Biomarker selection and cutoff estimation in drug development

Dem Fachbereich 03: Mathematik/Informatik
University of Bremen
for obtaining the academic degree
doctor rerum naturalium
(Dr. rer. nat.)

submitted Dissertation

by
Eleni Vradi, M.Sc.
born on 17.02.1987 in Korfu, Greece

Supervisor: Dr. Richardus Vonk
Prof. Dr. Werner Brannath

Submitted: 07.02.2019
Defended: 16.05.2019

Universität Bremen

Kompetenzzentrum für Klinische Studien Bremen

Acknowledgment

First of all, I would like to thank my supervisors, for being involved in this fruitful work. Their advice, positive mind and inspiration, made this thesis a great pleasure to work on. I would like to express my gratitute to my advisor Richardus Vonk for his support and all the trust he placed in me for this work. Thanks also to Research and Clinical Science Statistics department at Bayer AG for being a warm host during the last three years.

I am grateful to my academic advisor Werner Brannath, University of Bremen, for his guidance, encouragement and his influence on my research orientation. I am also thankful to Thomas Jaki for supporting me during my secondment at Lancaster University and for his valuable advice on practical considerations of biomarker selection methods as well as his great mentoring. My appreciation also goes to ESRs Pavel, Haiyan, Enya, Saswati and Arsenio for the nice moments during my secondments at Lancaster University and University of Bremen.

Special thanks go to my family and friends for their love and support. I am glad that I was part of the IDEAS European training network (http://www.ideas-itn.eu/) and had the opportunity to collaborate with great researchers.

Contents

1 Introduction 1

2 Model selection based on Combined penalties for Biomarker identification 11

3 A Bayesian approach to estimate the cutoff and the clinical utility of a biomarker assay 19

4 A Bayesian method for variable selection and classification under constrained clinical utility 31

Bibliography 43

Contributed Manuscripts 51

CHAPTER 1

Introduction

Drug development is a lengthy process that takes on average 12 years from discovery to the market for an innovative drug (Kola and Landis, 2004). This results in high costs for pharmaceutical companies (DiMasi et al. (2003)). The drug development process involves different phases: target discovery, animal studies, clinical development, and regulatory approval. Target discovery is the first step in the discovery of a medicine and refers to identifying the biological origin of a disease, and the potential targets for intervention. After the candidate drug is tested on animals to evaluate its safety, we move to the clinical development where the drug is tested on humans. In order to move from preclinical stages, which comprise *in-vivo* and *in vitro* experiments, to the clinical phase, consisting of studies in humans, regulations require specific pre-clinical safety and efficacy assessments. Regulatory agencies, such as the Food and Drug Administration (FDA) and the European Medicines Agency (EMA), are responsible for new drug approvals in the USA and European Union, respectively. These agencies examine the results of the safety and efficacy studies conducted during drug development.

Regardless of the high development costs, the rate of new drugs finally entering the

market remain low. Indeed, using data from ten large US and European pharma companies, Kola and Landis (2004) report that only 11% of all the submitted compounds had been approved by the European and/or the US regulatory authorities between 1991 and 2000. Most worrying is the failure rate for compounds already in Phase III trials, which is about 45%. In a recent follow-up study by DiMasi et al. (2016), the authors estimated around 12% for the overall probability of success (i.e. the likelihood that a drug in clinical phase is approved). The associated risks highly increase the costs per approved drug.

According to Kola and Landis (2004), the success rate varies considerably between different therapeutic areas: cardiovascular drugs, for instance, have 20% success rate, whereas only 8% of the compounds for central nervous system disorders successfully pass from first-in-man to registration stage. In other therapeutic areas, such as women's health, the failure rate at the registration stage is as high as 42%, and in oncology it is about 30%. Approximately 62% of the compounds undergoes attrition in Phase II trials. In this phase, cancer treatments face the highest failure rate, with more than 70% of the oncology compounds failing at this stage.

Major causes of attrition are the lack of efficacy accounting for approximately 30% and toxicology and clinical safety accounting for a further approximately 30% of failures. Denayer et al. (2014) and McGonigle and Ruggeri (2014) discuss extensively about the translational value of animal experiments. The authors emphasize that failures during phases II and III can be decreased by setting more stringent success criteria for the non-clinical stages and by generating more predictive animal models.

Drug development is even more complicated by the fact that, despite the efforts made to develop effective and safe medications, drugs often do not have the same outcomes in all patients. Some patients may respond well to a given treatment, others may not get any benefit, whilst for others the treatment can even be harmful. Individualized medicine offers a chance to pinpoint patients that respond differently to treatment because of specific biological and genetic features. In addition, individualized medicine allows to identify different patient populations since early stages of

drug development, and run trials that specifically target the responsive subgroup, potentially leading to significant results.

Biomarkers and precision medicine in drug development

In the era of personalized medicine, a growing variety of single and, more often, panels of biomarkers allow us to better understand health, risk factors, and disease mechanisms. According to the National Institutes of Health Biomarkers definition Group, a biomarker *"is a characteristic that is objectively measured and evaluated as an indicator of healthy biological processes, pathological processes, or pharmacological responses to therapeutic intervention"* (Biomarkers Definitions Working Group, 2001). Also the EMA has emphasized the importance of biomarkers (EMA, 2009), (EMA, 2009) stating that they play an increasingly important role in the development of new medicines and that their use is expected to streamline the access to new medicines.

In fact, today biomarkers are an integral part of drug development, as the design of new trials relies on biomarkers that highlight differences in patients' genetic profiles. These differences are then used to match a drug with those subjects that are more likely to benefit from it. Currently, biomarkers are notably employed in oncology research to select the most appropriate therapy according to the genomic characterization of individual tumours, and their use will likely increase in the near future. An example is breast cancer, where the treatment is often decided depending on a range of genetic traits, such as the status of the estrogen receptor gene, the amplification of the epidermal growth factor receptor 2 (HER2) gene and gene-expression profiles indicating the prognostic aggressiveness of the disease.

Buyse et al. (2011) provides definitions of different types of biomarkers used in drug development and cancer research. A prognostic biomarker gives information on the likely course of a disease in an untreated individual. That is, prognostic biomarkers foresee the prognosis of individual patients. Predictive biomarkers are used to predict how a patient will respond to a specific treatment. The baseline value of a

predictive biomarker must be shown to predict the efficacy or toxicity of a treatment, as assessed by a defined clinical end point. In other words, prognostic biomarkers predict the outcome in a natural cohort and predictive biomarkers predict the effect of an experimental treatment compared to a control group.

To validate a predictive biomarker, its ability to predict the effects of a drug (or lack thereof) should be demonstrated in multiple studies. Data from randomized trials that include patients with both high and low levels of biomarker are required to identify statistically significant predictive markers. Although retrospective analyses are sometimes sufficient to identify candidate predictive biomarkers and to incorporate them into trial design and clinical practice, prospective clinical trials may still be required to provide definitive evidence (Buyse et al., 2011).

According to Zhang et al. (2018), biomarker based clinical trial designs are appealing because they are more likely to succeed, by more accurately targeting the appropriate population as defined by the biomarker. In addition, they can lead to a shorter development process. From an industry point of view, Lavezzari and Womack (2016) argues that qualified biomarkers allow to enroll only specific patient subpopulations, resulting in faster, cheaper, and more informative clinical trials, thus increasing the speed to market of effective treatment options. In enrichment trials, defined as trials for which only patients tested as marker-positive are enrolled, biomarkers help to select a study population that will more likely to respond to a new drug. Therefore, enrichment designs may reduce drug development time and costs. However, several challenges, such as the predictability of biomarkers and the development of robust biomarker assays, must still be overcomed before advanced biomarker-based designs can be implemented in the clinic.

The importance of the use of biomarkers in reducing attrition rates and enhance translatability, e.g. to infer the correct dosage or signal whether a molecular target has been hit, was emphasized by Wehling (2009) who developed a score to assess the predictive value of biomarkers *per se*. Attention was drawn in the quantification of the importance of biomarkers and personalized medicine in the assessment

of translatability. The biomarker score (detailed in Wehling (2006)) is part of an overall score assigned to the translatability assessment. This scoring system gives an estimate of the translatability of an early drug project by including several pieces of evidence from animal and human studies, biomarker validation, and pharmacogenomics, among others. In Wehling (2011), the author presents two case studies and further present eight case studies in Wendler and Wehling (2012), where the scoring system has been successfully implemented. However, the translatability score is based on subjective beliefs and awaits validation. The proposed scoring system produces results retrospectively, but the important element here is that it shows that the early development and use of powerful biomarkers can considerably decrease the risk in drug development.

Bayesian clinical trials

Since last decade, Bayesian methodology has been regarded as a useful statistical method in clinical research because it can be easily adapted to accrue information during trials. Bayesian approaches can ease the decision making process and allow the design of smaller more informative trials. In addition, they can also be used to make mid-course adjustments to the trial design, to stop a trial, or to shorten a study (Berry, 2006). Bayesian analysis also allows to integrate historical information and synthesize results of several trials.

The FDA issued the "Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials" to guide the design of clinical trials for medical devices (FDA, 2010). Neither European nor Japanese regulatory authorities have introduced such documents for Bayesian statistical methods in clinical trials. Although the FDA's guidance indicates some additional conditions to use Bayesian statistics for medical devices i) prior information should be discussed with the FDA prior to the initiation of a study, and ii) indications of the device may be impacted by modifications at the interim analysis, the Bayesian framework offers clear advantages over its frequentist counterpart. For instance, traditional statistical methods take into account informa-

tion from previous studies only at the design stage. Bayesian statistics instead can incorporate information gathered before, during, and outside the trial. Furthermore, Bayesian methods allow to monitor the trial more frequently and to make interim decisions as soon as sufficient evidence is obtained.

The use of Bayesian methodology has been encouraged and has increased in the pharmaceutical industry, especially in early clinical settings. However, the methodology seems to have far less impact on preclinical *in vivo* studies. This is rather surprising in the context of regularly repeated in vivo studies where there is a considerable amount of data from historical control groups which has potential value. Besides preclinical and clinical research, the Bayesian approach can also be applied during post-marketing surveillance and for meta-analysis. Moreover, recent advancements in both computational algorithms and computer hardware have considerably increased the feasibility of Bayesian computational methods.

Contribution and Purpose of the thesis

In this dissertation, we present different statistical topics related to the usability and applicability of biomarkers in clinical drug development. The development of methods aiming to be used in translational research, has been a strong motivation for this thesis. Even though the chapters cover a wide range of statistical methodologies, it is important to emphasize the features in common to the topics discussed. First, all results can be appliedin different phases of drug development, both in clinical and pre-clinical areas. That means that we focused on methods that can be used to incorporate preclinical reslults into the selection of biomarkers in clinical development. Second, this thesis provides a framework to quantify the uncertainty in variable selection and consequently in decision-making (e.g. decisions on selection of a cutoff for distinguishing biomarker negative and positve patients)

The main objectives of this cumulative thesis are summarized below:

1. We propose a method to select biomarkers, discuss the trade-off between parsi-

mony and classification performance of the resulting model and explain under which conditions the method can be applied in clinical practice.

2. Based the observed score for each individual, that is derived from the selected biomarkers, we aim to develop an efficient estimation method to select cut-points that take into account the clinical utility of a diagnostic test. The methodology can be generalized and also applied to patient screening or patient selection in enrichment studies.

3. We propose a Bayesian variable selection method that simultaneously perform variable selection and cutoff estimation (of the selected variables) by controlling the clinical utility, in a way that the predictors in the final model are selected under constrained predictive values.

4. Apply this methodology to real data sets of clinical practice and finally to suggest new interesting lines of future research.

The thesis is structured as follows. In the first part, we introduce a new method for biomarker identification in clinical environment. We present a method to select biomarkers that aims to reduce the complexity of the model, i.e. the dimensionality of the model, while keeping the classification accuracy of the model as high as possible. Our method encourages sparsity through the combination of the L_0 with L_1-norm regularization of the model parameters. Although the combination of the L_0 with L_1-norm regularization was firstly introduced by Liu and Wu, the use of their method was restricted to moderate-sized data sets ($p << n$) (Liu and Wu (2007)). Instead, by considering a stepwise method for variable selection, we are able to apply the combined penalty function to high-dimensional settings. We also consider the combination of the L_0 with L_2-norm to account for the grouping effect. This means that, if in the dataset there is a group of correlated predictors, the L_2 norm tends to include or exclude simultaneously the group of the correlated variables. In contrast, the L_1 norm will select only a "representative" variable from a group of correlated predictors.

In early clinical development with limited sample size, statistical methods for incorporating results from animal models in early clinical trials are needed. It is of high importance that the development of biomarker assays should be initiated early in drug development process to be able to bridge later phases of clinical biomarker assessment. If prior knowledge from preclinical research is available, it should be utilized and incorporated in the selection process of biomarkers. A Bayesian framework that facilitates incorporation of prior information is considered in Chapter 3 and Chapter 4.

In Chapter 3, we discuss the cutoff selection of a biomarker assay. Continuous diagnostic tests (biomarkers or risk scores) are often used to discriminate between healthy and diseased populations. For the clinical application of such tests, the key aspect is how to select an appropriate cut-off, or threshold value, which defines a population that is more likely to respond to a treatment and a group of non-responders. For example, in enrichment studies, estimating a cutoff value such that the subsequent patient enrollment in a trial will depend on that value, is a very critical step.

As mentioned by Simon (2010) and adopted in this thesis, the term *clinical utility* refers to the applicability of a biomarker test to improve the outcome for patients. Improved outcome means that patients live longer, or that the treatment yields to the same effect of another drug but shows fewer adverse events. Clinical validation of a test is often accompanied by calculating the sensitivity of the test for the identification of responders and the specificity for identifying nonresponders, respectively. Sensitivity and specificity are useful measures to quantify the diagnostic ability of a test, however, they do not provide relevant information when making a clinical decision. By contrast, predictive values, the Positive Predictive Value (PPV) and the Negative Predictive Value (NPV), which are functions of sensitivity, specificity and the prevalence of the disease, are more useful measures to make clinical decisions and quantify the clinical utility of the biomarker-based test. Overall, the clinical utility of a test is a more informative and useful measure (than clinical validation) to make therapeutic decisions.

In the second part of the thesis, we introduce a new approach to derive the distribution of the cut-off and predictive values of a continuous biomarker. When applying novel biomarkers into routine standard care, it is important to consider risk thresholds to ensure the best possible decisions for every patient. Especially in early oncology trials, integration of biomarker information to better demonstrate efficacy of a treatment and guide the design of following studies, is crucial. A good example of how important is the choice of the cutoff value and how biomarkers can be applied to the clinic is described in Lunceford (2015), where he discusses cutoff selection as a criterion for patient enrolment in enrichment trials.

In our approach, we apply a Bayesian method to estimate the cutoff value of a biomarker assay by using the predictive values, and also determine the uncertainty around these estimates. We use a step function, which serves as an approximate model facilitating classification into two groups that have different response rates. The advantage of using the step function is that both the cutoff and the predictive values are parameters of the model. Even if the assumption of a step function is strong and the model is misspecified, the estimates of the assumed step function are consistent for the parameter values for which the assumed model minimizes the distance from the true distribution in terms of Kullback-Leibler divergence.

The proposed method works also well if we apply a constraint on the positive predictive value of the test, namely to belong to a predetermined interval of high values, e.g. between 80 and 100%. Equivalently the constraint can be applied to the NPV of the test. We illustrate this approach by considering the previously described Bayesian method and restricting the domain of the prior distribution to the desired constrained interval. This optimization strategy provides the best classifier with the PPV in the pre-determined interval.

In the fourth chapter, we discuss the simultaneous variable selection and cutoff estimation (of the selected variables) by controlling the clinical utility, which is expressed in terms of negative and positive predictive values. The selection of the predictors in the final model is done under the constraint that the predictive values

can take values in prespecified interval. To address this method, the chapter couples notions introduced in the second part of the thesis (estimation of the cutoff and predictive values) with ideas from the first section about biomarker selection.

In a setting where multiple markers are available, we face the challenge of selecting biomarkers for patient classification that belong to the most important class. This means that we select biomarkers such that the PPV (or analogously the NPV) belongs to a prespecified set of values. Here we will consider the PPV over the NPV as the important class. When using a Bayesian model and prior knowledge is avaialble on which variables are more informative for the outcome, then this prior information can be easily incorporated in the selection procedure.

To this end, Bayesian variable selection methods that use shrinkage priors, such as the Laplace prior, the spike and slab, and the horseshoe prior, can be applied. When applying frequentistic regularization methods, the most important aspect is the tuning or selection of shrinkage parameters, as they control the trade-off between parsimony and predictability of the selected model. In a frequentist setting, techniques like cross-validation are adopted to tune the shrinkage parameters. Instead of relying on cross validation for the tuning parameters, taking a Bayesian perspective has the advantage that the penalty parameter can be marginalized over the posterior distribution.

