ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2019)
68, *Part* 1, *pp.* 121–139

Informing a risk prediction model for binary outcomes with external coefficient information

Wenting Cheng, Jeremy M. G. Taylor, Tian Gu, Scott A. Tomlins and Bhramar Mukherjee

University of Michigan, Ann Arbor, USA

[Received June 2017. Final revision June 2018]

Summary. We consider a situation where rich historical data are available for the coefficients and their standard errors in an established regression model describing the association between a binary outcome variable Y and a set of predicting factors $\mathbf{X}$, from a large study. We would like to utilize this summary information for improving estimation and prediction in an expanded model of interest, $Y|\mathbf{X}, B$. The additional variable B is a new biomarker, measured on a small number of subjects in a new data set. We develop and evaluate several approaches for translating the external information into constraints on regression coefficients in a logistic regression model of $Y|\mathbf{X}, B$. Borrowing from the measurement error literature we establish an approximate relationship between the regression coefficients in the models $\Pr(Y = 1|\mathbf{X}, \beta)$, $\Pr(Y = 1|\mathbf{X}, B, \gamma)$ and $E(B|\mathbf{X}, \theta)$ for a Gaussian distribution of B. For binary B we propose an alternative expression. The simulation results comparing these methods indicate that historical information on $\Pr(Y = 1|\mathbf{X}, \beta)$ can improve the efficiency of estimation and enhance the predictive power in the regression model of interest $\Pr(Y = 1|\mathbf{X}, B, \gamma)$. We illustrate our methodology by enhancing the high grade prostate cancer prevention trial risk calculator, with two new biomarkers: prostate cancer antigen 3 and TMPRSS2:ERG.

Keywords: Bayesian methods; Constrained estimation; Logistic regression; Prediction models

1. Introduction

Risk prediction models for different binary disease end points are abundant in the clinical and epidemiological literature. Examples of established models are the breast cancer risk calculator (Gail *et al.*, 1989) and the Framingham risk score (D'Agostino *et al.*, 2001) which can be used to assess an individual's risk of experiencing a future health event and to make decisions concerning screening and prophylactic prevention. As a motivating example in this paper, the prostate cancer prevention trial (PCPT) risk calculator (Thompson *et al.*, 2006) is an on-line assessment tool which provides a personalized risk estimate of detecting prostate cancer based on risk factors such as age, prostate-specific antigen (PSA) and digital rectal examination (DRE) findings.

Whereas these established models are often based on standard epidemiologic and behavioural risk factors, wider availability of high throughput data and novel assay technologies are generating candidate biomarkers for potential inclusion in existing prediction models. It is very likely that the new biomarkers are assessed only on subjects in a study of moderate size and cannot be measured on the much larger population that is used for the well-established model. Investigators could directly estimate the expanded model in the new data set, but results from this expanded prediction model based solely on a limited number of subjects could be highly

Address for correspondence: Wenting Cheng, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.
E-mail: chengwt@umich.edu

 0035–9254/19/68121

variable. It is natural to consider using the information from the well-established model to increase the accuracy of the expanded model.

Substantial research has been done on the problem of enhancing risk prediction models with supplemental external information. The external information may be used to combine estimates from previous studies with the regression coefficients that are estimated in the new data set. Steyerberg *et al.* (2000) described a method to adjust the multivariate logistic regression model's coefficients estimated in a data set based on univariate regression models' coefficients in the literature. Newcombe *et al.* (2012) presented two possible approaches incorporating the effect estimates of a set of predictors: the first was by adding a composite weighted risk score based on these estimates and the second was by specifying informative priors for the coefficients of these variables in a Bayesian logistic regression model. Chatterjee *et al.* (2016) developed a general method for incorporating external coefficients, derived from constrained estimating equations. Other related approaches used constrained maximum likelihood and empirical likelihood (Imbens and Lancaster, 1994; Qin, 2000; Qin *et al.*, 2015). Cheng *et al.* (2018) developed and compared various approaches for the situation when the outcome variable is continuous. They established exact relationships between the parameters in the model of interest that includes the new biomarker and the parameters in the established model, and then proposed both frequentist and Bayesian approaches. In the current paper we adapt and extend the approaches to the situation when the outcome variable is binary.

There are also some simple approaches. For the Gail model, Mealiffe *et al.* (2010) computed a multiplicative risk score based on previously published odds ratios of newly discovered biomarkers. They then multiplied the Gail risk estimate and the multiplicative risk score to give a combined risk score. Grill *et al.* (2015) proposed a simple method of incorporating new markers via Bayes theorem. They updated the posterior odds of developing cancer based on both standard risk factors and new markers by using the likelihood ratio incorporating dependence between the two sets of risk factors to adjust the prior odds of developing cancer based on standard risk alone. Grill *et al.* (2017) assessed the performance of a set of likelihood ratio approaches as well as the approach that was proposed in Chatterjee *et al.* (2016).

We consider a situation where the outcome is a binary indicator of disease and the well-established model is described in a published paper, in which the estimated regression coefficients and their standard errors are presented. The expanded model includes one additional biomarker as a potential predictor. To introduce notation, let Y denote the binary outcome, $\mathbf{X}$ is a set of p standard risk factors and B is a new biomarker. The association between Y and $\mathbf{X}$ is described through the following logistic model:

$$\text{logit}\{\Pr(Y=1|\mathbf{X})\} = \mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p. \tag{1}$$

We assume that we have available summary level information on the estimated regression coefficients $\bar{\boldsymbol{\beta}}$ and their standard errors $\bar{\mathbf{S}}$ in model (1). On the basis of the work that went into establishing this model, we shall assume that all the $\mathbf{X}$s are deemed to be important and need to be included in any model, and further that the above form provides at least a good approximation to the distribution of Y given $\mathbf{X}$.

The model of primary interest is a model that describes the joint effect of $\mathbf{X}$ and B on Y:

$$\text{logit}\{\Pr(Y=1|\mathbf{X}, B)\} = \mathbf{X}\boldsymbol{\gamma}_{\mathbf{X}} + B\gamma_B = \gamma_0 + \gamma_1 X_1 + \ldots + \gamma_p X_p + \gamma_{p+1} B. \tag{2}$$

Our goal is to obtain the best estimate we can of the γs in a model of this form, making use of all the available information from the established model and the small data set.

Another model that can be estimated from the current small data set is

$$E(B|\mathbf{X}) = g^{-1}(\mathbf{X}\boldsymbol{\theta}) = g^{-1}(\theta_0 + \theta_1 X_1 + \ldots + \theta_p X_p) \quad (3)$$

where g is the link function, which is the identity function $g(y) = y$ for Gaussian B and the logit function $g(y) = \log\{y/(1-y)\}$ for binary B. We propose to formulate the problem in an inferential framework where the historical information is translated in terms of non-linear constraints on the regression parameters. The distribution of B will greatly affect how we translate the historical information into constraints on the regression parameters. We consider the cases that B is either Gaussian or binary.

The following description is the structure of the remainder of this paper: in Section 2 we describe the PCPT risk calculator and the available data including the new biomarkers that might be able to enhance this calculator. In Section 3, we establish a relationship equation between the regression coefficients of models (1)–(3) when B is Gaussian. In Section 4, we consider the situation when B is binary and derive the corresponding constrained solutions. We present a simulation study in Section 5. In Section 6 we demonstrate the proposed approaches for the high grade PCPT risk calculator. Concluding remarks are presented in Section 7.

Two simulated data sets and the programs that were used to analyse them can be obtained from

`http://rss.onlinelibrary.wiley.com/hub/journal/14679876/series-c-datasets`

2. A motivating example: prostate cancer risk prediction

The PCPT was a phase III randomized placebo-controlled trial of drug finasteride for the prevention of carcinoma of the prostate. The PCPT randomly assigned 18882 men who were at least 55 years old and did not have prostate cancer with either finasteride or placebo for 7 years. At the end of the 7 years of the study, all men who had not been diagnosed with prostate cancer during the trial were asked to undergo an end-of-study prostate biopsy. The biopsy result could be no cancer, low grade cancer or high grade cancer, which was defined as a Gleason score of 7 or higher. Variables that were collected in this trial included family history of prostate cancer, age, race, previous biopsy result, PSA and DRE.

The use of PSA to screen for prostate cancer had been controversial because the test has low specificity and can lead to overtreatment. Therefore, improved tests that use additional information are needed. The PCPT risk calculator for prostate cancer, PCPTrc, and a separate calculator for high grade prostate cancer, PCPThg (Thompson *et al.*, 2006), were the first online prostate cancer risk assessment tools to allow an individual to assess his risk of prostate cancer. These calculators are well established and are frequently used. These calculators were developed from 5519 men in the placebo group of the PCPT who underwent prostate biopsy. The calculator PCPThg (version 1.0) predicts the chance of high grade prostate cancer based on PSA level, age, DRE findings, prior biopsy result and race:

$$\log\left(\frac{p_i}{1-p_i}\right) = -6.25 + 0.03\,\text{age}_i + 0.96\,\text{race}_i + 1.29\log(\text{PSA}_i) + 1.00\,\text{DRE}_i - 0.36\,\text{biopsy}_i \quad (4)$$

where p_i is the probability of observing high grade prostate cancer for subject i. If we plug in a person's age, race, PSA level, DRE result and previous biopsy information, we can estimate the probability of detecting high grade prostate cancer. The estimated logistic model's coefficients and the 95% confidence intervals are available in Thompson *et al.* (2006). The estimated coefficients and covariance–variance matrices were also accessible as an R code document at `http://deb.uthscsa.edu/URORiskCalc/Pages/calcs.jsp`.

