects, which can be desirable in exploratory data analyses. By contrast, in FLAC, any linear transformations of explanatory variables (including different coding) will not affect predictions. Nevertheless, FLAC will apply more shrinkage in larger models, for example, when interaction effects are included. Thus, with limited data, it will penalize complex models more than simple ones.

Summarizing, being interested in accurate effect estimates and predictions in the presence of rare events, we recommend to use

- RR if inference is not required and optimization of the RMSE of coefficients and predicted probabilities is given the priority, and if no separation occurs,
- FLAC or CP as offering low variability and bias in effect estimates as well as predictions, if valid CIs for effects are needed. We have a slight preference for FLAC because of the invariance properties outlined earlier.

## Appendix A

### *A.1. Basic properties of Firth's logistic regression with added covariate*

- **With FLAC, the average predicted probability is equal to the observed proportion of events.** Recall that FLAC amounts to ML for an augmented data set taking into account an indicator variable, which discriminates between original and pseudo-observations. With ML, the average predicted probability is not only equal to the observed proportion of events in total but also within each category of a binary explanatory variable $x_j$. This can be easily seen from the score equation for the intercept $\partial l/\beta_0 = \sum_{i=1}^{N} y_i - \pi_i = 0$ and from the score equation for the binary variable of interest $\partial l/\beta_j = \sum_{i=1}^{N} (y_i - \pi_i)x_{ij} = 0$, where the binary variable $x_j$ takes the values 0 and 1. Consequently, in the FLAC algorithm, the ML estimation step yields average predicted probabilities equal to the observed event rates within both categories of the binary indicator variable. In particular, we conclude that the average predicted probability by FLAC agrees with the observed event rate.
- **In univariate analysis with a binary explanatory variable $x$, the estimated odds ratio by FLAC is closer to 1 than the odds ratio estimated by ML.** Let $\hat{\beta}_{\mathrm{ML},1}$ and $\hat{\beta}_{\mathrm{FLAC},1}$ denote the log odds ratios estimated by ML and FLAC, respectively. Then, the condition implies that $|\hat{\beta}_{\mathrm{FLAC},1}| \leqslant |\hat{\beta}_{\mathrm{ML},1}|$, which can be proven by transforming the score equations. Assume that we observe the counts

|  |  | x |  |
|---|---|---|---|
|  |  | 0 | 1 |
| y | 0 | $n_{00}$ | $n_{10}$ |
|  | 1 | $n_{01}$ | $n_{11}$ |

then the score equations for FLAC are equivalent to the following system of equations

$$\begin{aligned} n_{01} + 0.5 &= (n_{00} + n_{01})\mathrm{expit}(\hat{\beta}_{\mathrm{FLAC},0}) + \mathrm{expit}(\hat{\beta}_{\mathrm{FLAC},0} + \hat{\beta}_{\mathrm{ind}}) \\ n_{11} + 0.5 &= (n_{10} + n_{11})\mathrm{expit}(\hat{\beta}_{\mathrm{FLAC},0} + \hat{\beta}_{\mathrm{FLAC},1}) + \mathrm{expit}(\hat{\beta}_{\mathrm{FLAC},0} + \hat{\beta}_{\mathrm{FLAC},1} + \hat{\beta}_{\mathrm{ind}}) \\ 1 &= \mathrm{expit}(\hat{\beta}_{\mathrm{FLAC},0} + \hat{\beta}_{\mathrm{ind}}) + \mathrm{expit}(\hat{\beta}_{\mathrm{FLAC},0} + \hat{\beta}_{\mathrm{FLAC},1} + \hat{\beta}_{\mathrm{ind}}), \end{aligned}$$

where $\hat\beta_{\mathrm{FLAC},0}$ is the intercept, $\hat\beta_{\mathrm{FLAC},1}$ is the regression coefficient for $x$, $\hat\beta_{\mathrm{ind}}$ is the coefficient of the indicator variable discriminating between original and pseudo data, and $\mathrm{expit}(t) = (1+\exp(-t))^{-1}$. Because $1-\mathrm{expit}(t) = \mathrm{expit}(-t)$, we find from the third equation earlier that $\mathrm{expit}(-\hat\beta_{\mathrm{FLAC},0} - \hat\beta_{\mathrm{ind}}) = \mathrm{expit}(\hat\beta_{\mathrm{FLAC},0} + \hat\beta_{\mathrm{FLAC},1} + \hat\beta_{\mathrm{ind}})$, that is, $-\hat\beta_{\mathrm{FLAC},0} - \hat\beta_{\mathrm{ind}} = \hat\beta_{\mathrm{FLAC},0} + \hat\beta_{\mathrm{FLAC},1} + \hat\beta_{\mathrm{ind}}$. This implies that $\hat\beta_{\mathrm{FLAC},0} + \hat\beta_{\mathrm{ind}} = -\hat\beta_{\mathrm{FLAC},1}/2$ and that $\hat\beta_{\mathrm{FLAC},0} + \hat\beta_{\mathrm{FLAC},1} + \hat\beta_{\mathrm{ind}} = \hat\beta_{\mathrm{FLAC},1}/2$, which allows us to rewrite the first two equations as