The Lasso estimates from a frequentist analysis (Tibshirani, 1996) can be interpreted as posterior estimates, when the β's have independent identical Laplace priors. Spike and slab priors, the horseshoe, and the appropriate thresholding for inclusion probabilities are also discussed in the third chapter. By assuming a step function to model the probability of response, we incorporate in the selection algorithm the parameters of interest: the cutoff and predictive values. Whilst in Chapter 3 we have considered a single biomarker that was predictive for the outcome, in the multivariate problem presented in Chapter 4, the proposed method simultaneously performs variable selection and cutoff estimation for the risk score taken here as the linear predictor $X\beta$, where β is the vector of coefficients of the selected biomarkers X.

CHAPTER 2

Model selection based on Combined penalties for Biomarker identification

Contributed material

Eleni Vradi, Werner Brannath, Thomas Jaki & Richardus Vonk (2017). Model selection based on combined penalties for biomarker identification. *Journal of Biopharmaceutical Statistics*, $28 : 4, 735 - 749$, *DOI: 10.1080/10543406.2017.1378662*

In Chapter 2, we discuss model selection in a penalized regression framework. We explore different penalization methods that are broadly used in the literature for both regression and classification. This is of particular importance as some of the thorniest issues in drug development can potentially benefit from the use of high dimensional data for biomarker selection. Indeed, the high costs and the long duration for clinical development, coupled with high attrition rates, require the quantification of the risk associated with the transition from early- to late-stage development and biomarkers play an important role in this quantification. Thus, biomarkers panels must be carefully selected at early stages of drug development, as this would affect

the later stages of the process.

Inspired by the work of Liu and Wu (2007) that proposed a model selection approach that combines known penalty functions, we extent this idea to high-dimensional settings by introducing a stepwise forward variable selection algorithm. We briefly discuss different suitable optimization methods and conclude with some simulation results that can be seen in the contributed material.

Preliminaries on penalization methods

From a statistical point of view, biomarker selection translates into statistical modeling for variable selection. In practice, a large number of candidate predictors are available for modeling. Keeping only the most relevant variables in the model improves the interpretation and may enhance the predictability of the resulting model. The correct classification of variables as having (nearly) zero or non-zero effects is important in the field of clinical development concerned with biomarker selection. Inclusion of regressors with zero effect will result in reduced predictive performance of the resulting model and loss in estimation precision, whereas omitting regressors with non-zero effect will lead to biased estimates. Therefore, the ideal biomarker selection method should aim to get rid of any irrelevant biomarkers and include in the model the most important ones.

Especially in the framework of regularization methods, various penalty functions are used to perform variable selection. Frank and Friedman (1993) proposed the bridge regression by introducing the penalty of the form $L_q = \sum_{j=1}^{d} |\beta_j|^q$,$q > 0$, for the vector of regression coefficients $\beta = (\beta_1, \beta_2, \ldots, \beta_d) \in R^d$. When $q \leq 1$, the penalty performs variable selection. The case in which $q = 1$ is the L_1 penalty, and it corresponds to the Least Absolute Shrinkage and Selection Operator (denoted as 'Lasso') (Tibshirani, 1996). In this case, the penalty performs continuous shrinkage and variable selection at the same time.

On the other hand, when $q = 2$, we get the ridge estimator (Hoerl and Kennard,

1970) that shrinks coefficients towards zero, without performing variable selection. The limit of the L_q as $q \to 0$ gives the L_0 penalty, which penalizes the number of non-zero coefficients and, for this reason, is appealing for model selection, if sparse models are of advantage. However, due to its non-convexity and discontinuity at the origin, the corresponding optimization problem becomes difficult to implement in high dimensions.

In genomic research, an L_1 penalty is routinely used due to its convexity and optimization simplicity with regard to its numerical minimization also in high dimensions. The Lasso in the context of parametric models consists of optimizing the log-likelihood function subject to a constrained L_1-norm of the model parameters. It was originally introduced by Tibshirani (1996) for regression and it was proved to simultaneously perform shrinkage and feature selection in linear regression models. The L_1-norm penalization shrinks irrelevant predictors to zero and thereby provides sparse classifiers.

Lasso variable selection has been shown to be inconsistent in certain scenarios (Zou, 2006; Leng et al., 2006; Zhao and Yu, 2006; Yuan and Lin, 2007). Therefore, a new version of the Lasso was introduced, called the adaptive Lasso, where adaptive weights are used for penalizing different coefficients in the L_1 penalty. Zou (2006) showed that the adaptive Lasso enjoys the oracle properties namely it performs as well as if the true underlying model was given in advance. For the latter model, the objective function becomes $-logL + \lambda \sum_{j=1}^{d} w_j L_1(\beta_j)$ with $logL$ the likelihood function, $w_j = \frac{1}{|\beta_j^*|^\gamma}$ being the adaptive weights and β_j^* the ridge regression estimator.

However, the result of the L_1 type regularization may not be sparse enough for practical purposes and a good interpretation. Hence, the development of new methods to obtain sparser solutions has become an essential goal in research on classification and feature selection. A variable selection method that combines the L_1 and L_0 penalties was proposed by Liu and Wu (2007). The authors used a mixed integer

programming algorithm to optimize the objective function

$$-logL + \lambda CL^{\epsilon}_{\alpha} \tag{2.1}$$

with $CL^{\epsilon}_{\alpha} = \alpha L_1 + (1-\alpha)L^{\epsilon}_0$. The L^{ϵ}_0 term is a continuous approximation of L_0 to ease the optimization. The results showed that the method achieved sparser solutions than Lasso, as well as more stable solutions that the L_0 regularization. However, the application was limited to moderate dataset sizes, due to computational inefficiency for large-scale analysis. So far, other combinations of L_q penalties have been proposed, for example by Zou and Hastie (2005) and more recently by Huang et al. (2016), with each of these methods using different optimization algorithms to approach the solution.

In feature selection problems, when biomarkers (e.g. genes) are involved in the same biological pathway, the correlation between them can be high (Segal et al., 2003). Zou and Hastie (2005) proposed a new regularization term, the elastic net penalty, namely $(1-\alpha)L_1 + \alpha L_2$, with $\alpha \in [0,1]$, which is a convex combination of the L_1 and L_2 penalties. Zou and Hastie's work showed that in a regression framework a group of highly correlated predictors can be selected with the elastic net. On the other hand, Lasso fails to select the whole group of correlated variables, as it can only select a subset of independent variables. If the aim is a sparse model, Lasso is more appropriate method, as it will select one variable among a group of correlated predictors.

Optimization Methods

The optimization of the objective function $-logL + \lambda CL_{\alpha}$ is rather challenging since $CL_{\alpha}(\beta = \alpha L_1 + (1-\alpha)L_0)$, is non-convex and non-differentiable at certain points of the parameters' space. The same challenge holds when we consider the objective function $-logL + \lambda CL2_{\alpha}$ with $CL2_{\alpha}(\beta) = \alpha L_2 + (1-\alpha)L_0$).The solution to these multivariable optimization problems is often found by gradient-free algorithms.

Such a solution can be implemented when gradient evaluations are difficult, or in fact when gradients of the underlying optimization problem do not exist. Different methods that offer this feature have been suggested, including Simulated Annealing (Kirkpatrick et al., 1983), Differential evolution (Storn and Price, 1997), Nelder-Mead (Nelder and Mead, 1965), and Hooke−Jeeves (Hooke and Jeeves, 1961).

Furthermore, in the class of numerical methods, the Broyden (1970)- Fletcher (1970)- Goldfarb (1970)- Shanno (1970) (BFGS) variable metric (quasi Newton) method was shown to work well in the optimization of non-smooth and non-convex functions (Lewis and Overton, 2013). However, there is no guarantee that the algorithm converges to the optimal solution. A more detailed review of gradient-free algorithms can be found in Rios and Sahinidis (2013) and Lewis et al. (2000). Among the methods mentioned above, we used the Hooke−Jeeves (HJ) and the BFGS to solve our optimization problem and we report our empirical results which motivated us to develop the proposed stepwise forward method for variable selection.

In the class of direct search methods, the method of Hooke−Jeeves is based on function evaluations using discrete steps and fixed search stepsizes, which are reduced depending on the success of the steps of the algorithm. Repeated searches are performed according to a cyclic coordinate search pattern, followed by a search pattern that is defined by the difference between the starting and ending points of each cyclic coordinate search. On the other hand, the BFGS method uses an approximation of the Hessian matrix in order to find the stationary points of the function to be minimized. The ability to capture the curvature information of the considered function makes the BFGS method very efficcient.

We applied both the HJ and BFGS algorithms, and we compared their performance in the global optimization and in the stepwise forward method. Regarding the minimization of (2.1), our empirical results showed that when we used the HJ algorithm, the value of the objective function was always smaller than the one resulting from the BFGS method, keeping the values of the regularization parameters (α, λ) fixed. However, the BFGS method gave more stable solutions to the optimization problem.

Indeed, the BFGS method always converged to the same solution when repeating the optimization of (2.1), whereas the HJ algorithm was not stable, providing different results every time we optimized (2.1).

As mentioned by Frommlet and Nuel (2016), when the number of predictors d grow large (i.e. $d > 20$), it is not possible to apply algorithms that guarantee to find the optimal solution (see also Furnival and Wilson, 1974). Instead, heuristic search strategies, such as stepwise procedures, may be considered. By using heuristic techniques, we can approximate the optimum solution for the non-smooth, non-convex and NP-hard optimization problem of the equation (2.1), where exact algorithms are not applicable.

Moreover, the issues observed in the behaviour of both optimization methods led us to consider a heuristic stepwise algorithm to deal with model selection in high-dimensional settings. Regarding the stepwise method, using any of the two optimization algorithms, either HJ or BFGS, to minimize the penalized likelihood, our empirical results showed that the values of the objective function were the same in any case, resulting in the same solution.

The stepwise forward approach that we used for variable selection using the penalized likelihood criterion for feature selection, means that candidate predictors are sequentially included in the model, if the inclusion of a variable, on top of the set of variables already in the model, improves the model fit. The algorithm is described in details in Vradi et al. (2017). For the stepwise model selection, we chose to use the BFGS algorithm since it showed a faster convergence than the HJ algorithm. In this stepwise forward selection framework, at each step we optimize the objective function $-logL + \lambda\alpha \sum_{j=1}^{d} L_1(\beta_j)$ using the BFGS algorithm, and perform model selection using the L_0 penalty.

The optimal regularization parameters were tuned by 10-fold cross-validation on the two-dimensional surface (α, λ) using a grid of values. The choice of the optimal parameters was done in a way ackowledging the fact that we are aiming for a com-

promise between good classification performance and low complexity of the model (see details in Vradi et al. (2017)). To evaluate the classification performance of the methods we used the Brier score (Brier, 1950) as a measure for the accuracy of predictions which is the squared distance between the observed statuses y_i and the predicted probabilities $\hat{p}_i$.

Conclusion

In the contributed paper, we have outlined a method for variable selection, which penalizes the likelihood function with a linear combination of L_0 and L_1 or L_2 penalties $(CL, CL2)$ in a stepwise forward variable selection procedure. Our aim was to obtain a sparser model than the one that can be generated by a method that considers the L_1 penalty alone. At the same time, we aimed to achieve a good predictive performance. Therefore, we implemented a stepwise variable selection approach, which at each step performed both shrinkage by using the L_1 penalty and model selection using the L_0 criterion.

An advantage of the proposed method is that we no longer need to consider the continuous approximation to the discontinuous L_0 function and, thus can eliminate the continuity parameter ϵ. Moreover, the importance of our stepwise approach is highlighted by the fact that none of the state-of-the-art global optimization algorithms for non-smooth and non-convex functions has so far achieved satisfactory results.

Lastly, we showed in a simulation study and a real data application that our method generated sparse models while maintaining a good classification performance. This is an essential consideration for classification and screening applications, where the goal is to develop tests by using as fewer features as possible to enhance the interpretability and, potentially, the reproducibility of the results, as well as to control the costs of the test implementation. Our method showed satisfactory results with regard to sparsity and classification performance in terms of AUC of the ROC curves and Brier score. An R-package *stepPenal* (R Core Team, 2015) was developed

around this new method and is available on CRAN.

To conclude, we argue that our method provides a sparser model than the ones obtained by alternative methods, and, at the same time, it maintains similar prediction properties to the ones of other widely used methods, such as Lasso and adaptive Lasso. Even though the stepwise method is just an approximation to the true optimal solutions, it appears to approximate the true optimal solution as well as, and sometimes even better, than the global optimization routine. Moreover, it reduces the computational time considerably. Indeed, the fine tuning of the regularization parameters (α, λ) is an important aspect of penalization methods and can be computationally intensive and time-demanding.

CHAPTER 3

A Bayesian approach to estimate the cutoff and the clinical utility of a biomarker assay

Contributed material

Vradi E, Jaki T, Vonk R, Brannath W (2018). A Bayesian approach to estimate the cutoff and the clinical utility of a biomarker assay. *Statistical methods in medical research*. In press

In Chapter 3, we introduce a method for estimating the cutoff value of a biomarker assay. By using a Bayesian method, we can derive the posterior distribution for the cutoff and the predictive values of a diagnostic test. Firstly, we explain briefly methods that are used for optimal cutoff estimation, as well as the relation of prevalence, sensitivity and specificity with predictive values. The method considers a binary response but it can be extented to time-to-event outcomes as well and we give a real data example of the proposed method on survival data.

Preliminaries in diagnostic tests

In biomedical research, quantitative tests or biomarkers are often used for diagnostic or screening purposes. In fact, these tests simplify binary classifications by setting a cut-off value on the biomarker measurements. The Receiver Operating Characteristic (ROC) curve is a popular graphical method for displaying the discriminatory accuracy of a marker. It is employed to identify two different populations and examine the effectiveness of continuous diagnostic markers, as it distinguishes between diseased (positive test, T^+) and healthy individuals (negative test, T^-). The ROC curve is a plot of sensitivity $Se(c)$ and $1 - Sp(c)$ over all possible threshold values c of the marker. The area under the ROC curve (AUC) is used to evaluate the discriminatory ability of a given marker, i.e. how well the marker can distinguish between diseased and healthy individuals.

The classification of the true status of a patient (healthy or diseased) based on any diagnostic test is not error-free. Thus, it is necessary to measure the errors in order to assess the diagnostic validity of the test, that is, to evaluate its diagnostic accuracy or ability to differentiate between two populations. The test may fail in two different ways: by incorrectly classifying a healthy patient (a false positive, denoted by FP) or, alternatively, by defining a subject as healthy when he or she is in fact diseased (a false negative, denoted by FN).

Table 3.1: Classification of disease results by disease status

	True disease Status		
Test Results	Diseased (Y=1)	Healthy (Y=0)	**Total**
Positive (T^+)	True Positive TP	False Positive FP	TP+FP
Positive (T^-)	False Negative FN	True Negative TN	FN+TN
	TP+FN	FP+TN	N

The evaluation of a positive or negative test result is determined by a threshold, or cut-off value, that needs to be set on (the measurement scale of) the biomarker assay. The cutoff is a value that defines two different classes of observations according to

their biomarker score. For instance, a test result is positive if the biomarker value exceeds the cutoff, i.e. $X > c$ then T^+. On the other hand, the test is negative when the measured value is below the cutoff, i.e. $X \leq c$ then T^-. Optimal cut-off points are often set by using criteria derived from ROC curves. However, it is important to keep in mind that generally, one cannot talk in absolute terms of a "best choice" of cutoff c, as the optimal cut-off depends on the situation which is to be used. As discussed in Perkins and Schisterman (2006) optimal thresholds may vary depending on the underlying criteria.