The PCPT risk calculators are based on standard clinical, demographic and epidemiologic variables. None of the variables are related to the molecular mechanisms of carcinogenesis or prostate cancer disease progression. It is plausible to think that including other variables that are more related to the biology of cancer would lead to improved ability to detect prostate cancer. Prostate cancer antigen 3 (PCA 3) and TMPRSS2:ERG gene fusions are two prostate cancer biomarkers which have been shown to have better specificity for early detection of prostate cancer than PSA (Truong *et al.*, 2013; Tomlins *et al.*, 2015). Their transcripts are detectable and quantifiable in urine collected after DRE. To investigate whether PCA 3 and TMPRSS2:ERG could be combined with the PCPThg-calculator to give more accurate risk prediction, Tomlins *et al.* (2015) undertook a study in 679 men, in whom all the PCPThg-calculator variables and both a PCA 3 score and a TMPRSS2:ERG score were measured. In this data set the proportion with high grade prostate cancer is 26.4%. An independent validation study of 1218 men was also available. In this data set the proportion with high grade prostate cancer is 18.3%.

Tomlins *et al.* (2015) expanded the PCPThg-model by incorporating PCA 3 as an additional risk factor. They used the predicted risk score from PCPThg (i.e. $\widehat{\Pr}(Y_i = 1|X_i, \bar{\beta}_{\mathrm{PCPThg}}) \times 100$) directly as a predicting variable and estimated the joint effect of the PCPThg risk score and the PCA 3 value on the probability of high grade prostate cancer. They estimated the new model in the training data set and found that when applied to the validation data set the area under the curve (AUC) increased from 0.707 for the PCPThg-model to 0.752 for their model. They also constructed another expanded PCPThg-model by incorporating TMPRSS2:ERG and showed that the AUC increased from 0.707 to 0.754. We would like to propose more sophisticated statistical approaches that could potentially provide further improvement compared with these results.

3. Statistical approaches

3.1. Logistic regression approximation of the marginal $Pr(Y = 1|\mathbf{X})$

A difficulty in translating the summary information from modelling $\Pr(Y = 1|\mathbf{X})$ to modelling $\Pr(Y = 1|\mathbf{X}, B)$ is that a logistic model $\mathrm{logit}\{\Pr(Y = 1|\mathbf{X}, B)\}$ does not reduce to a logistic model $\mathrm{logit}\{\Pr(Y = 1|\mathbf{X})\}$ when marginalized over the distribution of B. To connect the regression coefficients in models (1)–(3), we need to approximate $\mathrm{logit}\{\Pr(Y = 1|\mathbf{X})\}$ written in terms of parameters γ and $\boldsymbol{\theta}$ and variables $\mathbf{X}$, and to equate that to $\mathrm{logit}\{\Pr(Y = 1|\mathbf{X})\} = \mathbf{X}\boldsymbol{\beta}$. To achieve this, we consider the following integral:

$$\begin{aligned}\Pr(Y = 1|\mathbf{X}) &= \int \Pr(Y = 1|\mathbf{X}, B)\, P(B|\mathbf{X})\, \mathrm{d}B \\ &= \frac{1}{(2\pi)^{1/2}\sigma_2} \int H(\mathbf{X}\boldsymbol{\gamma}_{\mathbf{x}} + B\gamma_{p+1}) \exp\left\{ -\frac{(B - \mathbf{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{B} - \mathbf{X}\boldsymbol{\theta})}{2\sigma_2^2} \right\} \mathrm{d}B \qquad (5)\end{aligned}$$

where $H(v) = \{1 + \exp(-v)\}^{-1}$, and $B|\mathbf{X}$ follows a Gaussian distribution $N(\mathbf{X}\boldsymbol{\theta}, \sigma_2^2)$. The integral in equation (5) does not have a closed form solution and approximations are necessary.

By a normal scale mixture representation of the logistic distribution function and computation of the logistic–normal integral (Monahan and Stefanski, 1992), we can find an approximated equation:

$$\Pr(Y = 1|\mathbf{X}) \approx H\left\{ \frac{\mathbf{X}\boldsymbol{\gamma}_x + (\mathbf{X}\boldsymbol{\theta})\gamma_{p+1}}{(1 + \gamma_{p+1}^2\sigma_2^2/1.7^2)^{1/2}} \right\}.$$

The derivation of the approximation is given in the on-line supplementary material appendix B. Using this approximation, we find an approximate relationship between $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$:

$$\beta_j \approx \frac{\gamma_j + \gamma_{p+1}\theta_j}{(1+\gamma_{p+1}^2\sigma_2^2/1.7^2)^{1/2}}, \qquad j=0,\ldots,p. \tag{6}$$

3.2. Firth correction in logistic regression

The Firth correction (Firth, 1993) is a general approach to reduce bias in maximum likelihood estimation by maximizing a penalized log-likelihood function, where the penalty is $\frac{1}{2}|\mathbf{I}|$ and $\mathbf{I}$ is the information matrix. In logistic regression, standard maximum likelihood estimates often experience serious bias or even non-existence due to separability and multicollinearity, and the Firth correction is suggested (Heinze and Schemper, 2002) as a way to improve the maximum likelihood estimates. In our constrained solution, we add the Firth correction to stabilize the estimates from standard logistic regression.

3.3. Unconstrained solutions

3.3.1. Direct regression

Without constraints, the unknown parameters $\boldsymbol{\gamma}$ in model (2) can be estimated by maximizing the likelihood. Specifically, the estimate solves

$$\max_{\boldsymbol{\gamma}}\left(\sum_{i=1}^{n}\left[Y_i\left(\sum_{j=0}^{p}\gamma_j X_{ij}+\gamma_{p+1}B_i\right)-\log\left\{1+\exp\left(\sum_{j=0}^{p}\gamma_j X_{ij}+\gamma_{p+1}B_i\right)\right\}\right]\right). \tag{7}$$

In addition, we implement Firth's penalized likelihood approach by using R package `logistf` (Heinze *et al.*, 2013). We use least squares to estimate $\boldsymbol{\theta}$ in model (3).

3.3.2. Standard Bayes method

Draws for the posterior distributions of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ are obtained by using flat conjugate priors for $\boldsymbol{\theta}$ and weakly informative Cauchy distribution priors for $\boldsymbol{\gamma}$, as described in the on-line supplementary materials appendix A.

3.4. Constrained solutions

3.4.1. Constrained maximum likelihood

The constrained maximum likelihood estimation maximizes the joint log-likelihood under the set of constraints generated on the basis of the approximate relationship equations (6). We shall require the parameter estimates for $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ to result in the derived value of $\boldsymbol{\beta}$ being within d standard errors of the old point estimates:

$$\begin{aligned}\min_{\boldsymbol{\gamma},\boldsymbol{\theta}}\Bigg(&\sum_{i=1}^{n}\left[-Y_i\left(\sum_{j=0}^{p}\gamma_j X_{ij}+\gamma_{p+1}B_i\right)+\log\left\{1+\exp\left(\sum_{j=0}^{p}\gamma_j X_{ij}+\gamma_{p+1}B_i\right)\right\}\right]\\ &+\sum_{i=1}^{n}\frac{\left(B_i-\sum_{j=0}^{p}\theta_j X_{ij}\right)^2}{2\hat{\sigma}_2^2}\Bigg) \qquad \text{subject to}\\ &\frac{\gamma_j+\gamma_{p+1}\theta_j}{(1+\gamma_{p+1}^2\sigma_2^2/1.7^2)^{1/2}}\in[\bar{\beta}_j-d\bar{S}_j,\bar{\beta}_j+d\bar{S}_j], \quad j=0,\ldots,p.\end{aligned} \tag{8}$$

In this optimization problem, $\hat{\sigma}_2^2$ is a plug-in estimator and is the ordinary least squares residual

variance from fitting $E(B|\mathbf{X})$ and d is a scale parameter representing the strength of external information. From simulations, we find that fixing d as $d=1$ leads to decent properties of the estimates of γ. A modified version that includes the Firth correction is also considered. Further details of these methods are provided in the on-line supplementary materials appendix A. We use the bootstrap as described in supplementary material appendix D to estimate the standard errors.