$$\frac{n_{01} + 0.5 - \mathrm{expit}(-\hat\beta_{\mathrm{FLAC},1}/2)}{n_{00}+n_{01}} = \mathrm{expit}(\hat\beta_{\mathrm{FLAC},0}) \tag{A.1a}$$

$$\frac{n_{11} + 0.5 - \mathrm{expit}(\hat\beta_{\mathrm{FLAC},1}/2)}{n_{10}+n_{11}} = \mathrm{expit}(\hat\beta_{\mathrm{FLAC},0} + \hat\beta_{\mathrm{FLAC},1}). \tag{A.1b}$$

The ML score equations can be brought into a similar form

$$\frac{n_{01}}{n_{00}+n_{01}} = \mathrm{expit}(\hat\beta_{\mathrm{ML},0}) \tag{A.2a}$$

$$\frac{n_{11}}{n_{10}+n_{11}} = \mathrm{expit}(\hat\beta_{\mathrm{ML},0} + \hat\beta_{\mathrm{ML},1}). \tag{A.2b}$$

Consider the case that $\hat\beta_{\mathrm{FLAC},1}$ is greater than 0. Then, the expression $0.5 - \mathrm{expit}(-\hat\beta_{\mathrm{FLAC},1}/2)$ is also greater than 0 and, comparing Equations A.1a and A.2a, we find that $\mathrm{expit}(\hat\beta_{\mathrm{FLAC},0})$ is greater than $\mathrm{expit}(\hat\beta_{\mathrm{ML},0})$. Because the function expit is strictly monotonically increasing, we conclude that $\hat\beta_{\mathrm{FLAC},0}$ is greater than $\hat\beta_{\mathrm{ML},0}$. Similarly, comparing Equations A.1b and A.2b, we find that $\hat\beta_{\mathrm{FLAC},0} + \hat\beta_{\mathrm{FLAC},1}$ is smaller than $\hat\beta_{\mathrm{ML},0} + \hat\beta_{\mathrm{ML},1}$. From the two inequalities, we can see easily that $\hat\beta_{\mathrm{FLAC},1}$ is smaller than $\hat\beta_{\mathrm{ML},1}$. In a similar way, one can show that if $\hat\beta_{\mathrm{FLAC},1} < 0$, then $\hat\beta_{\mathrm{ML},1} < \hat\beta_{\mathrm{FLAC},1}$.

### *A.2. Structure of explanatory variables in the simulation study*

Table A.1 gives information on the construction of the 10 explanatory variables used in the simulation study. First, 10 standard normal random variables $z_{ij} \sim \mathcal{N}(0,1)$, $j = 1, \dots, 10$, $i = 1, \dots N$, with correlation structure as listed in the second column of Table A.1 were generated. By applying the transformations described in the third column to the variables $z_{ij}$, four continuous, four binary, and two ordinal variables $x_{ij}$ were derived. The continuous variables $x_{i1}, x_{i4}, x_{i5}$, and $x_{i8}$ were truncated at the third quartile plus five times the interquartile distance in each simulated data set.

**Table A.1.** Structure of explanatory variables in the simulation study, following Binder, Sauerbrei und Royston [11]. Square brackets $[\dots]$ indicate that the non-integer part of the argument is removed. $\mathbb{1}$ denotes the indicator function, taking value 1 if its argument is true and 0 otherwise.