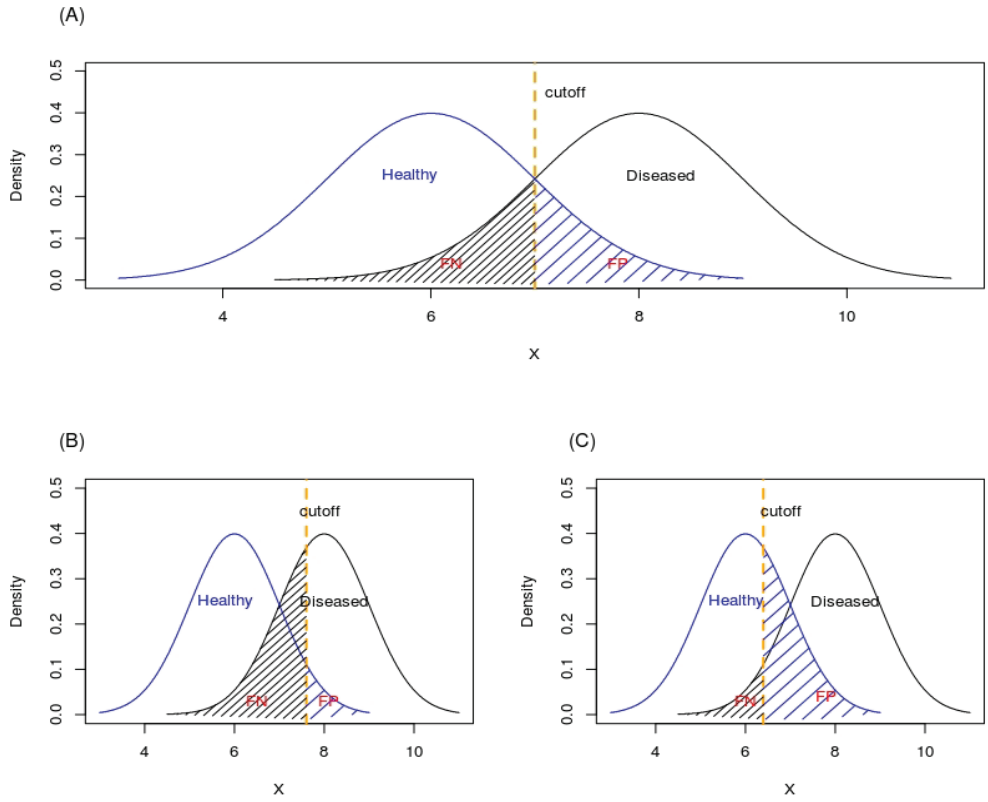

Figure 3.1: (A) Plot of probability density functions of healthy and diseased populations (B),(C) Different cut-off points result in different classification errors. The blue shadowed area depicts the False Positives (FP) and the black shadowed area the False Negatives (FN)

In diagnostic tests with continuous measurements, selecting the optimal cut-off is an important task. As shown in Figure 3.1, different cut-off scores c determine diverse frequencies of correct and incorrect diagnosis. In general, there is a range of potential test results for which the distributions of healthy and diseased subjects overlap. If we want to increase the probability of detecting diseased patients by moving the

cutoff to the left-hand side of the plot, the number of false positives also increases. On the other hand, if the cutoff is moved towards the right-hand side, the number of false positives decreases at the expense of more false negative results. Therefore, as sensitivity decreases, specificity increases, and vice versa. Hence, when selecting the "best" cutoff c, we seek for a balance between sensitivity and specificity measures.

In the literature, we find different optimality criteria that have been suggested to select the best threshold value. Indeed, there are several methods to determine optimal cut-off values. For example, there are different well-known criteria that use the ROC curve, such as the Youden index and the Euclidean index, see e.g. Fluss et al. (2005), Magder and Fix (2003). The Youden index (J), one of the most frequently used methods, maximizes the sum of the two correct classification probabilities, i.e $J = max_c\{Se(c) + Sp(c) - 1\}$

There are also measures that rely on the maximization of the diagnostic odds ratio (DOR) function Glas et al. (2003), defined as the ratio between the odds of TPR and FPR over all possible cut-point values of X. Böhning et al. (2011) has showed that the DOR strategy is no longer recommended since it might easily lead to the choice of cut-off values on the boundary of the parameter range of X.

However, ROC based methods do not provide information on the diagnostic accuracy for specific patients. The use of sensitivity and specificity was criticized for example by Moons and Harrell (2003), claiming that sensitivity and specificity are not the correct parameters to characterize diagnostic accuracy, as these parameters are of limited relevance to practice. According to the authors, the characteristics of a test should rather be evaluated based on the actual patient population. Especially when a diagnostic test is used for classification purposes, clinicians are mainly concerned about the predictive ability of the test.

The assessment of correct classifications can be facilitated by the use of positive and negative predictive values (PPV and NPV, respectively). PPV and NPV are functions of the accuracy of the test and of the overall prevalence of the disease. Therefore, they can be used to define the clinical utility of a diagnostic test for

classification purposes in a specific population. The positive and negative predictive values are defined as $PPV = P(Y = 1|T^+)$ and $NPV = P(Y = 0|T^-)$.

The positive predictive value (PPV), or the predictive value of a positive test, is the probability of developing a disease (or responding to a treatment) when the test result is positive. Hence, the positive predictive value can be estimated based on the proportion of patients with a positive test result that ultimately proved to be diseased (Y=1). From Bayes theorem, the PPV is expressed in terms of sensitivity (Se) and specificity (Sp) measures, and the prevalence π of the disease under study:

$$PPV = \frac{\pi Se}{\pi Se + (1 - Sp)(1 - \pi)} = \frac{TP}{TP + FP} \tag{3.1}$$

where the sensitivity $Se = P(T^+|y = 1)$ and $1 - Sp = P(T^+|y = 0)$. The prevalence of the disease is $\pi = P(Y = 1)$. Analogously, the NPV can be expressed in terms of sensitivity, specificity and prevalence as:

$$NPV = \frac{(1 - \pi)Sp}{\pi(1 - Se) + Sp(1 - \pi)} = \frac{TN}{FN + TN} \tag{3.2}$$

and the complementary 1-NPV is given by

$$1 - NPV = \frac{\pi(1 - Se)}{\pi(1 - Se) + (1 - Sp)(1 - \pi)} = \frac{FN}{FN + TN}$$

Similarly to the Youden index, that can be considered as a summary measure of the ROC curve, the predictive summary index (PSI) can be seen as a summary index of the predictive ROC (PROC) curve (Linn and Grunau (2006)). Analogously to the ROC, the PROC is a plot of the PPV versus 1-NPV over a range of cutoff values c. For any threshold c, the PSI is defined as:

$$PSI(c) = PPV(c) + NPV(c) - 1.$$

Obviously, $PSI \in [0, 1]$. The threshold value c^* that maximizes $PSI(c)$ may be selected as the optimal cutpoint, that is $c^* = argmax_c\{PSI(c)\}$. The PSI provides a criterion to choose an optimal cut-off value that takes the predictive values into account. The criterion describes the likelihood of correctly diagnosing a diseased patient after a positive test result, while at the same time it indicates how likely is to misdiagnose a patient with a disease after a negative test. It is important to note that a test with high sensitivity and specificity may have a low positive predictive value if applied to low prevalence populations.

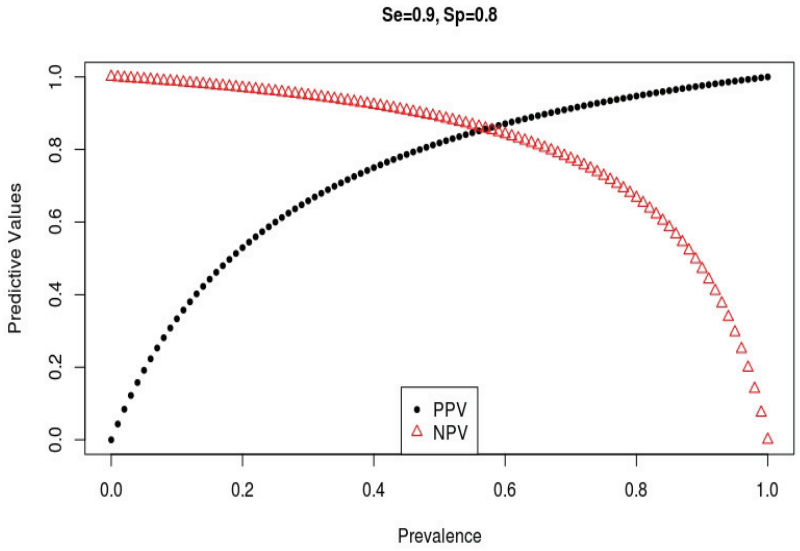

Figure 3.2: Plot of the predictive values for varying values of prevalence and fixed sensitivity=0.9 and specificity=0.8

In order to illustrate the impact of the prevalence on the clinical utility of a test, we have plotted in Figure 3.2 the PPV and NPV over a range of prevalence values for fixed sensitivity and specificity using the expressions in formulas (3.1) and (3.2). Assuming a test with 80% specificity and 90% sensitivity (both of which are theoretically independent of prevalence), we see that at low prevalence, a negative test result is more likely to be true than a positive result. This plot shows the reciprocal relationship between PPV and NPV: when prevalence increases, the PPV raises while the NPV decreases. It also points out the potential loss of PPV in case a test is used in a context of low disease prevalence, rather than in populations with a high disease prevalence.

Figure 3.3 shows how altering either sensitivity or specificity affects the predictive values of a test at three different values of prevalence ($\pi = 0.1$, $\pi = 0.5$ and $\pi = 0.7$). As we see, in case of a low prevalence disease, the PPV is less influenced by a loss of sensitivity than a loss of specificity. Moreover, there is a positive correlation between NPV and specificity (i.e the rate of false positives is low).

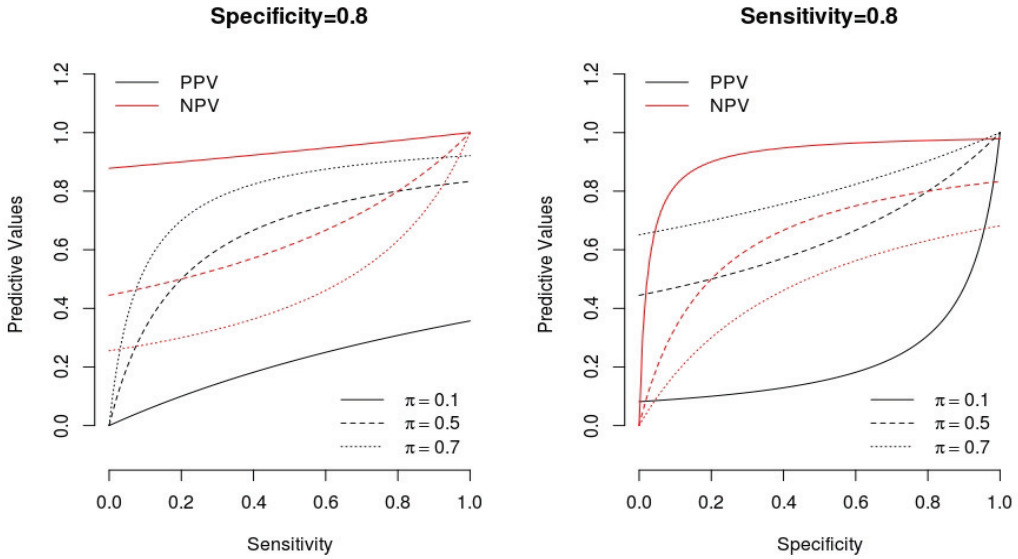

Figure 3.3: Plot of the predictive values for fixed specificity and varying sensitivity (left panel) and fixed sensitivity and varying specificity (right panel) where we consider a prevalence of 0.1 (solid line), 0.5 (dashed line) and 0.7 (dotted line). PPV and NPV are depicted in black and red lines, respectively.

Bayesian methods for cutoff estimation

Lately, Subtil and Rabilloud (2010) introduced a Bayesian method for estimating both the optimal cut-point of a quantitative marker and its credible interval, when the diagnostic test is based on optimizing a utility function of parameters including prevalence, sensitivity, specificity, net benefit, and net costs of the test. First, the authors assumed that the biomarker values follow normal distributions for the two populations, i.e. $N(\mu_H, \sigma_H^2)$ for the healthy and $N(\mu_D, \sigma_D^2)$ for the diseased subjects.

Assuming non-informative priors for the parameters (mean and variance for the diseased and non-diseased groups), they used an MCMC algorithm to sample from

the posterior distribution of $\mu_H, \mu_D, \sigma_H, \sigma_D$. Then, for each obtained sample in the MCMC chain, they optimized the utility function

$$U(c) = 1 - \Phi(\frac{c - \mu_D}{\sigma_D}) + R\Phi(\frac{c - \mu_H}{\sigma_H})$$

where Φ denotes the standatrd normal distribution function and R is a function of the prevalence and net cost and net benefit. The estimated optimal cutoffs over all samples form the posterior distribution of the optimal cutoff. To deal with variance or mean heterogeneity, Subtil and Rabilloud (2014) extended this Bayesian method to different biomarker population distributions like the student-t distribution or a mixture of Dirichlet distributions.

Lunceford (2015) discussed the estimation of the clinical utility of a biomarker assay in the context of predictive enrichment studies in the oncology field. According to the author, the key point is that enrichment involves an additional biomarker-based inclusion criterion with the aim to treat those who are more likely to respond. The motivation of this study was to select a cut-off value for a potential predictive biomarker that could be used as a criterion to enroll patients. The author proposed a method based on the integral representations of clinical utility measures. By implementing a Bayesian approach to assess clinical utility measures, he facilitates cutoff decision-making, without considering the actual cutoff estimation.

In the contributed paper, we proposed a Bayesian method to estimate the cutoff of a biomarker assay, and more importantly the uncertainty around this estimate. We assumed that the probability of response can be modeled by a step function to facilitate the decision-making process regarding, for example, the selection of patients that were more likely to respond to a personalized treatment. We used a step function to model the probability of response, which served as an approximate model to facilitate the classification of patients into two groups with pronounced differences in their response rates. For a single biomarker X and a vector of n responses Y, we modeled the response probability as $p(x) = P(Y = 1|X) = \mathbb{1}_{\{x \leq cp\}} p_1 + \mathbb{1}_{\{x > cp\}} p_2$, where p_1 , p_2 and cp are considered as unkown parameters.

The advantage of using the step function is that both the cutoff and the predictive values are parameters of the model. Even in case the assumption of a step function is strong and the model is misspecified, the estimates of the assumed step function are consistent for the parameter values, for which the assumed model minimizes the distance from the true distribution in terms of Kullback-Leibler divergence (Kullback and Leibler, 1951); (Huber et al., 1967); (Bunke et al., 1998).

We suggested a mixture prior for the cutoff to acknowledge potential prior-data conflict. One of the components of the mixture prior is an informative precise prior, i.e. the the true cutoff lies in an interval of high probability, the other component of the mixture prior is an uninformative prior taking values in the span of the biomarker measurements. For the predictive values we assumed Uniform distributions on the unit interval. We run the analysis using MCMC Metropolis-Hastings (Metropolis et al., 1953); (Hastings, 1970), to approximate the posterior distribution.

Application to survival data - Cox regression

In the contributed paper Vradi et al. (2018), we applied our proposed method to two real datasets. The first set included data on the use of the Prostate Specific Antigen as a diagnostic marker for prostate cancer, the second example is concerned with the time-to-event data where the thickness of the tumor is a marker associated with increased risk of death from melanoma.

In addition, we introduce an application of our method on survival data. The data is from a study for a cancer treatment for which a composite biomarker score has been developed, which is believed to predict the overall survival (OS). To maintain confidentiality, the study is anonymized. The aim of this analysis is to estimate a cut-off value of the composite score such that the patients below and above the cutoff have a pronounced difference in their survival probabilities.

In a survival setting, we assume the following: Let T_i^* denote the event time for subject i, $i = 1, ..., n$. Due to censoring, instead of observing T_i^*, we observe the

bivariate vector (T_i, Δ_i) where $T_i = min(T_i^*, C_i)$ and $\Delta_i = \mathbb{1}_{\{T_i^* \leq C_i\}}$ with $\mathbb{1}$ the indicator function and C_i is the censoring time. In conclusion, we assume that we observe the i.i.d. pairs (T_i, Δ_i) for n individuals. We apply the Cox proportional hazards model, the most widely used semi-parametric regression model analysis for survival times. The estimator of β is given by optimizing the partial likelihood for the Cox model given by:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{e^{X^T \beta}}{\sum_{l=1}^{n} Y_l(t_i) e^{X^T \beta}} \right)^{\Delta_i}$$

with $Y_i(t) = \mathbb{1}_{\{T_i \geq t\}}$ denotes the indicator of whether the individual i is at risk at time t and $Y(t) = \sum_{i=1}^{n} Y_i(t)$ denotes the total number from the sample that is at risk at time t and X denotes the vector of covariates for individual i.

Back to our data application, we consider the composite score X_i measured on $i = 1, ..., 499$ patients. In order to fit the Cox model for cut-off estimation, we assume a dummy variable $Z = \mathbb{1}_{\{X > cp\}}$. Conditional on the biomarker measurement, a proportional hazards model is assumed to hold. Specifically the hazard at time t for patient i is taken as

$$h(t, Z) = h_0(t) e^{\beta Z} = \begin{cases} h_0(t), & \text{if } X \leq cp \\ h_0(t) e^{\beta}, & \text{if } X > cp \end{cases}$$

where β quantifies the log difference in the hazards between the two groups of patients (above and below cp). Assuming a unifrom prior for all the parameters, results regarding the posterior density of the cutoff and the survival curves for the patients below and above the estimated cutoff are shown in Figure 3.4. The posterior median for *cp* is 1.63 with 95% credible interval $(0.39 - 2.13)$ and the log-rank test (Breslow et al., 1984) results in a p-value equal to 0.006, indicating a significant difference between the two survival curves for the two group of patients. The hazard ratio was found to be equal to 1.35 (95% CI, 1.08-1.68), i.e. the patients above the cutoff are 35% more likely to die than patients below the cutoff value.