3.4.2. *Informative full Bayes method*

In informative full Bayes methods, starting with the joint likelihood $L(Y|\mathbf{X}, B)\, L(B|\mathbf{X})$ we translate the constraints in approximation (6) to prior information. The first step is to write down the joint likelihood function with priors on $\gamma, \boldsymbol{\theta}$ and σ_2^2:

$$
\begin{aligned}
p(\gamma, \boldsymbol{\theta}, \sigma_2^2|\text{data}) &\propto L(Y|\mathbf{X}, B, \gamma)\, L(B|\mathbf{X}, \boldsymbol{\theta}, \sigma_2^2)\, \pi(\gamma, \boldsymbol{\theta}, \sigma_2^2) \\
&= \left[\prod_{i=1}^{n} \frac{\exp\left\{ \left(\sum_{j=0}^{p} \gamma_j X_{ij} + \gamma_{p+1} B_i \right) Y_i \right\}}{1 + \exp\left(\sum_{j=0}^{p} \gamma_j X_{ij} + \gamma_{p+1} B_i \right)} \frac{1}{\sqrt{(2\pi\sigma_2^2)}} \exp\left\{ -\frac{1}{2\sigma_2^2} \left(B_i - \sum_{j=0}^{p} \theta_j X_{ij} \right)^2 \right\} \right] \\
&\quad \times \pi(\gamma, \boldsymbol{\theta}, \sigma_2^2) \qquad (9)
\end{aligned}
$$

The logistic regression approximation result (6) suggests that $\theta_j = (1/\gamma_{p+1})\{\beta_j \sqrt{(1+\gamma_{p+1}^2 \sigma_2^2/1.7^2)} - \gamma_j\}$, $j = 0, \ldots, p$. We reparameterize equation (9) in terms of γ, $\boldsymbol{\beta}$ and σ_2^2, and include a Jacobian transformation matrix denoted by $\mathbf{J}$, where $|\mathbf{J}| = (1/|\gamma_{p+1}^{p+1}|)(1 + \gamma_{p+1}^2 \sigma_2^2/1.7^2)^{(p+1)/2}$. We use a non-informative inverse gamma prior for σ_2^2 and independent weakly informative Cauchy priors for γ (Gelman *et al.*, 2008). For parameters $\boldsymbol{\beta}$, we use the constraints as priors:

$$
\beta_j = \frac{\gamma_j + \gamma_{p+1}\theta_j}{(1 + \gamma_{p+1}^2 \sigma_2^2/1.7^2)^{1/2}} \sim N(\bar{\beta}_j, \bar{S}_j^2), \qquad j = 0, \ldots, p. \qquad (10)
$$

Then we can rewrite the joint distribution in terms of $\gamma, \boldsymbol{\beta}$ and σ_2^2 as

$$
\begin{aligned}
p(\gamma, \boldsymbol{\beta}, \sigma_2^2|\mathbf{Y}, \mathbf{X}, B) \propto & \left[\prod_{i=1}^{n} \frac{\exp\left\{ \left(\sum_{j=0}^{p} \gamma_j X_{ij} + \gamma_{p+1} B_i \right) Y_i \right\}}{1 + \exp\left(\sum_{j=0}^{p} \gamma_j X_{ij} + \gamma_{p+1} B_i \right)} \right] \\
& \times \left(\prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi\sigma_2^2)}} \exp\left[-\frac{1}{2\sigma_2^2} \left\{ B_i - \frac{\beta_0 \sqrt{(1 + \gamma_{p+1}^2 \sigma_2^2/1.7^2)} - \gamma_0}{\gamma_{p+1}} \right. \right. \right. \\
& \left. \left. \left. - \sum_{j=1}^{p} \frac{\beta_j \sqrt{(1 + \gamma_{p+1}^2 \sigma_2^2/1.7^2)} - \gamma_j}{\gamma_{p+1}} X_{ij} \right\}^2 \right] \right) \pi(\boldsymbol{\beta})\, \pi(\gamma)\, \pi(\sigma_2^2)\, |\mathbf{J}|.
\end{aligned}
$$

Further details of the priors and the implementation of a Metropolis–Hastings algorithm are given in the on-line supplementary materials appendix B. We note that in the algorithm the full conditional distributions of $\gamma_0, \ldots, \gamma_{p+1}$ and σ_2^2 do not have closed form expressions and, because of the non-linear relationship between the parameters, the Metropolis–Hasting algo-

rithm cannot be performed efficiently and thus it is computationally slow to obtain uncorrelated draws from the posterior distributions.

3.4.3. Transformation approach

As the informative full Bayes method is computationally intensive, we propose an approximate Bayesian approach that can produce draws that fall into the constrained space but reduces the computational burden of the informative Bayes method. The motivation for this stems from the Bayesian transformation approach incorporating monotone or unimodal constraints in posterior inference as proposed in Gunn and Dunson (2005), which we modify to the scenario of a regression model with external coefficient information.

Suppose that the draws from the non-informative standard Bayes method as described in Section 3.3.2 are $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$. The corresponding posterior covariance matrices are $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. We extract the posterior variances from $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ and denote them by $s^2_{\gamma_0}, \ldots, s^2_{\gamma_p}, s^2_{\gamma_{p+1}}, s^2_{\theta_0}, \ldots, s^2_{\theta_p}$. The ordinary least squares residual variance from fitting $E(B|\mathbf{X})$ is $\hat{\sigma}^2_2$. Then a constrained draw $\boldsymbol{\gamma}^*$ and $\boldsymbol{\theta}^*$ is obtained from an unconstrained draw by solving the optimization problem

$$\min_{\boldsymbol{\gamma}^*, \boldsymbol{\theta}^*} \{ d^2_{\mathrm{NED}}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^*) + d^2_{\mathrm{NED}}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \} = \sum_{j=0}^{p+1} \frac{(\gamma_j - \gamma_j^*)^2}{s^2_{\gamma_j}} + \sum_{k=0}^{p} \frac{(\theta_k - \theta_k^*)^2}{s^2_{\theta_k}}$$

$$\text{subject to } \frac{\gamma_j^* + \gamma_{p+1}^* \theta_j^*}{(1 + \gamma_{p+1}^{*2} \hat{\sigma}^2_2 / 1.7^2)^{1/2}} \in [\bar{\beta}_j - d\bar{S}_j, \bar{\beta}_j + d\bar{S}_j], \quad j = 0, \ldots, p, \tag{11}$$

where d_{NED} stands for normalized Euclidean distance. For the transformation of a single draw, we generate d from a half-normal distribution: $d \sim |N(0,1)|$. The intuition behind this transformation procedure is that it will produce values $\boldsymbol{\gamma}^*$ and $\boldsymbol{\theta}^*$ subject to the box constraints that are closest in normalized distance to the unconstrained values $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$.

The transformation is computationally efficient since we have a fast algorithm to solve the optimization problem (11). We fix γ^*_{p+1} and divide the minimization function (11) into $p+1$ two-dimensional constrained minimization problems in which the solutions can be re-expressed as functions of γ^*_{p+1}. After that, the whole minimization problem is reduced to an easy-to-solve one-dimensional optimization problem in γ^*_{p+1}. The constrained draws that are produced by the transformation approach are not draws from the true posterior distribution; however, we found in a limited number of simulations that they are reasonable approximations that can be generated much faster.

3.4.4. Constrained approach of Chatterjee et al. (2016)

For comparison we include a constrained maximum likelihood method that uses the integrated score equations (Chatterjee *et al.*, 2016). The method assumes that the model for $Y|\mathbf{X}, B$ is correct; it does not make any explicit assumptions about the distribution of $B|\mathbf{X}$, but it does require the distribution of $\mathbf{X}$ to be the same in the current sample as in the data that were used to develop the model for $Y|\mathbf{X}$. The method uses only the point estimates $\bar{\beta}$ and does not take into account the standard errors of those estimates.

3.4.5. Logistic regression plug-in method

We also included a simple method which consists of obtaining predicted probabilities by fitting a logistic regression model with two covariates: B and $\log\{\bar{p}_i/(1-\bar{p}_i)\}$, where $\bar{p}_i$ is the prediction from the established model for $Y|\mathbf{X}$. It is easy to show that this method does give a final model

for $Y|\mathbf{X}, B$ that has a logistic link function and is linear in $\mathbf{X}$ and B, and with some algebra the estimates of γ can be obtained.