| Underlying variable | Correlation of underlying variables | Explanatory variable | Type | Correlation of explanatory variables |
|---|---|---|---|---|
| $z_{i1}$ | $z_{i2}(0.8), z_{i7}(0.3)$ | $x_{i1} = [10z_{i1} + 55]$ | continuous | $x_{i2}(-0.6), x_{i7}(0.2)$ |
| $z_{i2}$ | $z_{i1}(0.8)$ | $x_{i2} = \mathbb{1}_{\{z_{i2}<0.6\}}$ | binary | $x_{i1}(-0.6)$ |
| $z_{i3}$ | $z_{i4}(-0.5), z_{i5}(-0.3)$ | $x_{i3} = \mathbb{1}_{\{z_{i3}\geqslant -1.2\}} + \mathbb{1}_{\{z_{i3}\geqslant 0.75\}}$ | ordinal | $x_{i4}(-0.4), x_{i5}(-0.2)$ |
| $z_{i4}$ | $z_{i3}(-0.5), z_{i5}(0.5), z_{i7}(0.3)$ $z_{i8}(0.5), z_{i9}(0.3)$ | $x_{i4} = [\max(0, 100\exp(z_{i4}) - 20)]$ | continuous | $x_{i3}(-0.4), x_{i5}\ (0.4), x_{i7}(0.2),$ $x_{i8}(0.4), x_{i9}(-0.2)$ |
| $z_{i5}$ | $z_{i3}(-0.3), z_{i4}(0.5), z_{i8}(0.3),$ $z_{i9}(0.3)$ | $x_{i5} = [\max(0, 80\exp(z_{i5}) - 20)]$ | continuous | $x_{i3}(-0.2), x_{i4}(0.4), x_{i8}(0.2),$ $x_{i9}(-0.2)$ |
| $z_{i6}$ | $z_{i7}(-0.3), z_{i8}(0.3)$ | $x_{i2} = \mathbb{1}_{\{z_{i6}<-0.35\}}$ | binary | $x_{i7}(0.2), x_{i8}(-0.2)$ |
| $z_{i7}$ | $z_{i1}(0.3), z_{i4}(0.3), z_{i6}(-0.3)$ | $x_{i7} = \mathbb{1}_{\{z_{i3}\geqslant 0.5\}} + \mathbb{1}_{\{z_{i3}\geqslant 1.5\}}$ | ordinal | $x_{i1}(0.2), x_{i4}(0.2), x_{i6}(0.2)$ |
| $z_{i8}$ | $z_{i4}(0.5), z_{i5}(0.3), z_{i6}(0.3)$ $z_{i9}(0.5)$ | $x_{i8} = [10z_{i8} + 55]$ | continuous | $x_{i4}(0.4), x_{i5}(0.2), x_{i6}(-0.2),$ $x_{i9}(-0.4)$ |
| $z_{i9}$ | $z_{i4}(0.3), z_{i5}(0.3), z_{i8}(0.5)$ | $x_{i9} = \mathbb{1}_{\{z_{i9}<0\}}$ | binary | $x_{i4}(-0.2), x_{i5}(-0.2), x_{i8}(-0.4)$ |
| $z_{i10}$ | – | $x_{i10} = \mathbb{1}_{\{z_{i10}<0\}}$ | binary | – |

## Acknowledgements

We thank Michael Schemper, Mohammad Ali Mansournia, the editor, and two referees for helpful comments. The data on complications in cardiac surgery were kindly provided by Paulo A. Amorim, Alexandros Moschovas, Gloria Färber, Mahmoud Diab, Tobias Bünger, and Torsten Doenst from the Department of Cardiotheracic Survergy at the University Hospital Jena. The research leading to these results has received funding from the Austrian Science Fund (FWF): I 2276 and from the Slovenian Research Agency (ARRS) within project N1-0035.

## References

1. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika* 1993; **80**(1):27–38.
2. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Statistics in Medicine* 2002; **21**(16):2409–2419.
3. Elgmati E, Fiaccone RL, Henderson R, Matthews JNS. Penalised logistic regression and dynamic prediction for discrete-time recurrent event data. *Lifetime Data Analysis* 2015; **21**(4):542–560.
4. Greenland S, Mansournia MA. Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine* 2015; **34**(23):3133–3143.
5. Gelman A, Jakulin A, Pittau MG, Su Y-S. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2008; **2**(4):1360–1383.
6. King G, Zeng L. Logistic regression in rare events data. *Political Analysis* 2001; **9**(2):137–163.
7. Greenland S. Simpson's paradox from adding constants in contingency tables as an example of bayesian noncollapsibility. *The American Statistician* 2010; **64**(4):340–344.
8. Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. *Applied Statistics* 1992; **41**(1):191–201.
9. Verweij PJM, Van Houwelingen JC. Penalized likelihood in Cox regression. *Statistics in Medicine* 1994; **13**(23-24): 2427–2436.
10. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; **21**(1):128–138.
11. Binder H, Sauerbrei W, Royston P. Multivariable model-building with continuous covariates: 1. Performance measures and simulation design, Technical Report FDM-Preprint 105, University of Freiburg Germany, 2011.
12. Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. *Journal of the American Statistical Association* 1997; **92**(438):548–560.
13. Heinze G, Ploner M. A SAS macro, S-PLUS library and R Package to perform logistic regression without convergence problems, Technical Report 2, Medical University of Vienna Austria. http://tinyurl.com/fllogistfTR, 2004.
14. Coveney J. *FIRTHLOGIT: Stata module to calculate bias reduction in logistic regression*, 2008. http://EconPapers.repec.org/RePEc:boc:bocode:s456948. (accessed on 23 February 2017).
15. Fijorek K, Sokolowski A. Separation-resistant and bias-reduced logistic regression: STATISTICA macro. *Journal of Statistical Software, Code Snippets* 2012; **47**:1–12.
16. Gelman A, Su Y-S. *arm: data analysis using regression and multilevel/hierarchical models*, 2015. http://CRAN.R-project.org/package=arm, R package version 1.9-3. (accessed on 10 January 2015).

## Supporting information

Additional supporting information may be found online in the supporting information tab for this article.