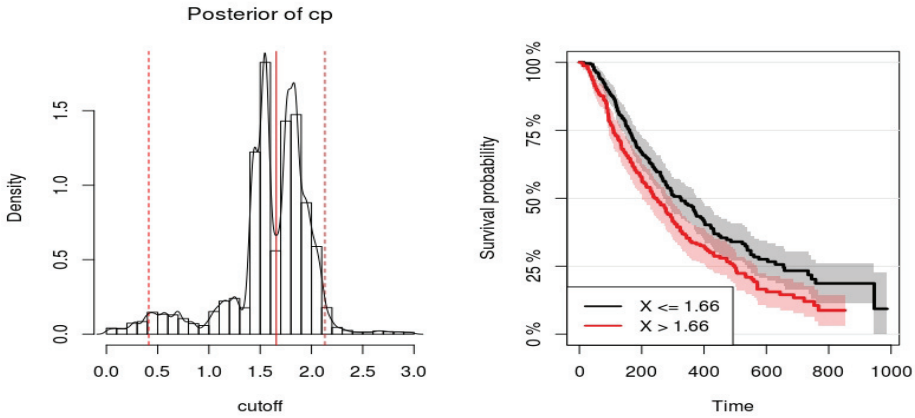

Figure 3.4: Left figure: Posterior density for the cutoff. The red vertical line represents the median of the posterior distribution and the dashed vertical lines the 95% credible intervals. Right figure: Kaplan Meier plot for the overall survival of the two groups defined by the estimated cutoff taken as the posterior median

Conclusion

To summarize, our results suggest that the proposed Bayesian method provides a practical strategy for cutoff selection and for clinical utility estimation of biomarker assays. In our simulation study, see Vradi et al. (2018), we have concluded that the Bayesian 95% credible intervals have good coverage probabilitiy, which are close to the nominal value both for the cutoff and the predictive values. In our simulations, we considered different prior specifications, including a very informative, an uninformative and mixtures of them. In all cases we obtained satisfactory results, and especially when precise prior information is available, the parameters were nearly unbiased with high precision. The Bayesian approach is very tractable in estimating the distribution of the parameters of interest. Moreover, our results show that point estimators (e.g. posterior mean) are essentially unbiased in every tested scenario, prior constellations and sample size assumptions.

An important aspect to consider when using Bayesian inference is that we can incorporate prior information into the cut-point through the prior distribution. In our work, we present four different prior specifications, including uninformative, informative, and mixture priors. In all cases, estimation gives satisfying results and,

especially when precise prior information is available, the estimated parameters are nearly unbiased and highly precise. We suggest using a mixture prior, since it works well in practice and it is robust with respect to potential prior-data conflicts.

A possible extension of the proposed method could be to take into account multiple cutoffs, i.e. consider a step function with more than one step to fit the data. Based on this assumption, we could identify different risk groups according to their probability of response. Of major interest, it would be to consider multiple biomarkers and investigate how to combine them in a single score. Then using the derived score we can apply the proposed method for patient classification. In the following Chapter 4, we introduce a Bayesian method for biomarker selection and classification by controlling the test's clinical utility as expressed by the predictive values.

CHAPTER 4

A Bayesian method for variable selection and classification under constrained clinical utility

Contributed material

Vradi E, Jaki T, Vonk R, Brannath W(2018). Bayesian variable selection and classification with control of predictive values. Manuscript submitted.

In Chapter 4, we bring together two topics: one is the idea of variable selection and the other is the topic of cutoff estimation for classifying patients according to their response probability. This work is motivated by the important task of selecting a set of predictive biomarkers (discussed in the first chapter), derive a risk score out of the selected markers and subsequently estimate a cutoff value (discussed in the second chapter), that will be used as a criterion for distinguishing two groups of patients, e.g. the classification of patients in resonders and non-responders' groups. In this chapter we propose a method for combining the two ideas using Bayesian methodology. This leads to a method to select variables for patient classification, taking the desired clinical utitily into account already at the selection stage of the

process.

Preliminaries in Bayesian variable selection

In the Bayesian setting, the model selection problem is transformed into the form of parameter estimation. Bayesian variable selection procedures are more informative than penalization methods, because they automatically address the model selection uncertainty. Rather than searching for a single optimal model, Bayesian analysis provides estimates of the posterior probability of all models within the considered class of models. From a Bayesian perspective, variable selection is performed by imposing prior distributions on the regression coefficients. There are two main alternatives for the choice of the prior: two component discrete mixture priors and a variety of continous shrinkage priors.

For the first approach, known as the Spike-and-Slab and suggested by Mitchell and Beauchamp (1988), assumes that the prior distribution for each regression coefficient β_j is a mixture of a point mass at zero ($\beta_j = 0$) and a diffuse uniform distribution elsewhere. George and McCulloch (1993) propose a stochastic seach variable selection (SSVS) where the subset selection is derived from a hierarchical normal mixture model. The Spike-and-Slab prior is equivalent to Bayesian Model Averaging (BMA) over the variable combinations and often has good performance in practice. One drawback is that the results can be sensitive to prior choices of the slab components or to prior inclusion probability. Moreover, inference with BMA can be computationally demanding with a large number of variables, due to the huge model space search.

Continuous shrinkage priors on the other hand, place absolutely continous distributions on the entire parameter vector to yield a sparse estimator. Within the class of shrinkage priors, scale mixtures of normals and normal-gamma distributions for the coefficients β have received extensive attention (Andrews and Mallows (1974), West (1987)) and more recently by Fernandez and Steel (2000), Griffin et al. (2010), Griffin and Brown (2012), Liang et al. (2008) and the references therein. Carvalho et al.

(2009) and Polson and Scott (2010) introduced the global-local shrinkage priors that adjust for sparsity via the *global* shrinkage and identify signals by *local* shrinkage parameters.

The Bayesian lasso (Park and Casella, 2008; Hans, 2009) typically refers to the use of double-exponential shrinkage prior for the regression coefficients β in the normal linear regression model $y = X\beta$ where y is the $n-$vector of observations and X is a $n \times d$ vector of predictor variables. The regressions coefficients are distributed as $\beta \sim DE(0, 1/\lambda)$. The posterior mode under the double-exponential prior is equivalent to the frequentist Lasso estimate by Tibshirani (1996).

The horseshoe prior (HS), that belongs to the family of global-local shrinkage priors, was introduced by Carvalho et al. (2010). It refers to a hierarchical-shrinkage prior for the regression coefficients where their standard deviation is the product of a local (λ_j) and global (τ) scaling parameter. It is given by

$$\beta_j|\lambda_j, \tau \sim Normal(0, \lambda_j^2\tau^2)$$

$$\lambda_j \sim Cauchy^+(0, 1) \qquad \text{and} \qquad \tau \sim Cauchy^+(0, 1)$$

where the $Cauchy^+(0, 1)$ is a standard half Cauchy distribution on the positive reals with location parameter 0 and scale 1. The intuition behind the horseshoe prior according to Piironen and Vehtari (2016) is the following: the global parameter τ pulls all the weights (β_j) globally towards zero, i.e. estimate the overall sparsity, while the thick half-Cauchy tails for the local shrinkage parameter λ_j allow some of the weights to escape shrinkage, i.e. is able to flag the non-zero elements of β. Different levels of sparsity are based on the value of τ, i.e. when $\tau \to 0$ will shrink all β_j to zero, whereas with large τ there is very little shrinkage towards zero. The advantage of the HS is that it is shown to be robust at handling unknown sparsity and large outlying signals.

Although the choice of independent half standard Cauchy priors for all λ_j is less

debated, the choice for the prior on τ has raised discussions (see (Piironen and Vehtari, 2016)). For example, it is recommented to use $\tau \sim C^+(0, \alpha^2)$, where $C^+(0, \alpha^2)$ denotes the half-Cauchy distribution with location 0 and scale α and the authors provide some suggestions on the choice of α. Van Der Pas et al. (2014) and Piironen and Vehtari (2016) discussed an intuitive way to design a prior for τ based on assumptions about the number of nonzero elements in the vector β. Theoretical justification why it is a desirable choice the half-Cauchy prior for scale parameters in hierarchical models is given in Gelman et al. (2006) and Polson et al. (2012).

Spike and Slab prior (SpSl) is a popular prior for sparse Bayesian estimation and refer to the prior of β written as a two-component mixture of priors. There are two different specifications for the spike component: spikes specified by an absolutely continous distribution and spikes defined by a point mass at zero, the so called Dirac spike. The slab component has its mass spread over a wide range of plausible values for the regression coefficients. The spike and slab prior is specified as $p(\beta_j) = (1 - \gamma_j)p_{spike}(\beta_j) + \gamma_j p_{slab}(\beta_j)$. A detailed overview regarding different spike-and-slab priors as well as comparisons amongst them can be found in Malsiner-Walli and Wagner (2011). However, the authors conclude that spike-and-slab priors do not discriminate between variables with zero and weak effects. The choice of a smaller variance for the slab component will not solve the problem but it will cause an increase in the posterior inclusion probabilities of all effects, even for the zero effects.

An absolutely continuous spike is in principle specified by any unimodal continuous distribution with mode at zero. Here we consider the prior for β_j as a spike taken to be a delta spike (Dirac spike) at the origin δ_0 combined with a normal slab. The prior inclusion probability is $\gamma_j \sim Bernoulli(\pi)$ and the prior for β_j is written as

$$\beta_j|\gamma_j, \sigma \sim (1 - \gamma_j)\delta_0 + \gamma_j Normal(0, \sigma^2)$$

$$\sigma^2 \sim Inv - Gamma(a_s, b_s)$$

Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990; Casella and George, 1992) is a sampling algorithm based on Markov Chain Monte Carlo (MCMC) techniques. In Bayesian inference, MCMC methods are used for obtaining a sequence of samples that can be used to approximate the posterior joint distribution of the parameters. Gibbs sampling is a method for sampling from the conditional densities of the parameters and is particularly well-suited to Bayesian networks (express the conditional dependence structure between random variables) which are typically specified as a collection of conditional distributions. The theory of MCMC guarantees that the stationary distribution of the samples (discarding the initial burnin samples) generated under Gibbs algorithm is the target joint posterior that we are interested in (Gilks et al., 1996).

Bayesian Variable Selection and Thresholding

From a Bayesian perspective, the variable selection property is *ad hoc*. Suppose there are d covariates which are candidates for inclusion in the model. Then each model $m \in M$ can be naturally represented by a $d-$vector γ of binary indicator variables determining whether or not a covariate is included in the model. Let γ be the $d-$vector where $\gamma_j = 1$ if the predictor variable X_j is included in the model and $\gamma_j = 0$ otherwise.

Under the continuous prior distribution, the probability of the event $\{\beta_j = 0\}$ is zero. In order to make posterior inferences about events such as $\{\beta_j = 0\}$ prior probability mass must be allocated to these events. In the Bayesian setting, placing prior mass on the events $\{\beta_j = 1\}$ is equivalent to assigning a prior distribution to the space of regression models that are to be considered. If $\pi(m)$ is the prior probability of model m, the posterior probability of the regression model is

$$\pi(m|y) = \frac{\pi(y|m)\pi(m)}{\sum_{m \in M} \pi(y|m)\pi(m)}$$

where M is the collection of all possible models and $\pi(y|m)$ is the marginal likelihood of the observed data y under model m. A review of Bayesian variable selection

methods for different prior considerations and thresholding procedures has been thoroughly discusssed in O'Hara et al. (2009).

Kuo and Mallick (1998) considered an approach for variable selection where they proposed a prior distribution $\pi(\beta)$ that is independent of γ (and therefore from model M) so that $\pi(\beta|\beta_{\setminus j}, \gamma) = \pi(\beta|\beta_{\setminus j})$, where $\beta_{\setminus j}$ denotes all terms of β except β_j. The advantage of this approach is that implementation is straitforward and requires only the specification of the prior for β. The full conditional posterior distribution is given by:

$$\pi(\beta_j|y, \gamma, \beta_{\setminus j}) \propto \begin{cases} \pi(y|\gamma, \beta)\pi(\beta_j|\beta_{\setminus j}) & , \gamma = 1 \\ \pi(\beta_j|\beta_{\setminus j}) & , \gamma = 0 \end{cases}$$

According to their proposal, the linear predictor for the generalized linear model determined by γ may be written by

$$\eta = \sum_{j=1}^{d} \gamma_j X_j \beta_j \tag{4.1}$$

The selection on the important predictors is done by examining the marginal posterior inclusion probabilities $\pi(\gamma_j = 1|y)$. In practice the most important variables are identified by their frequence of appearance in the Gibbs samplers. If selection of one model for prediction is desired, Barbieri et al. (2004) showed (by a mathematical proof) that the Median Probability model (denoted as MP), defined to be the model containing all variables with

$$\pi(\gamma_j = 1|y) \geq 1/2 \tag{4.2}$$

is the best model with regard to predictive performance. Another approach to identify non-important covariates can be based on posterior credible intervals, where selection of predictors is done by checking whether zero is included in the coefficients' credible interval or not (van der Pas et al., 2017; Li et al., 2010). The latter method has a poor performance in high dimensional settings and does not quantify the importance of each covariate (i.e. as inclusion probablities do). It is worth mentioning that the MP model suggested by Barbieri et al. (2004) assumes orthogonal

predictors, an assumption that is rarely apply in practice. Moreover, the MP model was derived assuming Gaussian noise and the theory was not applied to classification problems. However, Piironen and Vehtari (2017) applied the MP model on classification data with correlated predictors and showed through numerical experiments that the MP model performs well in terms of correct variable selection as well as predictive performance of the model.

When an absolutely continuous prior is used, the inclusion indicators are sampled conditionally on β_j. For the horseshoe prior, Carvalho et al. (2010) suggested that variable selection is done based on the posterior mean

$$\mathbb{E}(\beta_j|y) = (1 - \mathbb{E}(\kappa_i|y_i))y_i$$

where $\kappa_i = \frac{1}{(1+\lambda_j^2\tau^2)}$ is the shrinkage factor which describes how much coefficient β_j is shrunked towards zero. Carvalho et al. (2010) interpreted the shrinkage factor as inclusion probabilities by thresholding $1 - \kappa_j \geq 0.5$. According to Carvalho et al. (2010); Scott and Berger (2006) this thresholding approach induces a multiple testing rule, i.e. reject the null hypothesis $\beta_j = 0$ if $1 - \kappa_j \geq 0.5$, that controls the rate of type I error. The Spike-and-Slab prior with a dirac spike is modeling the inclusion probability directly and therefore variable selection is based on the marginal posterior of γ_j.

Bayesian variable selection and cutoff estimation with constrain on PPV

In some diagnostic situations it is essential to include the control of classification measures or most importantly, as we emphasized also in Chapter 3, the control of the clinical utility of the test. In this contributed paper, we are interested in the selection of the markers for a risk score, which lead to a prespecified lower bound of admissible values for PPV. For example, keep the biomarkers which, when combined, lead to a PPV greater or equal than 90%.

It is common that the evaluation of the clinical utility is done after the selection of a subset of biomarkers. That means that the important information requiring a

high PPV value is not considered in the optimization procedure and therefore the computed PPV is not necessarily satysfying the constraint. Here we aim to include the predictive values in the optimization algorithm. We used a step function to model the probability of response as proposed in Chapter 3. Therefore the predictive values (PPV and 1-NPV) and the cutoff *cp* are parameters of the model and the Bayesian methodology can be easily applied. The application of the step function in a setting with multiple predictors can be easily done by assuming a combined score, or a risk score, of the selected biomarkers (instead of a single marker as in Chapter 3).

Our aim is the classification of patients as well as identification of the important variables characterizing the different response groups. We achieve both goals simultaneously by combining the proposed step function model (for estimating the cutoff value) to discriminate and classify patients with Bayesian variable selection methods for the identification of important predictors. The modeling is done under the constraint that the selected variables provide a classifier with high positive (or negative) predictive value. In the Bayesian setting, the constrain that the PPV belongs to a predetermined interval of high values, i.e. $(90\% - 100\%)$, is achieved through the prior distributions on the predictive values. For example, using a shrinkage prior for β and the linear predictor as in equation (4.1), the probability of response is modeled as

$$p = P(y = 1|\beta X) = \begin{cases} p_1 = P(y = 1|\beta X \leq cp) \\ p_2 = P(y = 1|\beta X > cp) \end{cases}$$

The prior F for the coefficients β can be taken as one of the priors proposed in this chapter and the parameters cp, p_1 and p_2 are assumed to follow a Uniform distribution, where l is the lower bound of the constraint on the positive predictive value.

$$\begin{gathered} \beta \sim F \\ cp \sim Uniform(a, b) \\ p_1 \sim Unif(0, p_2) \qquad \text{and} \qquad p_2 \sim Unif(l, 1) \end{gathered} \tag{4.3}$$

The estimation of the cutoff on the risk score is done in two steps. Since the selection of biomarkers in a Bayesian model is *ad hoc*, the estimation of the parameters of the step function should be done accordingly in two steps. We are interested in the conditional distribution of the risk score for the finally selected biomarkers. Therefore, by fitting the model in (4.3) we obtain the marginal posterior density for the cutoff but not conditionally on the selected biomarkers. By fitting the model assuming the step function at the first step, we derive a vector of the estimated effects $\hat{\boldsymbol{\beta}}$, i.e. the posterior means of the β_j. To obtain the conditional distribution of the cutoff, we fit the model with the step function but now for fixed $\hat{\boldsymbol{\beta}}$ and data (X, y). This second step in our estimation procedure can be seen as an empirical Bayes approximation.