4. Statistical approaches when *B* is binary

4.1. The approximate relationship equation when B is binary

If B is a binary variable, the logistic regression approximation in Section 3 does not hold and the approximate relationship in equation (6) is not applicable. However, by Bayes theorem, there is a relationship equation connecting $\Pr(Y=1|\mathbf{X})$, $\Pr(Y=1|\mathbf{X}, B)$ and $f(B|\mathbf{X}, Y)$ (Grill *et al.*, 2015; Satten and Kupper, 1993):

$$\frac{\Pr(Y=1|\mathbf{X}, B)}{\Pr(Y=0|\mathbf{X}, B)} = \frac{f(B|\mathbf{X}, Y=1)}{f(B|\mathbf{X}, Y=0)} \frac{\Pr(Y=1|\mathbf{X})}{\Pr(Y=0|\mathbf{X})}. \tag{12}$$

Thus, when B is binary, we need to define a model for $B|\mathbf{X}, Y$ instead of a model for $B|\mathbf{X}$. Assume that $\text{logit}\{\Pr(B=1|\mathbf{X}, Y)\} = \Sigma_{j=0}^{p} \phi_j \mathbf{X}_j + \phi_{p+1} Y$. Some algebraic simplifications of equation (12) followed by a Taylor series expansion (as shown in the on-line supplementary material appendix C) result in an approximate relationship equation:

$$\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \approx \gamma_0 + \tfrac{1}{2}\phi_{p+1} + \tfrac{1}{4}\phi_0\phi_{p+1} + \tfrac{1}{8}\phi_{p+1}^2 + \sum_{j=1}^{p}\left(\gamma_j + \tfrac{1}{4}\phi_j\phi_{p+1}\right) X_j + (\gamma_{p+1} - \phi_{p+1}) B.$$

Then the approximate relationship between $\boldsymbol{\gamma}, \boldsymbol{\phi}$ and $\boldsymbol{\beta}$ is

$$\left.\begin{aligned} \beta_0 &\approx \gamma_0 + \tfrac{1}{2}\phi_{p+1} + \tfrac{1}{4}\phi_0\phi_{p+1} + \tfrac{1}{8}\phi_{p+1}^2, \\ \beta_j &\approx \gamma_j + \tfrac{1}{4}\phi_j\phi_{p+1}, \qquad j = 1, \ldots, p, \\ \gamma_{p+1} &= \phi_{p+1}. \end{aligned}\right\} \tag{13}$$

4.2. Unconstrained and constrained solutions

The two unconstrained solutions, direct regression and standard Bayes, can be performed in the same way as described in Section 3 regardless of the distribution of B.

To develop a constrained solution, we need first to define the likelihood function $L(B|\mathbf{X})$. It can be shown that $\Pr(B_i=1|X_i)$ is a weighted average of $\exp(\Sigma_{j=0}^{p} X_{ij}\phi_j)/\{1+\exp(\Sigma_{j=0}^{p} X_{ij}\phi_j)\}$ and $\exp(\Sigma_{j=0}^{p} X_{ij}\phi_j + \phi_{p+1})/\{1+\exp(\Sigma_{j=0}^{p} X_{ij}\phi_j + \phi_{p+1})\}$. We use estimates of the weights given by $w_{i,\bar{\boldsymbol{\beta}}} = 1/\{1+\exp(\mathbf{X}_i\bar{\boldsymbol{\beta}})\}$ and $1 - w_{i,\bar{\boldsymbol{\beta}}} = \exp(\mathbf{X}_i\bar{\boldsymbol{\beta}})/\{1+\exp(\mathbf{X}_i\bar{\boldsymbol{\beta}})\}$ where $\bar{\boldsymbol{\beta}}$ are the values for $\boldsymbol{\beta}$ from the established model. Then $L(B|\mathbf{X})$ can be written as

$$\begin{aligned} L(B|\mathbf{X}) &= \prod_{i=1}^{n} L(B_i|\mathbf{X}_i, \boldsymbol{\phi}) \\ &= \prod_{i=1}^{n} \left\{ \frac{\exp\left(\sum_{j=0}^{p} X_{ij}\phi_j\right)}{1+\exp\left(\sum_{j=0}^{p} X_{ij}\phi_j\right)} w_{i,\bar{\boldsymbol{\beta}}} + \frac{\exp\left(\sum_{j=0}^{p} X_{ij}\phi_j + \phi_{p+1}\right)}{1+\exp\left(\sum_{j=0}^{p} X_{ij}\phi_j + \phi_{p+1}\right)} (1 - w_{i,\bar{\boldsymbol{\beta}}}) \right\}^{B_i} \\ &\quad \times \left\{ \frac{1}{1+\exp\left(\sum_{j=0}^{p} X_{ij}\phi_j\right)} w_{i,\bar{\boldsymbol{\beta}}} + \frac{1}{1+\exp\left(\sum_{j=0}^{p} X_{ij}\phi_j + \phi_{p+1}\right)} (1 - w_{i,\bar{\boldsymbol{\beta}}}) \right\}^{1-B_i}. \end{aligned}$$

4.2.1. Constrained maximum likelihood

The constrained maximum likelihood estimation optimizes the following joint log-likelihood $L(Y|\mathbf{X}, B)\, L(B|\mathbf{X})$ with a set of constraints on $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$, namely

$$\min_{\gamma,\phi}\Bigg(\sum_{i=1}^{n}\Bigg[-Y_i\Bigg(\sum_{j=0}^{p}\gamma_j X_{ij}+\gamma_{p+1}B_i\Bigg)+\log\Bigg\{1+\exp\Bigg(\sum_{j=0}^{p}\gamma_j X_{ij}+\gamma_{p+1}B_i\Bigg)\Bigg\}\Bigg]$$

$$-\sum_{i=1}^{n}\Bigg[B_i\log\Bigg\{\frac{\exp\Big(\sum_{j=0}^{p}X_{ij}\phi_j\Big)w_{i,\bar{\boldsymbol{\beta}}}}{1+\exp\Big(\sum_{j=0}^{p}X_{ij}\phi_j\Big)}+\frac{\exp\Big(\sum_{j=0}^{p}X_{ij}\phi_j+\phi_{p+1}\Big)(1-w_{i,\bar{\boldsymbol{\beta}}})}{1+\exp\Big(\sum_{j=0}^{p}X_{ij}\phi_j+\phi_{p+1}\Big)}\Bigg\}$$

$$+(1-B_i)\log\Bigg\{\frac{w_{i,\bar{\boldsymbol{\beta}}}}{1+\exp\Big(\sum_{j=0}^{p}X_{ij}\phi_j\Big)}+\frac{1-w_{i,\bar{\boldsymbol{\beta}}}}{1+\exp\Big(\sum_{j=0}^{p}X_{ij}\phi_j+\phi_{p+1}\Big)}\Bigg\}\Bigg]\Bigg)$$

$$\text{subject to}\begin{cases}\gamma_0+\frac{1}{2}\phi_{p+1}+\frac{1}{4}\phi_0\phi_{p+1}+\frac{1}{8}\phi_{p+1}^2\in[\bar{\beta}_0-d\bar{S}_0,\bar{\beta}_0+d\bar{S}_0],\\ \gamma_j+\frac{1}{4}\phi_j\phi_{p+1}\in[\bar{\beta}_j-d\bar{S}_j,\bar{\beta}_j+d\bar{S}_j],\quad j=1,\ldots,p,\\ \gamma_{p+1}=\phi_{p+1}.\end{cases}\tag{14}$$

We also consider a modification that adds the Firth penalty term.

4.2.2. Informative full Bayes method

Analogously to the derivation of the informative full Bayes solution that was described in Section 3, we first write down the product of $L(Y|\mathbf{X}, B)$ and $L(B|\mathbf{X})$ with priors:

$$p(\boldsymbol{\gamma},\boldsymbol{\phi}|\text{data})\propto\Bigg[\prod_{i=1}^{n}\frac{\exp\Big\{\Big(\sum_{j=0}^{p}\gamma_j X_{ij}+\gamma_{p+1}B_i\Big)Y_i\Big\}}{1+\exp\Big(\sum_{j=0}^{p}\gamma_j X_{ij}+\gamma_{p+1}B_i\Big)}$$

$$\times\Bigg\{\frac{\exp\Big(\sum_{j=0}^{p}X_{ij}\phi_j\Big)w_{i,\bar{\boldsymbol{\beta}}}}{1+\exp\Big(\sum_{j=0}^{p}X_{ij}\phi_j\Big)}+\frac{\exp\Big(\sum_{j=0}^{p}X_{ij}\phi_j+\phi_{p+1}\Big)(1-w_{i,\bar{\boldsymbol{\beta}}})}{1+\exp\Big(\sum_{j=0}^{p}X_{ij}\phi_j+\phi_{p+1}\Big)}\Bigg\}^{B_i}$$

$$\times\Bigg\{\frac{w_{i,\bar{\boldsymbol{\beta}}}}{1+\exp\Big(\sum_{j=0}^{p}X_{ij}\phi_j\Big)}+\frac{1-w_{i,\bar{\boldsymbol{\beta}}}}{1+\exp\Big(\sum_{j=0}^{p}X_{ij}\phi_j+\phi_{p+1}\Big)}\Bigg\}^{1-B_i}\Bigg]\pi(\boldsymbol{\gamma},\boldsymbol{\phi}).\tag{15}$$

We can reparameterize expression (15) in terms of γ and β, and include a Jacobian corresponding to this transformation. We denote the Jacobian matrix by $\mathbf{M}$ where $|\mathbf{M}| = |4/\gamma_{p+1}|^{p+1}$. We specify independent weakly informative Cauchy priors for γ and use the constraints directly as priors for β. Then similarly to Section 3.4.2 we can rewrite the joint distribution in terms of γ and β as

$$
\begin{aligned}
p(\gamma, \beta | Y, \mathbf{X}, B) \propto & \left\{ \prod_{i=1}^{n} \frac{\exp\left\{ \left(\sum_{j=0}^{p} \gamma_j X_{ij} + \gamma_{p+1} B_i \right) Y_i \right\}}{1 + \exp\left(\sum_{j=0}^{p} \gamma_j X_{ij} + \gamma_{p+1} B_i \right)} \right. \\
& \times \left(\frac{w_{i,\bar{\beta}}}{1 + \exp\left[- \left\{ \frac{4\beta_0 - 4\gamma_0 - 2\gamma_{p+1} - \frac{1}{2}\gamma_{p+1}^2}{\gamma_{p+1}} + \sum_{j=1}^{p} X_{ij} \frac{4(\beta_j - \gamma_j)}{\gamma_{p+1}} \right\} \right]} \right. \\
& \left. + \frac{1 - w_{i,\bar{\beta}}}{1 + \exp\left[- \left\{ \frac{4\beta_0 - 4\gamma_0 - 2\gamma_{p+1} - \frac{1}{2}\gamma_{p+1}^2}{\gamma_{p+1}} + \sum_{j=1}^{p} X_{ij} \frac{4(\beta_j - \gamma_j)}{\gamma_{p+1}} + \gamma_{p+1} \right\} \right]} \right)^{B_i} \\
& \times \left[\frac{w_{i,\bar{\beta}}}{1 + \exp\left\{ \frac{4\beta_0 - 4\gamma_0 - 2\gamma_{p+1} - \frac{1}{2}\gamma_{p+1}^2}{\gamma_{p+1}} + \sum_{j=1}^{p} X_{ij} \frac{4(\beta_j - \gamma_j)}{\gamma_{p+1}} \right\}} \right. \\
& \left. \left. + \frac{1 - w_{i,\bar{\beta}}}{1 + \exp\left\{ \frac{4\beta_0 - 4\gamma_0 - 2\gamma_{p+1} - \frac{1}{2}\gamma_{p+1}^2}{\gamma_{p+1}} + \sum_{j=1}^{p} X_{ij} \frac{4(\beta_j - \gamma_j)}{\gamma_{p+1}} + \gamma_{p+1} \right\}} \right]^{1 - B_i} \right\} \pi(\beta)\, \pi(\gamma)\, |\mathbf{M}|.
\end{aligned}
$$

The full conditional distributions of $\gamma_0, \ldots, \gamma_{p+1}$ and $\beta_0, \ldots, \beta_p$ do not have closed form expressions and a Metropolis–Hastings algorithm is used.

4.2.3. *Transformation approach*

Suppose that the raw draws from the non-informative Bayes method for $Y|\mathbf{X}, B$ are γ and the raw draws from non-informative Bayes method for $B|\mathbf{X}, Y$ are ϕ. The posterior variances are $s^2_{\gamma_j}$, $j = 0, \ldots, p+1$, and $s^2_{\phi_k}$, $k = 0, \ldots, p+1$. Then γ^* and ϕ^* are obtained by solving the optimization problem

$$
\min_{\gamma^*, \phi^*} \left\{ d^2_{\mathrm{NED}}(\gamma, \gamma^*) + d^2_{\mathrm{NED}}(\phi, \phi^*) \right\} = \sum_{j=0}^{p+1} \frac{(\gamma_j - \gamma_j^*)^2}{s^2_{\gamma_j}} + \sum_{k=0}^{p+1} \frac{(\phi_k - \phi_k^*)^2}{s^2_{\phi_k}}
$$

$$
\text{subject to} \begin{cases} \gamma_0^* + \frac{1}{2}\phi_{p+1}^* + \frac{1}{4}\phi_0^*\phi_{p+1}^* + \frac{1}{8}\phi_{p+1}^{*2} \in [\bar{\beta}_0 - d\bar{S}_0, \bar{\beta}_0 + d\bar{S}_0], \\ \gamma_j^* + \frac{1}{4}\phi_j^*\phi_{p+1}^* \in [\bar{\beta}_j - d\bar{S}_j, \bar{\beta}_j + d\bar{S}_j], \quad j = 1, \ldots, p, \\ \gamma_{p+1}^* = \phi_{p+1}^*. \end{cases} \tag{16}
$$

5. Simulation study

To evaluate the performance of the various approaches we conduct a simulation study. The results for Gaussian B are presented here, and the results for binary B and other scenarios are presented in the on-line supplementary materials appendix E. The simulation scenario has three predicting variables, X_1, X_2 and B, and the sample size of each data set is 55. 500 replicate data sets are generated. Y_i is Bernoulli distributed with $\text{logit}\{\Pr(Y_i = 1|X_{i1}, X_{i2}, B_i)\} = 2 + 3X_{i1} + 3X_{i2} + 2B_i$. X_{i1} and X_{i2} are independently and identically distributed on $U(-0.75, 0.25)$ and B_i is simulated as $B_i = 0.5X_{i1} + 0.5X_{i2} + N(0, 0.75^2)$. A logistic regression based on a large data set of 10000 subjects gives estimates for the model $\text{logit}\{\Pr(Y = 1|\mathbf{X})\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. The estimates and standard errors from this fit are $\bar{\beta}_0 = 1.50, \bar{S}_0 = 0.04, \bar{\beta}_1 = 2.95, \bar{S}_1 = 0.09, \bar{\beta}_2 = 3.01$ and $\bar{S}_2 = 0.09$.

We report three evaluation metrics that are related to estimation accuracy: the average of the estimated coefficients, relative efficiency of the estimated coefficient and mean-squared error (MSE) across 500 replicates. The average of the estimated coefficients is defined as $\bar{\gamma}_j = (1/500)\Sigma_{m=1}^{500}\hat{\gamma}_{m,j}$; the relative efficiency of the estimated coefficients is defined as $V(\hat{\gamma}_{j,\text{direct}})/V(\hat{\gamma}_{j,\text{method}})$ where $V(\hat{\gamma}_{j,\text{direct}}) = (1/500)\Sigma_{m=1}^{500}(\hat{\gamma}_{m,j} - \bar{\gamma}_j)^2$ estimated by direct regression; the MSE of the estimated coefficients is defined as $(1/500)\Sigma_{m=1}^{500}(\hat{\gamma}_{m,j} - \gamma_j)^2$, $j = 1, \ldots, p+1$.

The predictive ability of logistic prediction models can be assessed by using a variety of methods and metrics on a validation data set (Steyerberg *et al.*, 2010). In this simulation study, we assess the predictive ability of the model on a validation data set of size 800 by using the scaled Brier score $(\Sigma_{i=1}^{n}(Y_i - \hat{p}_i)^2/\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2)$. We assess the variability of the predictions on the validation data set by using the standard deviation of the predicted probabilities, $\{(1/799)\Sigma(\hat{p}_i - \bar{\hat{p}})^2\}^{1/2}$. The discriminatory performance of the model is assessed by using the area under the receiver operating characteristic curve, AUC. These three performance measures are also calculated for the model based on $\bar{\beta}$ which does not use B, and for the best achievable model that uses the true values of the γs.

Table 1 presents the results. In this setting what is achievable with the true model is noticeably better, as measured by the Brier score and AUC, than using the established model, and the established model does not give the correct standard deviation of the predicted probabilities. There is bias in $\hat{\gamma}_1$, $\hat{\gamma}_2$ and $\hat{\gamma}_3$ for the direct regression approach, which is reduced by including the Firth penalty. We find that the constrained methods greatly improve the estimation efficiency of the coefficients γ_1 and γ_2 of $\mathbf{X}$. The constrained methods can reduce the MSE of X_1 and X_2 by 70% or more, and give four or five times more efficient point estimates of γ_1 and γ_2. In contrast, these constrained solutions can only reduce the MSE of γ_3 by 10–40% and generally have similar efficiency to that of direct regression plus the Firth correction. The new methods give similar variability of the predicted probabilities compared with the true model. All the methods that use the external information give similar Brier scores and AUCs which are very similar to the best that can be achieved. They are all better than not using B at all, and slightly better than the methods that do not use the external information. In terms of computational efficiency, the informative prior Bayesian approach and the transformation approach require more time than the other methods; however, the transformation approach takes about 18% the time of the informative full Bayes approach.