Step 1: In the first step we fit a Bayesian model with a shrinkage prior and approximate the posterior distribution

$$f(\beta, cp, p_1, p_2, \theta | X, y) \propto \mathcal{L}(\beta, cp, p_1, p_2 | X, y) \times f(cp) \times f(p_1) \times f(p_2) \times f(\beta | \theta)$$

where $f(\cdot)$ denotes the density function, $\mathcal{L}$ is the likelihood function and θ denotes the hyperparameters in the model (θ will be different for the different priors for β). The selection of the variables that are included in the final model is based on the thresholding criterion in (4.2), i.e. keep those β_j that the marginal inclusion probability is greater than 1/2. We take as point estimates the posterior means of the selected β_j and calculate the estimated risk score as $\hat{\beta} X$.

Step 2: Fit the model in (4.3) now for fixed $\hat{\beta}$ and approximate the posterior density $f(cp, p_1, p_2 | X, y, \hat{\beta})$ implementing the Gibbs sampling approach. The estimated cutoff and predictive values are taken as posterior means of the marginal posterior distributions for each of the parameter.

Conclusion

We propose a Bayesian method that simultaneously performs variable selection assuming a shrinkage or mixture prior (Laplace, Spike-and-Slab and Horseshoe), and cutoff estimation under the constraint that the PPV belongs to a prespecified interval of plausible values. The optimization with the constraint provides the best classifier with a PPV in the prespecified interval.

In the simulation study we conducted (see the contributed paper), we used the Laplace, the SpSl and the HS priors and evaluated their performance in terms of correct variables included in the model as well as in terms of classification performance. We constrained p_2 to have a lower bound of 0.8, i.e we used a Uniform distribution in the interval (0.8,1). We found that the True Positives (TP), defined as the number of times in the simulation runs that the important variables were selected in the model was at least 70%. The False Positives (FP), defined as the average number that a non-important variable was included in the final model, did not exceed 20% on average.

We compared our poposed method with the two-stage approach, described as follows; the algorithm firstly does variable selection assuming a logistic model and at the second stage, considering the variables selected from stage one, estimates the cutoff and predictive values by fitting the step function model. The proposed method differs from the two-stage appoach at the first stage, where we fit a step function instead of a logistic. By considering the step function at the first stage, we are able to include information about the cutoff and predictive values already at the selection stage.

The classification performance of the proposed method was assessed by calculating the Brier score (Brier, 1950) as well as the bias of the predictive values on a testing (validation) dataset. Regarding the classification error, the HS prior resulted in slightly higher errors for the proposed approach compared to the Laplace ans SpSl priors. We observed the higher error in the scenario where the predictors are highly

correlated.

Regarding bias, the proposed approach behaved similarly to the two-stage procedure for all the priors, resulting in nearly unbiased estimators for the predictive values. We showed that the bias of p_1 was on average close to zero for all the different priors, and was found slightly higher for p_2 which was on average about 10%. We observed a tendency to slightly overestimate p_1 and underestimate p_2. When the correlation among predictors is high (i.e. > 0.5), the bias of p_2 was slightly higher than in other scenarios, and on average 17%.

Bibliography

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–102.

Barbieri, M. M., Berger, J. O., et al. (2004). Optimal predictive model selection. *The annals of statistics*, 32(3):870–897.

Berry, D. A. (2006). Bayesian clinical trials. *Nature reviews Drug discovery*, 5(1):27.

Biomarkers Definitions Working Group (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*, 69(3):89–95.

Böhning, D., Holling, H., and Patilea, V. (2011). A limitation of the diagnostic-odds ratio in determining an optimal cut-off value for a continuous diagnostic test. *Statistical methods in medical research*, 20(5):541–550.

Breslow, N. E., Edler, L., and Berger, J. (1984). A two-sample censored-data rank test for acceleration. *Biometrics*, pages 1049–1062.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthey Weather Review*, 78(1):1–3.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms. *IMA journal of applied mathematics*, 6(3):222–231.

Bunke, O., Milhaud, X., et al. (1998). Asymptotic behavior of bayes estimates under possibly incorrect models. *The Annals of Statistics*, 26(2):617–644.

Buyse, M., Michiels, S., Sargent, D. J., Grothey, A., Matheson, A., and De Gramont, A. (2011). Integrating biomarkers in clinical trials. *Expert review of molecular diagnostics*, 11(2):171–182.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174.

Denayer, T., Stöhr, T., and Van Roy, M. (2014). Animal models in translational medicine: Validation and prediction. *New Horizons in Translational Medicine*, 2(1):5–11.

DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: new estimates of r&d costs. *Journal of health economics*, 47:20–33.

DiMasi, J. A., Hansen, R. W., and Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *Journal of health economics*, 22(2):151–185.

EMA (2009). Ich e16 genomic biomarkers related to drug response: context, structure and format of qualification submissions.

FDA (2010). Guidance for the use of bayesian statistics in medical device clinical trials.

Fernandez, C. and Steel, M. F. (2000). Bayesian regression analysis with scale mixtures of normals. *Econometric Theory*, 16(1):80–101.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The computer journal*, 13(3):317–322.

Fluss, R., Faraggi, D., and Reiser, B. (2005). Estimation of the youden index and its associated cutoff point. *Biometrical journal*, 47(4):458–472.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Frommlet, F. and Nuel, G. (2016). An adaptive ridge procedure for l0 regularization. *PloS one*, 11(2):e0148620.

Furnival, G. M. and Wilson, R. W. (1974). Regressions by leaps and bounds. *Technometrics*, 16(4):499–511.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical

models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). Introducing markov chain monte carlo. *Markov chain Monte Carlo in practice*, 1:19.

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., and Bossuyt, P. M. (2003). The diagnostic odds ratio: a single indicator of test performance. *Journal of clinical epidemiology*, 56(11):1129–1135.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24(109):23–26.

Griffin, J. E. and Brown, P. J. (2012). Structuring shrinkage: some correlated priors for regression. *Biometrika*, 99(2):481–487.

Griffin, J. E., Brown, P. J., et al. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications.

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

Hooke, R. and Jeeves, T. A. (1961). "direct search"solution of numerical and statistical problems. *Journal of the ACM (JACM)*, 8(2):212–229.

Huang, H.-H., Liu, X.-Y., and Liang, Y. (2016). Feature selection and cancer classification via sparse logistic regression with the hybrid l1/2+ 2 regularization. *PloS one*, 11(5):e0149675.

Huber, P. J. et al. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. Berkeley, CA.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. *science*, 220(4598):671–680.

Kola, I. and Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature reviews Drug discovery*, 3(8):711.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

Lavezzari, G. and Womack, A. (2016). Industry perspectives on biomarker qualification. *Clinical Pharmacology & Therapeutics*, 99(2):208–213.

Leng, C., Lin, Y., and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica*, pages 1273–1284.

Lewis, A. S. and Overton, M. L. (2013). Nonsmooth optimization via quasi-newton methods. *Mathematical Programming*, 141(1-2):135–163.

Lewis, R. M., Torczon, V., and Trosset, M. W. (2000). Direct search methods: then and now. *Journal of computational and Applied Mathematics*, 124(1-2):191–207.

Li, Q., Lin, N., et al. (2010). The bayesian elastic net. *Bayesian Analysis*, 5(1):151–170.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Linn, S. and Grunau, P. D. (2006). New patient-oriented summary measure of net total gain in certainty for dichotomous diagnostic tests. *Epidemiologic Perspectives & Innovations*, 3(1):11.

Liu, Y. and Wu, Y. (2007). Variable selection via a combination of the l 0 and l 1 penalties. *Journal of Computational and Graphical Statistics*, 16(4):782–798.

Lunceford, J. K. (2015). Clinical utility estimation for assay cutoffs in early phase oncology enrichment trials. *Pharmaceutical statistics*, 14(3):233–341.

Magder, L. S. and Fix, A. D. (2003). Optimal choice of a cut point for a quantitative diagnostic test performed for research purposes. *Journal of clinical epidemiology*, 56(10):956–962.

Malsiner-Walli, G. and Wagner, H. (2011). Comparing spike and slab priors for bayesian variable selection. *Austrian Journal of Statistics*, 40(4):241–264.

McGonigle, P. and Ruggeri, B. (2014). Animal models of human disease: challenges in enabling translation. *Biochemical pharmacology*, 87(1):162–171.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Moons, K. G. and Harrell, F. E. (2003). Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. *Academic radiology*, 10(6):670–672.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The computer journal*, 7(4):308–313.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Perkins, N. J. and Schisterman, E. F. (2006). The inconsistency of optimal cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American journal of epidemiology*, 163(7):670–675.

Piironen, J. and Vehtari, A. (2016). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. *arXiv preprint arXiv:1610.05559*.

Piironen, J. and Vehtari, A. (2017). Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics*, 9:501–538.

Polson, N. G., Scott, J. G., et al. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rios, L. M. and Sahinidis, N. V. (2013). Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56(3):1247–1293.

Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of bayesian multiple testing. *Journal of statistical planning and inference*, 136(7):2144–2162.

Segal, M. R., Dahlquist, K. D., and Conklin, B. R. (2003). Regression approaches for microarray data analysis. *Journal of Computational Biology*, 10(6):961–980.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656.

Simon, R. (2010). Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Personalized medicine*, 7(1):33–47.

Storn, R. and Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359.

Subtil, F. and Rabilloud, M. (2010). A bayesian method to estimate the optimal threshold of a longitudinal biomarker. *Biometrical Journal*, 52(3):333–347.

Subtil, F. and Rabilloud, M. (2014). Estimating the optimal threshold for a diagnostic biomarker in case of complex biomarker distributions. *BMC medical informatics and decision making*, 14(1):53.

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

Van Der Pas, S., Kleijn, B., Van Der Vaart, A., et al. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2):2585–2618.

van der Pas, S., Szabó, B., van der Vaart, A., et al. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis*, 12(4):1221–1274.

Vradi, E., Brannath, W., Jaki, T., and Vonk, R. (2017). Model selection based on combined penalties for biomarker identification. *Journal of Biopharmaceutical Statistics*, 28(4):735–749. PMID: 29072549.

Vradi, E., Jaki, T., Vonk, R., and Brannath, W. (2018). A bayesian model to estimate the cutoff and the clinical utility of a biomarker assay. *Statistical Methods in Medical Research*, 0(0). PMID: 29966502.

Wehling, M. (2006). Translational medicine: can it really facilitate the transition of research "from bench to bedside"? *European journal of clinical pharmacology*, 62(2):91–95.

Wehling, M. (2009). Assessing the translatability of drug projects: what needs to be scored to predict success? *Nature reviews Drug discovery*, 8(7):541.

Wehling, M. (2011). Drug development in the light of translational science: shine or shade? *Drug discovery today*, 16(23-24):1076–1083.

Wendler, A. and Wehling, M. (2012). Translatability scoring in drug development: eight case studies. *Journal of translational medicine*, 10(1):39.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648.

Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):143–161.

Zhang, W., Wang, J., and Menon, S. (2018). Advancing cancer drug development through precision medicine and innovative designs. *Journal of biopharmaceutical statistics*, 28(2):229–244.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Contributed Manuscripts

Journal of Biopharmaceutical Statistics

ISSN: 1054-3406 (Print) 1520-5711 (Online) Journal homepage: http://www.tandfonline.com/loi/lbps20

Model selection based on combined penalties for biomarker identification

Eleni Vradi, Werner Brannath, Thomas Jaki & Richardus Vonk

To cite this article: Eleni Vradi, Werner Brannath, Thomas Jaki & Richardus Vonk (2018) Model selection based on combined penalties for biomarker identification, Journal of Biopharmaceutical Statistics, 28:4, 735-749, DOI: 10.1080/10543406.2017.1378662

To link to this article: https://doi.org/10.1080/10543406.2017.1378662

Published online: 26 Oct 2017.

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
2018, VOL. 28, NO. 4, 735–749
https://doi.org/10.1080/10543406.2017.1378662

Taylor & Francis
Taylor & Francis Group

Model selection based on combined penalties for biomarker identification

Eleni Vradi [a], Werner Brannath[b], Thomas Jaki [c], and Richardus Vonk[a]

[a]Department of Research and Clinical Sciences Statistics, Bayer AG, Berlin, Germany; [b]Institute of Statistics, Competence Center for Clinical Trials Bremen, Faculty 3, University of Bremen, Bremen, Germany; [c]Department of Mathematics and Statistics, Medical and Pharmaceutical Statistics Research Unit, Lancaster University, Lancaster, United Kingdom

ABSTRACT

The growing role of targeted medicine has led to an increased focus on the development of actionable biomarkers. Current penalized selection methods that are used to identify biomarker panels for classification in high-dimensional data, however, often result in highly complex panels that need careful pruning for practical use. In the framework of regularization methods, a penalty that is a weighted sum of the L_1 and L_0 norm has been proposed to account for the complexity of the resulting model. In practice, the limitation of this penalty is that the objective function is non-convex, non-smooth, the optimization is computationally intensive and the application to high-dimensional settings is challenging. In this paper, we propose a stepwise forward variable selection method which combines the L_0 with L_1 or L_2 norms. The penalized likelihood criterion that is used in the stepwise selection procedure results in more parsimonious models, keeping only the most relevant features. Simulation results and a real application show that our approach exhibits a comparable performance with common selection methods with respect to the prediction performance while minimizing the number of variables in the selected model resulting in a more parsimonious model as desired.

KEYWORDS

Biomarker panels; combined penalties; model selection; penalized regression; regularization; sparsity; stepwise variable selection; treatment responder

Introduction

The high costs and long duration of clinical development, paired with high levels of attrition, require the quantification of the risk when moving from early to late stage clinical development, and biomarkers may play an important role in this quantification. However, only rarely the number of variables (biomarkers) in the resulting panel plays an active role in selection procedures. Variable selection is an important aspect in the determination of such panels in the framework of high-dimensional statistical modeling. In practice, a large number of candidate predictors are available for modeling. Keeping only the relevant variables in the model enhances the interpretation and may increase the predictability of the resulting model.

Particularly in the framework of regularization methods, various penalty functions are used to perform variable selection. Frank and Friedman (1993) proposed the bridge regression by introducing the penalty of the form $L_q = \sum_{j=1}^{d} \left|\beta_j\right|^q,\ q>0$, for the vector of regression coefficients $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \ldots, \beta_d\right) \in \mathbb{R}^d$. When $q \leq 1$ the penalty performs variable selection. The case where $q = 1$ is the L_1 penalty and corresponds to the Least Absolute Shrinkage and Selection Operator (Lasso) (Tibshirani, 1995). It performs continuous shrinkage and variable selection at the same time, whereas for $q = 2$ we get the

CONTACT Eleni Vradi, MSc eleni.vradi@bayer.com Research and Clinical Sciences Statistics, Bayer AG, Muellerstrasse 178, 13342 Berlin, Germany.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lbps.

ridge estimator (Hoerl and Kennard, 1970) that shrinks coefficients toward zero but it does not perform variable selection. The limit of the L_q as $q \to 0$ gives the L_0 penalty, which penalizes the number of non-zero coefficients and thus is appealing for model selection, if sparse models are of advantage. However, due to its non-convexity and discontinuity at the origin, the corresponding optimization problem becomes difficult to implement in high dimensions.

In genomic research, an L_1 penalty is routinely used due to its convexity and optimization simplicity. However, the result of the L_1 type regularization may not be sparse enough for a good interpretation. The development of methods to obtain sparser solutions than through L_1 penalization methods is becoming essential part in the classification and feature selection area. A variable selection method that combines the L_1 and L_0 penalties was proposed by Liu and Wu (Liu and Wu, 2007). They used a mixed integer programming algorithm for optimization of the objective function. The results showed that their method achieved sparser solutions than Lasso and more stable solutions that the L_0 regularization. However, the application was limited to moderate data sizes, due to computational inefficiency for large-scale problems. Other combinations of L_q penalties have been proposed so far (Zou and Hastie, 2005) and recently (Huang et al., 2016) with each of these methods using a different optimization algorithm to approach the solution.

In this article, we propose a method for variable selection that penalizes the likelihood function with a linear combination of L_0 with L_1 or L_2 penalties (*CL, CL2*) in a stepwise forward variable selection procedure. The aim is to obtain a model that is sparser than the model with the L_1 penalty alone and at the same time achieve a good predictive performance. Moreover, a strong motivation for the proposed stepwise forward variable selection method is that state-of-the-art global optimization algorithms for non-smooth and non-convex functions do not provide satisfactory results. In Section 2, we define the *CL* and *CL2* penalties and present the algorithm for solving the penalized logistic regression problem with these combined penalties. In Section 3, we use simulated data to evaluate the performance of our method, and we compare it to Lasso and adaptive Lasso both in terms of correct variable selection (true covariates with $\beta_j \neq 0$) as well as predictive performance. Finally, we show an application of our method for classification and variable selection on a real dataset with protein measurements to identify the least number of predictors that can best classify responders and non-responders to a treatment.