The results for the method of Chatterjee *et al.* (2016) show some similarities to those of the other methods that use the external information, but also show some differences. The method shows a similar amount of bias or even slightly greater bias than do the other methods. The method does result in some gain in efficiency in the point estimates, and also smaller MSE,

Table 1. Simulation results for Gaussian B†

Method	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	*Scaled Brier score*	*AUC*	$\hat{p}$ *mean*	*Time (s)*
True value	3	3	2	0.605	0.864	0.49 (0.333)	—
Established model using known $\bar{\beta}$	—	—	—	0.796	0.761	0.51 (0.239)	—
Direct regression	3.37 (1)	3.40 (1)	2.35 (1)	0.661	0.852	0.49 (0.344)	1.3
MSE	3.36	3.48	0.96				
Direct regression + Firth correction	2.89 (1.55)	2.92 (1.52)	1.99 (1.69)	0.651	0.852	0.49 (0.324)	2.4
MSE	2.10	2.18	0.49				
Non-informative Bayes	2.72 (1.79)	2.75 (1.78)	2.04 (1.77)	0.647	0.852	0.49 (0.307)	3.6
MSE	1.88	1.93	0.47				
Constrained maximum likelihood	3.08 (3.60)	3.17 (3.91)	2.30 (1.11)	0.628	0.857	0.49 (0.313)	44.9
MSE	0.90	0.88	0.84				
Constrained maximum likelihood + Firth correction	2.88 (6.01)	2.97 (6.32)	1.96 (1.85)	0.622	0.857	0.49 (0.303)	78.2
MSE	0.55	0.53	0.45				
Informative full Bayes	2.87 (4.93)	2.98 (5.15)	2.30 (1.33)	0.624	0.857	0.49 (0.301)	9097.6
MSE	0.64	0.67	0.72				
Transformation	2.90 (6.80)	3.00 (6.94)	1.93 (1.75)	0.622	0.857	0.49 (0.298)	888.2
MSE	0.48	0.48	0.48				
Chatterjee *et al.* (2016)	3.17 (2.91)	3.29 (3.06)	2.34 (1.01)	0.631	0.859	0.49 (0.342)	43.2
MSE	1.13	1.17	0.94				
Simple logistic ($\bar{p}$, B)	3.27 (1.64)	3.34 (1.62)	2.28 (1.11)	0.644	0.858	0.49 (0.339)	0.9
MSE	2.04	2.16	0.84				

†For each method, we report the mean (relative efficiency with respect to direct regression), MSE, average Brier score, average AUC, average $\hat{p}$ (with standard deviations in parentheses) and computing time for 500 data sets of size 55.

compared with direct regression, but not as much gain as for the other new methods. For $\hat{\gamma}_3$ there is no gain in efficiency compared with direct regression, and even a loss in efficiency compared with the direct regression with Firth correction. The method does tend to give slightly more variability to the predicted probabilities than the other methods. The method of Chatterjee *et al.* (2016) has comparable values of the AUC and Brier score with those of the other methods that use external information.

The simple logistic regression plug-in method is found to have better performance than direct regression, but not as good performance as the more sophisticated approaches for using the external information. It also can result in biased estimates of the γs.

The other results which are presented in the on-line supplementary materials give similar conclusions.

6. Application to the prostate cancer data

We demonstrate our methodology by enhancing the PCPT risk calculator for high grade prostate cancer. Using the data from Tomlins *et al.* (2015) we shall illustrate the methods that are described in this paper to develop a logistic model that includes all the PCPThg-variables and PCA 3. We estimate the new model from the training data set of 679 men, incorporating the known coefficients and their standard errors from the PCPThg-calculator. After a transfor-

mation ($\log_2(\text{PCA3}+1)$) the distribution of PCA 3 is roughly normally distributed in both cohorts and thus the approximate relationship equations (6) are applicable. The distribution of TMPRSS2:ERG looks like a truncated normal whose value is bounded below at 0, with many observations equal to 0. We dichotomized TMPRSS2:ERG by splitting at the median and develop a logistic model that includes PCPThg-variables and dichotomized TMPRSS2:ERG. The approximate relationship equations (13) would be appropriate in this case.

These two expanded PCPThg-models will be estimated by both the unconstrained methods and the constrained methods that were described in Section 3 and Section 4. For comparing coefficient estimation across different methods, we report the estimated coefficients and their standard errors calculated from the training data set. For comparing prediction power, we calculate the Brier score and the AUC based on the validation data set. We also present the original PCPThg-model and the expanded model developed by Tomlins *et al.* (2015). We give the calibration plots for the original PCPThg-model, the expanded model by Tomlins *et al.* (2015), the expanded PCPThg-model estimated without constraints (direct regression) and the expanded PCPThg-model estimated with constraints (transformation approach). The calibration plot contains the predicted and the observed risk of high grade cancer in 10 groups which are defined by sorting the predicted probabilities from lowest to highest and then separated into 10 groups of approximately equal size. For each group, the expected numbers of events is the sum of the predicted probability in the group. Perfect predictions should be on the 45° line.

Table 2 presents the expanded PCPThg-model incorporating these two biomarkers fitted to the training data set. For the expanded PCPThg-model incorporating the PCA 3 score, if we compare the standard errors across different methods, it is easily seen that the constrained methods can reduce the standard errors of regression coefficients compared with direct regression. For example, the informative full Bayes solution can substantially reduce the standard errors in parameters of variables PSA (0.08 *versus* 0.19), age (0.008 *versus* 0.013), DRE findings (0.17 *versus* 0.27), prior biopsy history (0.16 *versus* 0.28) and race (0.23 *versus* 0.31). The constrained maximum likelihood with Firth penalty can reduce the standard errors of the parameters of variables PSA, age, prior biopsy history and race by at least 50%.

Among the 1218 validation study patients, the AUC for the PCPThg-model and the expanded PCPThg-score plus PCA 3 model are 0.707 and 0.752. By incorporating the PCA 3 score in the PCPThg-model, the AUC increases to 0.767 in direct regression. However, the constrained methods do not further increase the AUC. All the new methods except for the transformation approach give predicted probabilities that are on average too high, suggesting that they are not well calibrated. For calibration, as measured by the Brier score, the original PCPT calculator performs better than direct regression and of the new methods only the transformation approach gives an improved Brier score. In Fig. 1 we can see that the expanded PCPThg-model incorporating PCA 3 tends to overestimate the risk of developing high grade prostate cancer among those patients with high risk. However, the overall calibration ability of the expanded PCPThg-model estimated by the transformation approach still outperforms that of the original PCPThg-model, the expanded PCPThg-score plus PCA 3 model or the expanded PCPThg-model estimated by direct regression.

The expanded PCPThg-model incorporating binary TMPRSS2:ERG fitted to the training data set again shows that the constrained methods can reduce the standard errors of regression coefficients compared with direct regression. In Fig. 1 we can see that the expanded PCPThg-model incorporating binary TMPRSS2:ERG tends to overestimate the risk of developing high grade cancer among those patients with high risk. However, the transformation approach predicts the risk well for the high risk groups.

Table 2. Expanded PCPThg-model†

Model	*PSA*	*Age*	*DRE findings*	*Prior biopsy history*	*Race*	*PCA 3 or TMPRSS2: ERG score*	*Scaled Brier score*	*AUC*	*$\hat{p}$ mean*
Original PCPThg	1.29 (0.09)	0.031 (0.012)	1.00 (0.17)	−0.36 (0.18)	0.96 (0.27)	—	0.933	0.707	0.14 (0.132)
Estimated PCPThg	1.06 (0.18)	0.033 (0.012)	1.15 (0.26)	−1.44 (0.27)	0.44 (0.29)	—	0.975	0.716	0.27 (0.174)
Expanded model with PCA 3 score									
PCPThg-score + PCA 3	—	—	—	—	—	—	0.950	0.752	0.27 (0.201)
Direct regression	1.00 (0.19)	0.009 (0.013)	1.07 (0.27)	−1.30 (0.28)	0.04 (0.31)	0.56 (0.08)	0.950	0.767	0.28 (0.221)
Direct regression + Firth correction	0.97 (0.19)	0.009 (0.013)	1.06 (0.27)	−1.27 (0.27)	0.05 (0.31)	0.56 (0.08)	0.953	0.767	0.28 (0.219)
Non-informative Bayes	0.98 (0.18)	0.009 (0.013)	1.05 (0.27)	−1.27 (0.27)	0.04 (0.30)	0.56 (0.08)	0.950	0.767	0.28 (0.218)
Constrained maximum likelihood	1.20 (0.09)	0.010 (0.007)	1.08 (0.14)	−0.55 (0.13)	0.30 (0.19)	0.59 (0.08)	0.948	0.766	0.27 (0.225)
Constrained maximum likelihood + Firth correction	1.19 (0.09)	0.012 (0.006)	1.08 (0.14)	−0.54 (0.13)	0.47 (0.11)	0.53 (0.07)	0.947	0.764	0.27 (0.218)
Informative full Bayes	1.23 (0.10)	0.009 (0.008)	0.99 (0.17)	−0.73 (0.17)	0.26 (0.22)	0.60 (0.08)	0.946	0.767	0.27 (0.222)
Transformation	1.23 (0.07)	0.008 (0.009)	0.96 (0.14)	−0.50 (0.13)	0.41 (0.19)	0.55 (0.08)	0.883	0.765	0.22 (0.191)
Chatterjee *et al.* (2016)	1.22 (0.08)	0.007 (0.005)	0.86 (0.10)	−0.20 (0.08)	0.58 (0.11)	0.56 (0.10)	0.888	0.759	0.15 (0.168)
Simple logistic ($\bar{p}$, B)	0.82 (0.18)	0.023 (0.000)	0.64 (0.11)	−0.23 (0.01)	0.61 (0.10)	0.55 (0.08)	0.940	0.759	0.27 (0.204)
Expanded model with binary TMPRSS2:ERG									
PCPThg-score + TMPRSS2:ERG	—	—	—	—	—	—	0.932	0.732	0.26 (0.153)
Direct regression	1.01 (0.18)	0.032 (0.012)	1.03 (0.26)	−1.44 (0.28)	0.57 (0.29)	0.77 (0.20)	0.929	0.745	0.26 (0.179)
Direct regression + Firth correction	0.98 (0.18)	0.032 (0.012)	1.02 (0.26)	−1.41 (0.27)	0.57 (0.29)	0.76 (0.20)	0.930	0.744	0.27 (0.177)
Non-informative Bayes	0.99 (0.18)	0.032 (0.012)	1.01 (0.26)	−1.40 (0.27)	0.55 (0.29)	0.76 (0.20)	0.926	0.745	0.27 (0.175)
Constrained maximum likelihood	1.14 (0.07)	0.032 (0.004)	1.06 (0.14)	−0.52 (0.11)	0.81 (0.18)	0.74 (0.21)	0.928	0.742	0.25 (0.176)
Constrained maximum likelihood + Firth correction	1.14 (0.07)	0.032 (0.004)	1.06 (0.14)	−0.52 (0.11)	0.80 (0.17)	0.72 (0.20)	0.931	0.742	0.26 (0.176)
Informative full Bayes	1.14 (0.09)	0.033 (0.007)	0.95 (0.14)	−0.76 (0.16)	0.77 (0.21)	0.73 (0.19)	0.922	0.744	0.25 (0.175)
Transformation	1.17 (0.07)	0.030 (0.007)	0.94 (0.12)	−0.50 (0.11)	0.89 (0.16)	0.74 (0.14)	0.889	0.742	0.21 (0.152)
Chatterjee *et al.* (2016)	1.25 (0.03)	0.029 (0.002)	0.85 (0.05)	−0.37 (0.04)	1.06 (0.05)	0.77 (0.27)	0.911	0.736	0.14 (0.129)
Simple logistic ($\bar{p}$, B)	0.98 (0.18)	0.023 (0.000)	0.76 (0.11)	−0.28 (0.01)	0.73 (0.10)	0.80 (0.19)	0.918	0.739	0.25 (0.155)