Methods

Regularization

Suppose we have data ($\boldsymbol{X}$, $\boldsymbol{y}$), where $y = (y_1, y_2, \ldots, y_n)$ is the vector of responses and $\boldsymbol{X}$ is an $n \times d$ matrix of predictors. We will assume that the observations are independent and the predictors standardized. With linear predictor $\eta = \boldsymbol{X}^T\boldsymbol{\beta}$ and link function g the generalized linear model is expressed as

$$g(E(\boldsymbol{y}|\boldsymbol{X})) = \eta \tag{2.1}$$

Under the regularization framework, the estimated coefficients $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_d\right) \in \mathbb{R}^d$ are obtained by minimizing the objective function $-logL + \lambda P(\boldsymbol{\beta})$, and are given by:

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{argmin}\{-logL + \lambda P(\boldsymbol{\beta})\}$$

where $P(\beta)$ is a regularization term. The parameter $\lambda > 0$ is a tuning parameter and $-logL$ is the negative log-likelihood. One of the most popular and commonly used regularization method is the L_1 regularization (Lasso), where $P(\beta) = \sum_{j=1}^{d}\left|\beta_j\right|$. Setting $\lambda = 0$ reverses the Lasso to maximum likelihood estimation. On the other hand, a very large λ will completely shrink β to zero thus leading to the empty or null model. In general, moderate values of λ will cause shrinkage of the solutions toward zero, and some coefficients may be exactly zero.

Other types of L_1 regularization include the adaptive Lasso, where adaptive weights are used for penalizing different coefficients in the L_1 penalty and which was shown to have the oracle property (Zou, 2006). A variable selection and estimation procedure is said to have the oracle property i) if it selects the true model with probability tending to 1 and ii) if the estimated penalized coefficients are asymptotically normal, with the same asymptotic empirical variance as the estimator based on the true model.

However, the L_1 type regularization is consistent only under rather restrictive assumptions (Zhao and Yu, 2006) and the coefficient estimates are severely biased due to shrinkage (Meinshausen and Yu, 2009; Fan and Li, 2001). Although the L_0 norm, where $P(\beta) = \sum_{j=1}^{d} 1_{\beta_j \neq 0}$ and $1_{\beta_j \neq 0}$ is the indicator function of whether $\beta_j \neq 0$, tend to yield the sparsest solutions, its implementation in high-dimensional data becomes an NP hard optimization problem and is not computationally feasible. Classical information criteria like AIC (Akaike, 1974) or BIC (Schwarz, 1978) lie in the general class of the regularization $\lambda P(\beta) = \lambda \sum_{j=1}^{d} 1_{\beta_j \neq 0}$ for suitable choices of λ. In order to gain a more concise and sparse solution and while keeping a high predictive accuracy of the classification model, we propose a regularization term that combines the L_0 with L_1 or L_2 norms (Liu and Wu, 2007). Figure 1 plots

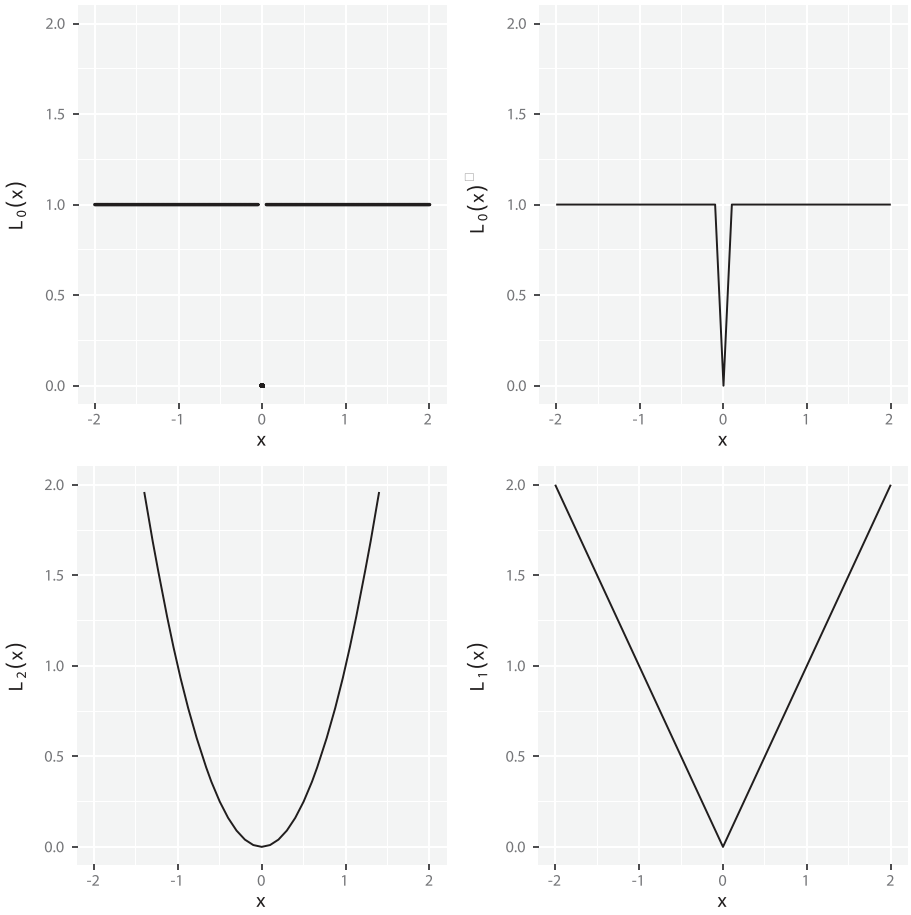

Figure 1. Plot of the $L_0, L_0\ \varepsilon, L_2, L_1$ *norms with* $\varepsilon = 0.1 and d = 1$

the penalty functions L_1 and L_2 in the bottom panel and the L_0 penalty in the top panel. Unlike L_2, the penalty terms L_1 and L_0 are singular at the origin and thus perform variable selection (Fan and Li, 2001).

The combined $L_0 + L_1$ penalty

Following Liu and Wu (Liu and Wu, 2007) the penalization term is defined as $CL_\alpha{}^\varepsilon(\beta) = (1-a)L_0{}^\varepsilon + aL_1$, where $0 \le a \le 1$ is a weighting parameter between $L_0{}^\varepsilon$ and L_1 penalties, with $L_0{}^\varepsilon$ given by:

$$L_0{}^\varepsilon(\beta) = \begin{cases} 1, & |\beta| \ge \varepsilon \\ \frac{|\beta|}{\varepsilon}, & |\beta| < \varepsilon \end{cases} \tag{2.2}$$

Clearly $CL_1{}^\varepsilon = L_1$ $(a = 1)$ and $CL_0{}^\varepsilon = L_0{}^\varepsilon$ $(a = 0)$ are special cases of $CL_\alpha{}^\varepsilon$. Discontinuity at the origin of L_0 makes the optimization difficult, and therefore we consider the continuous approximation to L_0 defined by (2.2). The limit of $L_0{}^\varepsilon(\beta)$ when $\varepsilon \to 0$ is $L_0(\beta)$ itself. When $\varepsilon > 0$ is small, $L_0{}^\varepsilon(\beta)$ is a good approximation to $L_0(\beta)$ (Figure 1 top right). The estimated coefficients are obtained by minimizing the objective function

$$-logL + \lambda \sum_{j=1}^{d} CL_\alpha{}^\varepsilon\left(\beta_j\right) \tag{2.3}$$

and are given by

$$\hat{\beta}_{CL_\alpha{}^\varepsilon} = \underset{\beta}{argmin}\left\{-logL + \lambda \sum_{j=1}^{d} CL_\alpha{}^\varepsilon\left(\beta_j\right)\right\}.$$

The combined $L_0 + L_2$ penalty

We now consider another combination, the L_0 norm with L_2. The motivation for combining the L_0 norm with L_2, is to consider a penalty that will join the nice properties of the L_2 and those of the L_0 norm, which is to perform variable selection (L_0) and keep in the model groups of variables that are correlated (L_2). In theory, a strictly convex function provides a sufficient condition for such grouping of variables and the L_2 penalty guarantees strict convexity. The grouping effect refers to the simultaneous inclusion (or exclusion) of correlated predictors in the model.

The penalization term is now defined $CL2_\alpha{}^\varepsilon(\beta) = (1-a)\ L_0{}^\varepsilon + a\ L_2$, where $0 \le a \le 1$. The $L_0{}^\varepsilon$ term introduced above is for variable selection and the L_2 penalty shrinks the coefficients toward zero with no contribution to variable selection. Figure 2 gives a graphical representation of the regularization terms $CL_{0.3}{}^{0.1}$, L_1, L_2, $CL2_{0.5}{}^{0.1}$.

The stepwise forward procedure

In their paper, Liu and Wu (2007) proposed a global algorithm to solve the corresponding difficult non-convex problem (mixed integer linear programming). However, the applicability was restricted to moderate datasizes. As mentioned by Frommlet F. and Nuel G. (Frommlet and Nuel, 2016), when the number of predictors grows large ($d > 20$) it is not possible to apply algorithms which guarantee to find the optimal solution (Furnival and Wilson, 1974) and instead heuristic search strategies like stepwise procedures may be considered. By using heuristic techniques, we can approximate the solution of the non-smooth, non-convex, and

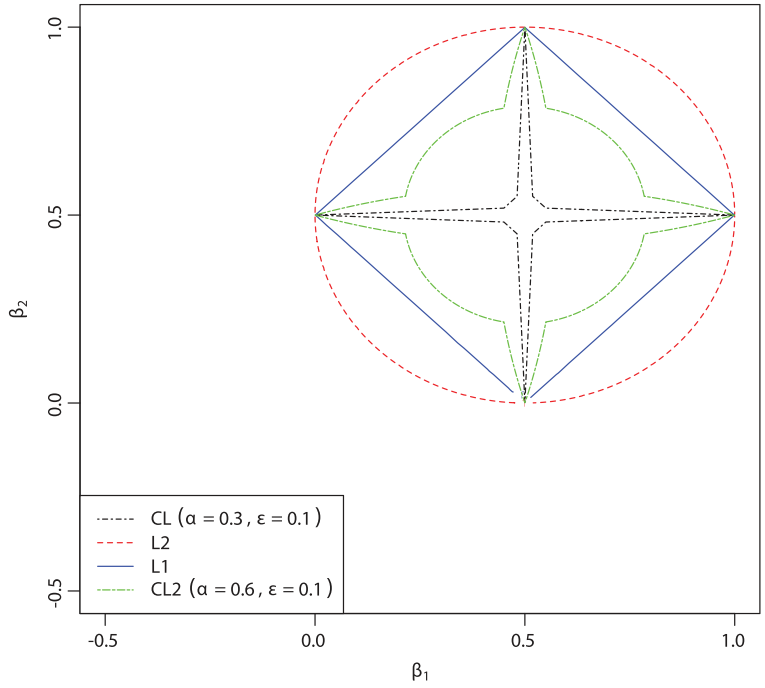

Figure 2. Two dimensional contour plot for the regularization terms $CL_{0.30.1}, L_1, L_2, CL2_{0.50.1}$ *with* $\varepsilon = 0.1$*andd* $= 2$. The black dashed curve represent the CL penalty ($\alpha = 0.3, \varepsilon = 0.1$) and the green shaped curve is the contour of the CL2 ($\alpha = 0.5, \varepsilon = 0.1$). The outer contour shows the shape of the ridge penalty, while the blue solid line is the Lasso penalty.

NP-hard optimization problems like the one in equation (2.3), where exact algorithms are not applicable for such minimization problems.

The optimization of the objective function (2.3) is rather challenging since $CL_\alpha{}^\varepsilon(\beta)$ and $CL2_\alpha{}^\varepsilon(\beta)$ are non-convex and non-differentiable at some points of the parameters' space. We apply the Broyden (1970), Fletcher (1970), Goldfarb (1970), and Shanno (1970) (BFGS) variable metric (quasi-Newton) method, which is shown to work well for the optimization of non-smooth and non-convex functions (Lewis and Overton, 2009 ; Lewis and Overton, 2013; Curtis and Que, 2015). The BFGS method uses an approximation of the Hessian matrix to find the stationary points of the function to be minimized. Its ability to capture the curvature information of the considered function renders the method so effective.

We propose to use a stepwise forward variable selection using the previously introduced penalized likelihood criterion for feature selection that can be used effectively in high-dimensional data. In this stepwise forward selection framework, at each step we optimize the objective function $-logL + \lambda\, a\, L_1$ using the BFGS algorithm and obtain

$$\hat{\beta}_{L_1} = \underset{\beta}{argmin}\left\{ -logL + \lambda\, a \sum_{j=1}^{d} L_1\left(\beta_j\right)\right\}.$$

The selected model is based on the criterion that minimizes the value of

$$-\, logL\left(\hat{\beta}_{L_1}\right) + \lambda a L_1\left(\hat{\beta}_{L_1}\right) + \lambda\ (1-a)\ L_0\left(\hat{\beta}_{L_1}\right) \tag{2.4}$$

The suggested algorithm is described as follows:

Step 1

- Given a set of d standardized predictors $X_1, X_2, \ldots, X_d$ and a response $y_i \in \{0,1\},\ i = 1, \ldots n$ we consider all possible univariate models $(M_1, M_{2,} \ldots,\ M_d)$

$$M_1:\ Y \sim \beta_0 + \beta_1 X_1, \ldots\ M_2:\ Y \sim \beta_0 + \beta_2 X_2, \ldots\ M_3:\ Y \sim \beta_0 + \beta_3 X_3, \ldots, M_d:\ Y \sim \beta_0 + \beta_d X_d$$

- Estimate $\hat{\beta}_{L_1}^{M_1}, \ldots,\ \hat{\beta}_{L_1}^{M_d}$ and keep $M_j, j \in \{1, \ldots, d\}$ that gives the smallest value of the function (2.4), e.g. keep variable X_2

Step 2

- With the model chosen in step 1 (e.g. M_2) and in an additive way we consider all the $d-1$ models (M') by adding the remaining $d-1$ variables one at a time to the model M_2.

$$\begin{aligned} M'_1\ &: \mathrm{Y} \sim \beta_0 + \beta_2 \mathbf{X}_2 + \beta_1 X_1 \\ M'_2 &: \mathrm{Y} \sim \beta_0 + \beta_2 \mathbf{X}_2 + \beta_3 X_3 \\ &\vdots \\ M'_d &: \mathrm{Y} \sim \beta_0 + \beta_2 \mathbf{X}_2 + \beta_d X_d \end{aligned}$$

- Keep the model that minimizes the function in (2.4),

Step 3

- Continue adding single variables until the value of the function (2.4) in the current step is bigger than its value in the previous step.

The advantage of using the function (2.4) instead of (2.3) in the optimization is that we no longer need to consider the continuous approximations to the discontinuous L_0 function and therefore we eliminate the number of parameters by the continuity parameter ε. The reason why we can do so is that within each step the L_0-penality term remains constant (since the dimension of the model is fixed) and hence play no role in the determination of the regression coefficients. The L_0-penality term only plays a role for the stopping criterium.

Sparse logistic regression with combined penalties

As a particular example, we consider the binary linear regression model (2.1), where $y \in \{0,1\}$, is a vector of n observed binary outcomes, $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \ldots,\ \beta_d\right) \in \mathbb{R}^d$ is the vector of coefficients. The link function is the logit function $logit(p) = \log\left(\frac{p}{1-p}\right)$, where p is the conditional event probability and is given by

$$p = P(\boldsymbol{y} = 1|\boldsymbol{X}) = \frac{e^{\eta}}{1 + e^{\eta}} \tag{2.5}$$

The coefficient estimates are obtained by minimizing (2.3) with the log-likelihood

$$logL = L(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i)\log(1 - p_i)$$

Results

In this section, we examine via simulations the performance of logistic regression when models are selected and estimated with the above introduced combined penalties (*CL, CL2)* by either stepwise forward selection as introduced in Section 2 *(stepCL and stepCL2)* or by global minimization (*CL, CL2)*. In the stepwise model selection scheme, we also examine the performance of the stepwise adaptive L_1 model with the $\lambda\ (1-a)\ L_0$ selection criterion *(stepAdaCL)*. For that model, the objective function to minimize is $-logL+\lambda\, a\ \sum_{j=1}^{d} w_j L_1\left(\beta_j\right)$, where $w_j=\frac{1}{\left|\beta_j^*\right|^{\gamma}}$ are the adaptive weights and $\left|\beta_j^*\right|$ is the ridge regression estimator. The estimation is done with the stepwise algorithm described in Section 2.4.

Although the proposed method is a stepwise variable selection procedure, we did not consider the comparison with other stepwise methods like the stepwise BIC or AIC, as they tend to perform poorly when the dimension is large relative to the sample size and are usually too liberal, that is, they tend to select a model with many spurious covariates (Chen and Chen, 2008). As mentioned by Zhang and Shen (Zhang and Shen, 2010) these criteria may be inadequate due to their nonadaptivity to the model space and infeasibility of exhaustive search.

We include the global minimization in spite of the disadvantages mentioned in Section 2 for a comparison. We also consider the results from Lasso (L_1 penalty) and the adaptive Lasso. We compare the different methods in terms of the fraction of correctly selected variables and the prediction classification accuracy. The real data come from a biomarker study of protein measurements with the objective to select biomarkers that potentially discriminate between responders and non-responders.