†For each method, point estimates (with standard errors in parentheses) from the training data set, and the Brier score, the AUC and the mean and standard deviation (in parentheses) of predicted probabilities from the validation data set. The sample size of the training data set is 679. The sample size of the validation data set is 1218.

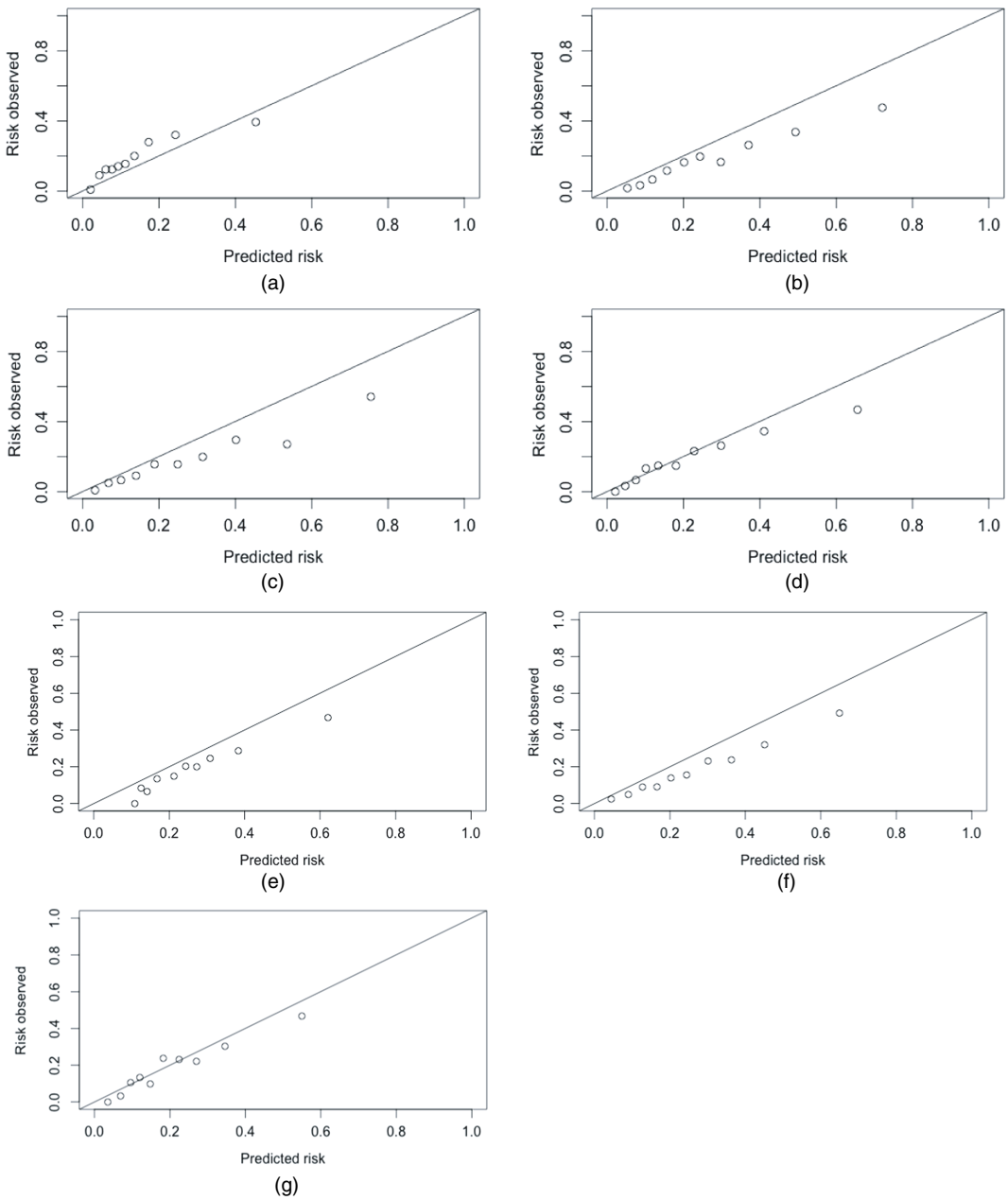

Fig. 1. Calibration plot of the original high grade PCPT risk calculator PCPThg and calibration plots of the expanded PCPThg-model by incorporating PCA 3 score and dichotomized TMPRSS2:ERG: (a) PCPThg-model; (b) PCPThg-score + PCA 3, Tomlins *et al.* (2015); (c) PCPThg-covariates + PCA 3, direct regression; (d) PCPThg-covariates + PCA 3, Bayesian transformation approach; (e) PCPThg-score + dichotomized TMPRSS2:ERG; (f) PCPThg-covariates + dichotomized TMPRSS2:ERG, direct regression; (g) PCPThg-covariates + dichotomized TMPRSS2:ERG, Bayesian transformation approach

For the prostate cancer example with either new biomarker, the method of Chatterjee *et al.* (2016) gave parameter estimates and standard errors that are different from those of both of the methods that do not use the external information and from the other methods that do use the external information. In general, they tend to be closer to those of the original PCPT calculator. Also the method gives a much lower predicted population proportion than all the other methods and lower than the observed prevalence of 18.3% in the validation data set. This is possibly because the method of Chatterjee *et al.* (2016) does not take into account the uncertainty in the estimates coefficients. It is notable, in contrast with what was seen in the simulation studies, that the method of Chatterjee *et al.* (2016) tends to give smaller standard errors than do all the other methods for the coefficients of the **X**-variables, but larger standard errors for the coefficient of the new biomarker. The method does assume that the **X**-distribution is the same in the original data set and the data set with the new biomarkers, but in fact there are considerable differences between these **X**-distributions; it is unclear how this is affecting the estimates and standard errors. The results (as shown in appendix F of the on-line supplementary materials) for the constrained maximum likelihood estimate, with a range of values for d, demonstrate that the choice of d can be quite impactful, both on the estimates and on their standard errors. As expected the results for $d = 0.1$ are quite close to those of the method of Chatterjee *et al.* (2016), because neither method incorporates the uncertainty in the parameter estimates.

7. Discussion

We propose several strategies for translating the external coefficient information that is obtained from outside the data set into constraints on regression coefficients in the setting of a logistic regression model describing $\Pr(Y = 1|\mathbf{X}, B)$. Simulation studies show that the external coefficient information from the established model can help to improve the efficiency of estimation and to enhance the predictive power in the expanded model.

In terms of computational efficiency, in simulation studies the transformation approach shows advantage over the informative full Bayes method because in the transformation approach the raw draws are first obtained in a fast way and then transformed into draws that obey the constraints based on an efficient optimization algorithm, whereas the informative full Bayes solution produces constrained draws inefficiently. When the dimensionality of the predictors **X** increases, the computational cost of the transformation approach solution will not increase much because the high dimensional optimization problem will always reduce to a one-dimensional optimization problem based on our algorithm regardless of the dimensionality of the predictor space. Furthermore, the correlation between the samples in the Markov chain for the informative full Bayes approach is very high and effective samples are more difficult to obtain when the dimensionality increases (additional simulation results that validate this finding are not shown). As a consequence, the discrepancy of these two constrained solutions in computational cost will be more apparent in higher dimensions. In general the Bayesian approaches are much more computationally time consuming than the other approaches. Although it is conceivable that better algorithms and improved programming could speed these up considerably, it is very unlikely that they will ever have speed comparable with that of the constrained maximum likelihood methods or the approach of Chatterjee *et al.* (2016). A general overview of the computational and implementation details for all the methods that are described in this paper are given in on-line supplementary materials appendix G.