Simulation study

We simulate data for varying number of predictors. We consider two settings, one high-dimensional data where the number of predictors (d) exceeds the number of samples (n), and a setting where the sample size is smaller than the dimensionality of the data. We assume multivariate normal predictors $X_1,\ldots,\ X_d$ with pairwise correlation ρ (compound symmetry). Let ρ denote the correlation between variables $X_m,\ X_l$ where $m,l\ \in\{1,\ldots,\mathrm{d}\}$, $m\neq l$.

The true model that was used to generate the outcome has k informative covariates $X_k,\ k\in\mathbb{Z},\ 1<k<d$. We consider a classification problem with $\boldsymbol{y}$ a binary response and standard normally distributed predictors $\boldsymbol{X}\sim MVN(0,\Sigma)$, where Σ is the covariance matrix. We consider the logistic model with logit link function, $logit(p)=X^T\beta$, as described above with p the probability of $y = 1$ given X as defined in (2.5). In other words, each component of the response vector y is viewed as a realization of a Bernoulli random variable with probability of success p.

Four scenarios will be presented here, each with $n = 100$ samples.

1. Scenario 1: $d\leq n$, correlation $\rho=0.5$.

We consider $d = 50$ covariates, with $k = 3$ informative predictors and coefficient vector $\beta=\left(3,1.5,2,\underbrace{0,\ldots,0}_{47}\right)$. The correlation between the three informative predictors is $\rho=corr(X_m,X_l)=0.5$, $m\neq l$ and $m,l=1,2,3$.

2. Scenario 2: High-dimensional setting $d\geq n$, correlation $\rho=0.5$

We consider $d = 150$ covariates, with $k = 15$ informative predictors with $\beta=\left(\underbrace{3,\ldots,}_{3},\underbrace{-3.5,\ldots,}_{3},\underbrace{1.5,\ldots,}_{3},\underbrace{5,\ldots,}_{4},\underbrace{-2,..,}_{2}\underbrace{0,\ldots,0}_{135}\right)$. The correlation ρ between the 15 informative predictors is $\mathrm{corr}(X_m,\ X_l)=0.5,\ m\neq l$ and $m,l=1,\ldots,15$.

3. Scenario 3: High-dimensional setting $d\geq n$, correlation $\rho=0.7$.

The dataset consists of $n = 100$ samples and $d = 200$ covariates, with $k = 15$ informative predictors with $\beta = \left(\underbrace{2, \ldots,}_{4} \underbrace{-3, \ldots,}_{3} \underbrace{1.5, \ldots,}_{4} \underbrace{-2, \ldots,}_{3} \underbrace{0, \ldots, 0}_{185} \right)$. The correlation ρ between the 15 informative predictors is $\text{corr}(X_m,\ X_l) = 0.7,\ m \neq l$ and $m, l = 1, \ldots, 15$.

4. Scenario 4: High-dimensional setting $d \geq n$, block correlation.

We consider $d = 200$ covariates, with $k = 16$ informative predictors with $\beta = \left(1, \underbrace{4, \ldots,}_{4} \underbrace{-3, \ldots,}_{3} \underbrace{1.5, \ldots,}_{4} \underbrace{-2, \ldots,}_{4} \underbrace{0, \ldots, 0}_{185} \right)$. In this scenario, there are two groups (blocks) of correlated predictors and one single independent feature. The coefficients of $d–k = 184$ variables were set to zero, $\beta_r = 0,\ r = 184, \ldots,\ 200$. The correlation between predictors in block 1 is $\text{corr}(X_m,\ X_l) = 0.4,\ m \neq l$ and $m, l = 1, \ldots, 7$ and the correlation among predictors in block 2 is $\text{corr}(X_m,\ X_l) = 0.7,\ m \neq l$ and $m, l = 8, \ldots, 16$.

Tuning of parameters

All analyses were done in R version 3.2.3 (R Core Team, 2015). For the Lasso and adaptive Lasso, the *glmnet* library was used and all the functions that were used for the combined penalty approach can be found in the R-package "*stepPenal*", available on the CRAN. For the adaptive lasso, weights were estimated by ridge regression and then used for a weighted L_1 penality in estimation of β. The optimal regularization parameters for the methods *stepCL, stepAdaCL, CL, CL2, stepCL2* were tuned by 10-fold cross-validation on the two-dimensional surface (a, λ) using a grid of values. The choice of the optimal parameters was done in the following way. For each configuration of (a, λ) in the grid, the AUC of the ROC curves on the validation set was computed in each of the 10 validation sets. The average of the 10 AUCs was reported together with its standard deviation.

Selection of (a, λ) was based on the interval $A = [maxAUC - sdAUC,\ maxAUC)$ where maxAUC is the maximum average AUC and sdAUC is the standard deviation of the AUCs corresponding to the (a, λ) with maximum average AUC. The (a, λ) that corresponds to the median of the AUCs in the interval A was chosen for the final model fitting. In case that more than one configurations yields the median of the AUCs, we select the configuration with the largest λ and smallest a, to obtain the sparsest model. The use of the interval A acknowledges the sample variability and the fact that we are aiming for a compromise between good classification performance and complexity of the model. In other words, a small decrease in the AUC of the ROC curve is acceptable in return to a less complex model. The Lasso and adaptive Lasso were also tuned by 10-fold cross-validation on the one-dimensional space (λ), using the default settings in R in the function *cv.glmnet* and the measure type *"auc"*.

Simulation results

The different classifiers were built by the estimated tuning parameters on the training set. Then, the obtained classifiers were applied to the testing set for classification and prediction. For the testing set, we simulated data from the same distribution as the training set for $n = 1000$ samples. We simulated 1000 datasets on which we applied all methods. For each method, we computed the mean classification performance of the models on the testing sets measured by the AUC of the ROC curve (test AUC). This is a measure for the discrimination ability of the model to correctly distinguish the two classes of the response. The complexity of the resulting model was measured by the ratio of correctly selected variables (true covariates with $\beta_j \neq 0$) to the total variables selected by the model. We will call this ratio *RCV* for the rest of the paper.

Table 1. Simulation results for all four scenarios based on 1000 replications. The table summarizes the complexity of the models in terms of correct non-zero estimates. The first column to the left shows the median value (1st and 3rd quantile) of the correct variables included in the model. The second column is the median (1st and 3rd quantile) of the total variables each model selects. The third column is the average RCV (standard deviation), the ratio of the correct variables over the total variables selected. The *stepCL, stepAdaCL* and *stepCL2* are the stepwise methods and *CL, CL2* are the global optimization methods.

	Scenario											
Methods	1	2	3	4	1	2	3	4	1	2	3	4
(n = 100)	Number of true $\beta_j \neq 0$ Median (1QN-3QN)				Total Variables Median (1QN-3QN)				RCV[1] Mean (sd)			
stepCL	3	5	4	5	3	6	6	7	0.85	0.78	0.71	0.79
	(2–3)	(4–6)	(4–4)	(5–6)	(2–4)	(5–7)	(5–6)	(6–8)	(0.18)	(0.17)	(0.15)	(0.15)
stepCL2	3	6	5	6	3	7	6	9	0.80	0.77	0.82	0.75
	(3–3)	(5–7)	(4–5)	(5–7)	(3–3)	(6–8)	(6–7)	(8–10)	(0.18)	(0.16)	(0.11)	(0.15)
stepAdaCL	3	5	4	5	3	8	7	6	0.82	0.70	0.62	0.71
	(2–3)	(4–6)	(4–5)	(4–6)	(3–3)	(5–10)	(6–8)	(5–9)	(0.18)	(0.20)	(0.14)	(0.20)
CL (L0+L1)	3	6	7	7	4	8	9	9	0.72	0.77	0.82	0.75
	(3–3)	(5–7)	(7–8)	(6–8)	(3–5)	(7–10)	(8–10)	(8–11)	(0.19)	(0.13)	(0.12)	(0.13)
CL2 (L0+L2)	3	8	9	9	7	13	19	22	0.48	0.64	0.47	0.41
	(3–3)	(7–10)	(7–10)	(7–10)	(5–9)	(10–17)	(15–26)	(17–26)	(0.19)	(0.18)	(0.18)	(0.14)
Lasso (L1)	3	7	7	8	4	11	9	17	0.66	0.67	0.77	0.55
	(3–3)	(6–9)	(6–8)	(7–9)	(3–9)	(7–21)	(7–14)	(10–26)	(0.33)	(0.26)	(0.24)	(0.23)
Adaptive Lasso	3	12	8	9	3	15	8	15	0.79	0.81	0.88	0.65
	(2–3)	(9–13)	(6–9)	(8–11)	(3–6)	(11–17)	(6–11)	(11–20)	(0.30)	(0.14)	(0.15)	(0.17)

[1] $RCV = \frac{\#\text{ correctly selected variables with true } \beta_j \neq 0}{\#\text{ total variables selected}}$

Table 2. Simulation results for all four scenarios based on 1000 simulated datasets. The table summarizes the performance of the models for all scenarios. The Brier score and the area under the ROC curve are calculated on the testing set.

	Scenario							
	1	2	3	4	1	2	3	4
Methods	Brier score Mean (sd)				AUC (test) Mean (sd)			
stepCL	0.09	0.12	0.08	0.12	0.95	0.92	0.96	0.91
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.03)
stepCL2	0.10	0.11	0.07	0.11	0.94	0.92	0.97	0.92
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.01)	(0.03)
stepAdaCL	0.10	0.13	0.08	0.14	0.95	0.91	0.96	0.89
	(0.02)	(0.03)	(0.02)	(0.03)	(0.01)	(0.03)	(0.02)	(0.03)
CL (L0+L1)	0.10	0.11	0.07	0.12	0.96	0.94	0.98	0.93
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
CL2 (L0+L2)	0.12	0.12	0.09	0.15	0.94	0.93	0.97	0.90
	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.03)	(0.02)	(0.03)
Lasso (L1)	0.11	0.11	0.08	0.12	0.96	0.93	0.97	0.92
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.01)	(0.02)
Adaptive Lasso	0.11	0.09	0.07	0.11	0.95	0.95	0.98	0.93
	(0.03)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)

This ratio takes values between zero and one. When the model selects none of the informative variables it is zero and it becomes one, when the selected model includes only the k informative covariates. The closer the *RCV* is to one, the sparser the model is and the more it selects the true variables. The results in Table 1 summarize the performance of the different methods in terms of model complexity. An ideal model selection method would only select the k true features and set the coefficients of the other d-k variables equal to zero.

In most of the scenarios, the *stepCL* and *stepCL2* methods yield a higher *RCV* than the other methods and on average the *stepCL* and *stepCL2* models are sparser than the other methods. In scenarios 2 and 3 the adaptive Lasso yields the higher *RCV*, but the models are not as sparse as the stepwise methods. Although the stepwise methods (*stepCL, stepCL2, stepAdaCL*) result in including the least variables in the model, its discriminative ability in terms of AUC, as shown in Table 2, is comparable with the other methods that tend to select a larger model with more variables.

The *stepCL2* method also has remarkable performance both in terms of sparsity and predictive discrimination. Considering the trade-off between model complexity and performance, the proposed stepwise combined penalty approach achieves a good balance between parsimony, including less variables and maintaining a high predictive accuracy that is comparable with state-of-the-art methods. We should mention that in scenarios 2, 3, and 4 none of the methods select all of the k informative variables, however, for the stepwise method the AUC of the ROC curve on the testing set is greater than 90%, indicating a good discrimination accuracy by including the least variables in the model. In all scenarios, we found that the *stepCL2* method has comparable performance to *stepCL* and is superior to adaptive Lasso and Lasso.

In Table 2, we present results regarding the predictive classification accuracy of the methods by the AUC of the ROC curves. We report the Brier score (Brier, 1950) as a measure of the accuracy of predictions, defined as

$$Brier = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{p}_i - y_i\right)^2$$

It is given by the squared distance between the patients observed status y_i and the predicted probability $\widehat{p_i}$. The decision space for the Brier score is the interval [0,1] and generally the lowest the Brier score, the better the classification rule. If the predicted probability is 0.5 for each individual, the Brier score of 0.25 would indicate that the classification rule is a random one.

Empirical results from our simulations show that even for settings where $n > d$, the stepwise method gives the sparsest solutions while maintaining classification performance measures as good as Lasso and adaptive Lasso. The *CL2* method tends to select big models, due to the L_2 norm which shrinks coefficients toward zero without variable selection. Thus when a is close to 1, the model will behave similar to ridge regression and the resulting model will be complex in terms of the number of predictors. On the other hand, when a is closer to 0, the *CL2* and *stepCL2* methods will borrow more of the characteristics of the L_0 norm and will result in sparser models.

In our simulations, we also considered the case where there is no correlation among predictors (results not shown in the table as we do not consider it a realistic scenario). We repeated scenario 2 with the only alteration of setting $\rho = 0$. Results were in the same direction as in scenario 2 shown in Table 1 and Table 2. That is, the stepwise methods perform better than all the other methods in terms of model complexity resulting in the sparsest models with a high classification performance.

Furthermore, we examined the situation where there are no predictors in the data associated with the outcome. In that case that the true model is the null model, none of the methods identified the true model. Again, running through the second scenario with $d = 150$ predictors with none being informative for the outcome, the *stepCL* method selected a median of five variables, whereas the other methods selected between 14 (*AdaLasso, CL*) and 19 (*CL2*). We observed the same pattern in the results for repeating the first scenario with $d = 50$ uninformative predictors, where none of the methods selected the true model but the stepwise methods produced the sparsest solutions.

Application- real data analysis

To illustrate the applicability of the proposed method, we applied the stepwise method on a real example involving protein measurements. The dataset contained $n = 53$ patients with baseline measurements of $d = 187$ proteins. To maintain confidentiality, the names of the proteins are not revealed. For the presentation of the results and keeping the study anonymized we renamed the proteins to $X_1, X_2, \ldots, X_{187}$. The objective is to extract potential candidate markers discriminating responders from non-responders based on patients' protein levels. We apply our proposed stepwise combined penalty approach with the aim to select a small set of proteins that can sufficiently predict response to the treatment. We compare our approach with the commonly used Lasso and adaptive Lasso, but also with the global optimization penalized methods *CL* and *CL2*.

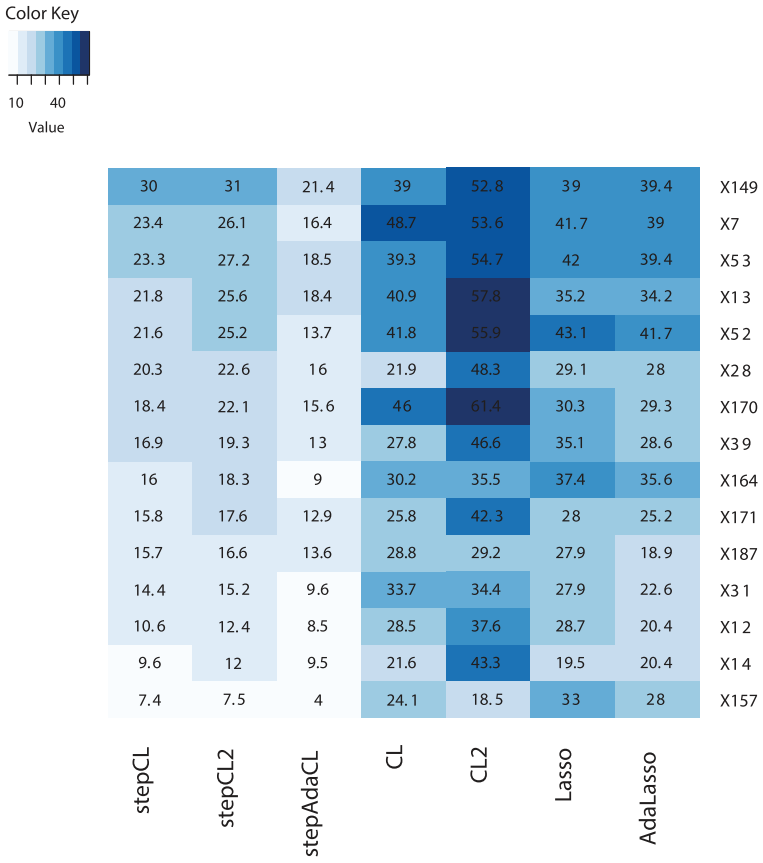

Figure 3. Heatmap of the frequencies (%) of the top 10 variables selected over B=1000 bootstrapped datasets of the protein data. The first column (stepCL) is ordered by decreasing frequencies of the top 10 variables selected by all methods. In dark colour are the higher frequencies (>30%) and the lighter colours depict the lower frequencies.

The regularization parameters were not tuned by cross-validation, due to the relatively small sample size ($n = 53$). The tuning was done using the bootstrap method in the following way; for a grid of values of (a, λ), we trained the models on $B = 100$ bootstrapped datasets (drawing samples with replacement from the original data) and evaluate their classification performance (in terms of AUC) on the original data. For each combination of the tuning parameters (a, λ), the models were trained on B bootstrapped sets and validated on the testing set (original data) and the average AUC (over the B bootstrapped samples) was reported together with its standard deviation. The configuration of (a, λ) that corresponds to the median of the AUC in interval A, as described above in Section 3.2 'Tuning of parameters', was chosen.