The gain in efficiency in the expanded model of interest depends on the sample size that is

used to construct the established model and the sample size that is used to estimate the expanded model of interest. In our simulation studies the established models are based on large data sets with 10000 observations whereas the current data sets are very small. The relative gain in efficiency in the regression coefficient of variables $\mathbf{X}$ by incorporating the external coefficient information is significant and the predictive power in the validation data set is enhanced. However, when the sample size in the current data set is sufficiently large to estimate the expanded model, the constrained methods do not lead to much improvement in the predictive ability compared with direct regression, as was the case in the prostate cancer example. However, our numerical results suggest that improved precision of the coefficient estimates, as measured by standard errors, can be achieved even if the current data set is not small.

A situation that we did not consider in the simulation studies is when the event of interest is rare and the predicted probabilities are very low. Although there may be other methods that exploit this assumption, we hypothesize that the relative performance of the methods that we considered may be similar to those for the small samples sizes that we did evaluate. This is because both situations have limited information in the data from which to estimate regression coefficients. However, this hypothesis would need to be investigated.

The approaches that are proposed in this paper are based on establishing a relationship between the parameters in the $Y|\mathbf{X}$ model and the parameters in the $Y|\mathbf{X}, B$ model. Depending on the form of B and the structure of the models these relationships need to be analytically derived and are approximations. The methods also require an explicit model for $B|\mathbf{X}$. Although it would be desirable to avoid having to specify a parametric model for $B|\mathbf{X}$, we also note that the appropriateness of the model for $B|\mathbf{X}$ can be checked to some degree from the small data set. The differences in the distributions of $\mathbf{X}$ in the external and internal studies will not have much effect on the performance of our proposed constrained methods (the simulation results are not shown). This is because the coefficients' approximate relationship equations are constructed on the basis of the conditional distributions $Y|\mathbf{X}, B$ and $B|\mathbf{X}$. As long as these two conditional distributions are correctly specified in the internal study, the approximate relationship equations will hold regardless of the differences in the distributions of $\mathbf{X}$ in these two studies. The approaches also do have the feature that they can directly incorporate the uncertainty in the parameters of the $Y|\mathbf{X}$ model. An alternative method of using $\bar{p}$ as a covariate is appealing because of its simplicity and broad applicability, although it does appear to have slightly worse properties than the more sophisticated approaches. The approach of Chatterjee *et al.* (2016) is appealing because it is applicable to any form of B and it does not require an explicit model for $B|\mathbf{X}$; however, it does require the same distribution of $\mathbf{X}$ in the two populations and it does not incorporate the uncertainty in the parameters of the $Y|\mathbf{X}, B$ model. We also found that it was sometimes numerically unstable for small sample sizes.

One point of future consideration is the distribution of the new biomarker B. We develop the approximate relationship equation for the scenarios that B is Gaussian and binary. When B is multivariate Gaussian, based on the generalization of equation (5), assuming that $B|\mathbf{X}$ is multivariate normal with L dimensions, mean $\mathbf{X}\boldsymbol{\theta}$ and covariance matrix $\mathbf{V}_{L\times L}$, the approximate relationship between $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ is

$$\beta_j \approx \left(\gamma_j + \sum_{l=1}^{L} \gamma_{p+l}\theta_{lj}\right) \Big/ (1+\boldsymbol{\gamma}_B^{\mathrm{T}}\mathbf{V}\boldsymbol{\gamma}_B/1.7^2)^{1/2}, \qquad j=0,\ldots,p. \tag{17}$$

Then the strategies to incorporate the external coefficient information that were described in Section 3 can be easily extended in this case. However, if additional biomarkers follow other distributions, these approximate relationship equations will fail. Therefore, further

investigations are needed for the generalization of our proposed constrained solutions to adapt flexibly to other possible distributions of the new biomarker.

Acknowledgements

This research was supported by National Science Foundation grant DMS 1712933 and National Institutes of Health grant CA 129102.

References

Chatterjee, N., Chen, Y.-H., Maas, P. and Carroll, R. J. (2016) Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *J. Am. Statist. Ass.*, **111**, 107–117.

Cheng, W., Taylor, J. M. G., Vokonas, P. S., Park, S. K. and Mukherjee, B. (2018) Improving estimation and prediction in linear regression incorporating external information from an established reduced model. *Statist. Med.*, **37**, 1515–1530.

D'Agostino, R. B., Grundy, S., Sullivan, L. M., Wilson, P. and for the CHD Risk Prediction Group (2001) Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *J. Am. Med. Ass.*, **286**, 180–187.

Firth, D. (1993) Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.

Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C. and Mulvihill, J. J. (1989) Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natn. Cancer Inst.*, **81**, 1879–1886.

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y.-S. (2008) A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Statist.*, **2**, 1360–1383.

Grill, S., Ankerst, D. P., Gail, M. H., Chatterjee, N. and Pfeiffer, R. M. (2017) Comparison of approaches for incorporating new information into existing risk prediction models. *Statist. Med.*, **36**, 1134–1156.

Grill, S., Fallah, M., Leach, R. J., Thompson, I. M., Hemminki, K. and Ankerst, D. P. (2015) A simple-to-use method incorporating genomic markers into prostate cancer risk prediction tools facilitated future validation. *J. Clin. Epidem.*, **68**, 563–573.

Gunn, L. H. and Dunson, D. B. (2005) A transformation approach for incorporating monotone or unimodal constraints. *Biostatistics*, **6**, 434–449.

Heinze, G., Ploner, M., Dunkler, D. and Southworth, H. (2013) Firth bias reduced logistic regression. *R Package Version 1.21*.

Heinze, G. and Schemper, M. (2002) A solution to the problem of separation in logistic regression. *Statist. Med.*, **21**, 2409–2419.

Imbens, G. W. and Lancaster, T. (1994) Combining micro and macro data in microeconometric models. *Rev. Econ. Stud.*, **61**, 655–680.

Mealiffe, M. E., Stokowski, R. P., Rhees, B. K., Prentice, R. L., Pettinger, M. and Hinds, D. A. (2010) Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. *J. Natn. Cancer Inst.*, **102**, 1618–1627.

Monahan, J. and Stefanski, L. A. (1992) Normal scale mixture approximations to $F^*(z)$ and computation of the logistic-normal integral. In *Handbook of the Logistic Distribution* (ed. N. Balakrishnan). New York: CRC Press.

Newcombe, P. J., Reck, B. H., Sun, J., Platek, G. T., Verzilli, C., Kader, A. K., Kim, S.-T., Hsu, F.-C., Zhang, Z., Zheng, S. L., Mooser, V. E., Condreay, L. D., Spraggs, C. F., Whittaker, J. C., Rittmaster, R. S. and Xu, J. (2012) A comparison of Bayesian and frequentist approaches to incorporating external information for the prediction of prostate cancer risk. *Genet. Epidem.*, **36**, 71–83.

Qin, J. (2000) Combining parametric and empirical likelihoods. *Biometrika*, **87**, 484–490.

Qin, J., Zhang, H., Li, P., Albanes, D. and Yu, K. (2015) Using covariate-specific disease prevalence information to increase the power of case-control studies. *Biometrika*, **102**, 169–180.

Satten, G. A. and Kupper, L. L. (1993) Inferences about exposure-disease associations using probability-of-exposure information. *J. Am. Statist. Ass.*, **88**, 200–208.

Steyerberg, E. W., Eijkemans, M. J. C., Van Houwelingen, J. C., Lee, K. L. and Habbema, J. D. F. (2000) Prognostic models based on literature and individual patient data in logistic regression analysis. *Statist. Med.*, **19**, 141–160.

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

Thompson, I. M., Ankerst, D. P., Chi, C., Goodman, P. J., Tangen, C. M., Lucia, M. S., Feng, Z., Parnes, H. L. and Coltman, C. A. (2006) Assessing prostate cancer risk: results from the prostate cancer prevention trial. *J. Natn. Cancer Inst.*, **98**, 529–534.

Tomlins, S. A., Day, J. R., Lonigro, R. J., Hovelson, D. H., Siddiqui, J., Kunju, L. P., Dunn, R. L., Meyer, S., Hodge, P., Groskopf, J., Wei, J. T. and Chinnaiyan, A. M. (2015) Urine TMPRSS2:ERG plus PCA3 for individualized prostate cancer risk assessment. *Eur. Urol.*, **70**, 45–53.

Truong, M., Yang, B. and Jarrard, D. F. (2013) Toward the detection of prostate cancer in urine: a critical analysis. *J. Urol.*, **189**, 422–429.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Informing a risk prediction model for binary outcomes with external coefficient information: supplementary materials'.