The results show that the stepwise methods yield the sparsest models by selecting eight variables (*stepCL)* and nine (*stepCL2*) accordingly, whereas the other methods select between 22 (*CL2*) and 26 (*Lasso*). It is noticeable that the classification performance of the stepwise method is as good as the other variable selection methods, even including the least predictors. In order to evaluate the performance of the models and in the absence of an external validation dataset we use bootstrapping. We applied all the methods on another $B = 1000$ bootstrapped datasets of the protein data, by sampling with replacement, and the frequencies of the top 10 selected variables by all methods are reported in Figure 3. This resulted in 16 unique proteins.

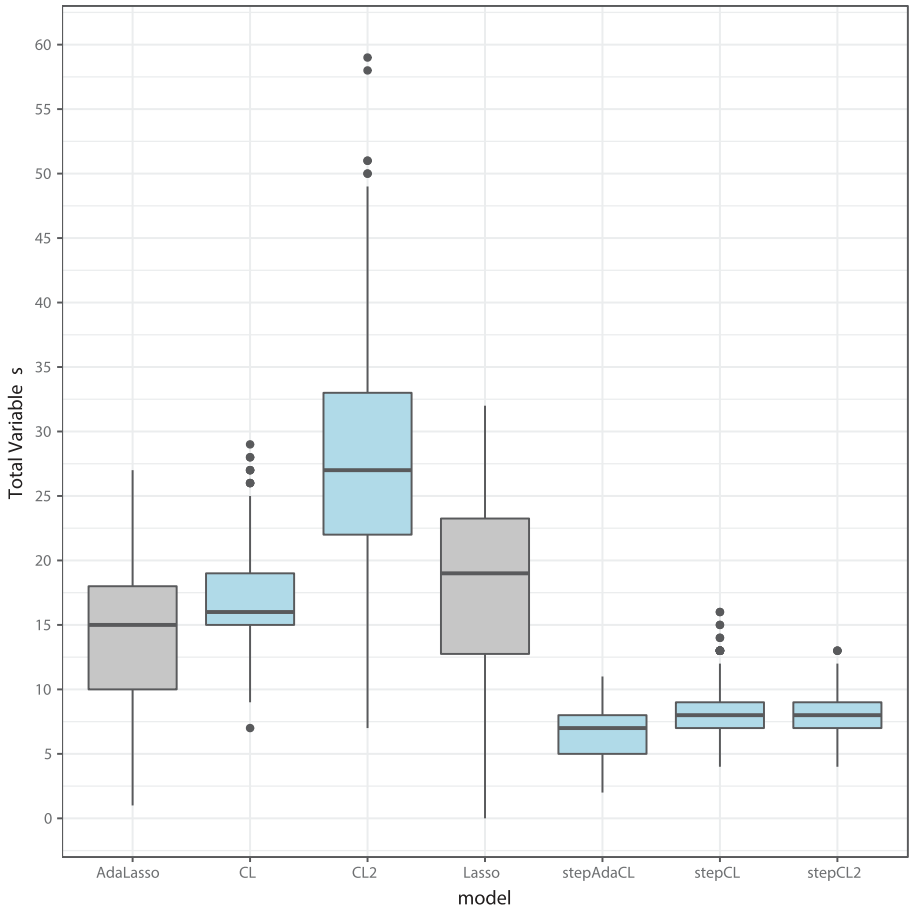

Figure 4. Boxplots of the total variables selected by all methods over the B=1000 bootstrapped datasets of the protein data.

This figure shows that the proteins that were frequently selected by the *stepCL* and *stepCL2* methods are also frequently selected by the Lasso and adaptive Lasso. Note that the stepwise methods have lower frequencies of the selected variables, because selection of larger models will automatically increase the number of selection for individual variables.

Figure 4 shows boxplots of the total number of variables included in the model over the bootstrap evaluations. The stepwise method yields consistent model selection by selecting a median of eight variables for *stepCL* and *stepCL2*, whereas the Lasso and adaptive Lasso have a big variability on the complexity of the model selected. The AUC of the ROC curves that is used as a measure of classification performance of the methods on the bootstrapped datasets and their distribution is shown in Figure 5. The stepwise methods tend to always select the most sparse models more systematically, while maintaining a very good classification performance.

Conclusion

In this paper, we have proposed a stepwise forward approach for model selection in the framework of penalized regression using a penalty that combines the L_0 norm, which is based on the number of coefficients, with L_1 norm which is based on the size of coefficients or L_2 norm which take into

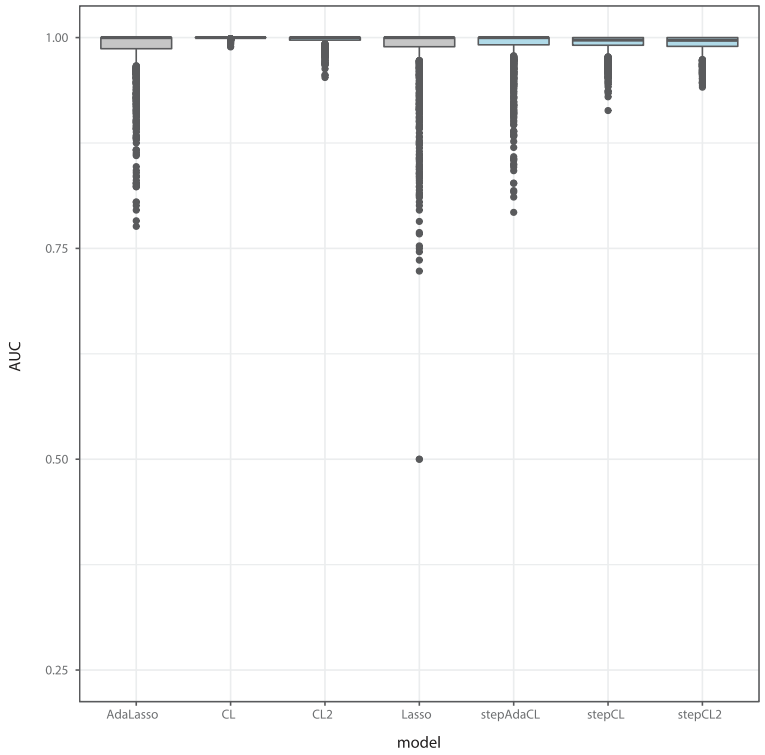

Figure 5. Boxplots of the AUC of the ROC curves by all methods over the B=1000 bootstrapped datasets of the protein data.

account the grouping effect. The aim of the proposed method is to find a model that includes as few and relevant variables as possible on one hand, while maintain good predictive performance on the other hand. The combined penalization term $CL_\alpha{}^\varepsilon(\beta)$ that was introduced by Liu and Wu (Liu & Wu, 2007) was limited to moderate datasets due to limitations of the optimization algorithm. Considering the heuristic stepwise forward approach, we can apply the penalization $CL_\alpha(\beta)$ and $CL2_\alpha(\beta)$ to high-dimensional data by using the BFGS algorithm which is found to work well in practice for non-convex and non-smooth functions (Lewis and Overton, 2009 ; Lewis and Overton, 2013; Curtis and Que, 2015). As a result, the practical implementation of the stepwise penalization method is simpler and more efficient.

We find that the combined penalty does achieve the goal of sparser selection compared to the Lasso and adaptive Lasso while at the same time retaining the same or very similar predictive performance. While the stepwise method is only an approximation to the true optimal solutions it appears to approximate the true optimal solution as well or on occasion even better than the global optimization routine and reduces the computational time considerably. Tuning of the regularization parameters $(a,\ \lambda)$ can, however, take a few minutes. Tuning is an important aspect of penalization methods and will be further explored and improved in future work.

Simulation results and a real data application show that the proposed method yields sparser models, while maintaining a good classification performance. This is an essential consideration for classification and screening applications where the goal is to develop a test using as few features as possible to enhance the interpretability and, potentially, the reproducibility of the results, as well as to control the cost of the implementation of the test. Overall, we found that our method provides a sparser model while maintaining similar prediction properties as compared to other methods. We

hope that this paper could be a first step to learn more about the theoretical properties of this method, which seems to be worth further investigation.

Furthermore, it would be of great interest to extend the forward stepwise method to the stepwise bidirectional approach, considering at each step of the algorithm which variables can be included and excluded (forward and backwards variable selection) in the model. As future work we also consider to apply our method to regression problems for variable selection with a continuous response as well as time-to-event data.

Declaration of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 633567 and is part of the IDEAS European training network (***http://www.ideas-itn.eu/***). This report is in part independent research arising from Prof Jaki's Senior Research Fellowship (NIHR-SRF-2015-08-001) supported by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

ORCID

Eleni Vradi http://orcid.org/0000-0003-0330-3309
Thomas Jaki http://orcid.org/0000-0002-1096-188X

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19* (6):716–723.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Mon Wea Reviews 78*:1–3.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms. *IMA Journal of Applied Mathematics 6*(1):76–90.

Chen, J., and Chen, Z., (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, pp.759–771.

Curtis, F. E., Que, X., (2015). A quasi-Newton algorithm for nonconvex, nonsmooth optimization with global convergence guarantees. *Mathematical Programming Computation*, 7(4):399–428.

Fan, J., Li, R. (2001). Varaiable selection via nonconcave penalized likelihood and its Oracle properties. *Journal of the American Statistical Association* 96:1348–1360.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal 13*(3):317–322.

Frank, L. E., Friedman, J. H., 1993. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Frommlet, F., Nuel, G. (2016). An adaptive ridge procedure for L0 regularization. *PLoS ONE 11*(2):e0148620. doi:10.1371/journal.pone.0148620

Furnival, G., Wilson, R. (1974). Regression by leaps and bounds. *Technometrics 16*(4):499–511.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation 24*(109):23–26.

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

Huang, -H.-H., Liu, X.-Y., Liang, Y. (2016). Feature selection and cancer classification via sparse logistic regression with the hybrid L1/1+L2 regularization. *PLoS One 11*(5):e0149675.

Lewis, A. S., Michael L. Overton, (2009). Nonsmooth optimization via BFGS. Submitted to *SIAM J. Optimiz.* : 1–35.

Lewis, A. S., Overton, M. L. (2013). Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming 141*(1–2):135–163.

Liu, Y., Wu, Y. (2007). Variable selection via a Combination of the L0 and L1 penalties. *Journal of Computational and Graphical Statistics 16*:782–798.

Meinshausen, N., Yu, B. (2009). Lasso type recovery of sparse representations for high-dimensional data. *Annals Statistical JSTOR* 37:246–270.
R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics* 6(2):461–464.
Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation 24*(111):647–656.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *JR Statist Social B*. 58:267–288.
Zhang, Y., and Shen, X., (2010). Model selection procedure for high-dimensional data. *Statistical analysis and data mining*, 3(5)350-358.
Zhao, P., Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research 7*:2541–2563.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *Taylor & Francis 101*:1418–1429.
Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2):301–320.

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–19

DOI: 10.1177/0962280218784778
journals.sagepub.com/home/smm
SAGE

A Bayesian model to estimate the cutoff and the clinical utility of a biomarker assay

Eleni Vradi,[1,2] Thomas Jaki,[3] Richardus Vonk[1] and Werner Brannath[2]

Abstract

To enable targeted therapies and enhance medical decision-making, biomarkers are increasingly used as screening and diagnostic tests. When using quantitative biomarkers for classification purposes, this often implies that an appropriate cutoff for the biomarker has to be determined and its clinical utility must be assessed. In the context of drug development, it is of interest how the probability of response changes with increasing values of the biomarker. Unlike sensitivity and specificity, predictive values are functions of the accuracy of the test, depend on the prevalence of the disease and therefore are a useful tool in this setting. In this paper, we propose a Bayesian method to not only estimate the cutoff value using the negative and positive predictive values, but also estimate the uncertainty around this estimate. Using Bayesian inference allows us to incorporate prior information, and obtain posterior estimates and credible intervals for the cut-off and associated predictive values. The performance of the Bayesian approach is compared with alternative methods via simulation studies of bias, interval coverage and width and illustrations on real data with binary and time-to-event outcomes are provided.

Keywords

Bayesian model, cutoff estimation, predictive values, step function, diagnostic tests, clinical utility, response rates

1 Introduction

The development of diagnostic tests using biomarkers is now an integral part of the drug discovery and development process. Biomarkers are used in enrichment to assist in patient selection and in the design of clinical trials.[1] In the field of oncology, for instance, biomarkers are used to develop tests aiming to identify and treat those who are more likely to respond and demonstrate a higher therapeutic benefit. The adaptation of these biomarkers-based tests for classification purposes requires the assessment of the test performance and, perhaps even more importantly, their clinical utility.

The evaluation of the diagnostic performance of a set of biomarkers is usually performed using Receiver Operating Characteristic (ROC) curves, which plot the true positive rate (sensitivity) versus the false positive rate (1-specificity) over all possible decision thresholds of the test. This is helpful in choosing the most discriminating marker or set of markers.[2] After choosing an accurate marker from a set of markers, an appropriate threshold, or cutoff value, must be determined such that it correctly classifies patients as required.

Several strategies exist for selecting a cutoff value. These may be based on numerical results around the sensitivity and specificity, but may also include criteria based on biological or physiological information. Thus, optimal thresholds may vary depending on the underlying criteria.[3] Most commonly, the optimal cutoff is chosen as the one that optimizes a utility function. For example, the cutoff that maximizes the number of correctly classified patients or the cutoff that minimizes the misclassification cost. Because a utility function also requires

[1]Department of Research and Clinical Sciences Statistics, Bayer AG, Berlin, Germany and Competence Center for Clinical Trials, University of Bremen, Germany
[2]Competence Center for Clinical Trials, University of Bremen, Bremen, Germany
[3]Department of Mathematics and Statistics, Lancaster University, Lancaster, Lancashire LA14YF, UK

Corresponding author:
Eleni Vradi, Bayer Pharma, AG Muellerstrasse, 178 Berlin 13342, Germany.
Email: eleni.vradi@bayer.com

information about cost or benefit, which is not always available, the optimal cutoff value is found by using criteria related to ROC curves. Confidence intervals around the cutoff value are obtained either using the delta method or, most commonly, by employing bootstrapping, though the coverage probabilities can be far from the desired level.[4]

ROC-based methods, however, do not provide information on the diagnostic accuracy for a specific patient. Particularly in situations where a diagnostic test is used for classification purposes, clinicians are mainly concerned with the predictive ability of the test, approaching the result of the test from the direction of the patients. The assessment of correct classifications can be facilitated by the use of positive and negative predictive values (PPV and NPV, respectively). These predictive values are functions of the accuracy of the test and the overall prevalence, and can be used to assess the clinical utility of a diagnostic test for classification purposes.

Lunceford[5] discussed the estimation of the clinical utility of a biomarker assay in the context of predictive enrichment studies. The aim of his research was to select a cutoff on a potentially predictive biomarker that can be used as an enrollment criterion for patient selection. By implementing a Bayesian approach in estimating clinical utility measures, he facilitates cutoff decision-making, but without considering the actual cutoff estimation.

In this paper, we are interested in estimating the cutoff and the clinical utility of a biomarker, but most importantly the uncertainty around the estimates of the parameters of interest. We propose a flexible Bayesian approach that can utilize prior information to estimate the cutoff of a biomarker and its credible interval. By modelling the probability of response with a step function using predictive values, we obtain estimates for the cutoff as well as for the predictive values of the test. Bayesian analysis allows us to assign probability distributions to our prior beliefs for the parameters of interest and combine these with the data likelihood to yield a posterior probability distribution representing our updated belief.

In section 2, we present the Bayesian model for estimating the cutoff of a (continuous or ordinal) biomarker for a binary outcome. The different prior specifications for the cutoff that we consider allow for some robustness of the method. The finite-sample performance of the proposed Bayesian approach is demonstrated through a series of simulations and compared with alternative frequentist methods like Maximum Likelihood approach and the PSI index in Section 3. We also present applications of our method in Section 4 on real data for a continuous biomarker and binary, as well as time-to-event endpoints. Finally, we conclude with a brief discussion.

2 Methods

2.1 Bayesian model for estimating the cutoff and its credible interval

In this section, we present a Bayesian model for estimating the posterior distribution of a cut-off value for a biomarker, as well as its predictive values. Let $X = (X_1, X_2, \ldots, X_n) \in \mathbb{R}$ denote the continuous biomarker measurements for n individuals and assume that X is available to be measured on all patients. Let $Y = (Y_1, Y_2, \ldots, Y_n)$ denote the binary response variable, where $Y_i \in \{0, 1\}$ for all $i = 1, \ldots, n$ is the response indicator (e.g. $Y_i = 0$ denotes the non-responders and $Y_i = 1$ the responder subjects). We do not make assumptions about the distribution of the biomarker X and by convention it will be assumed that high values of the marker X are associated with increased probability of response to a treatment.

We assume that the probability of response p can be modeled by a step function (Figure 1), in terms of positive predictive value (PPV) and negative predictive value (NPV) of the biomarker assay. The PPV is defined as the

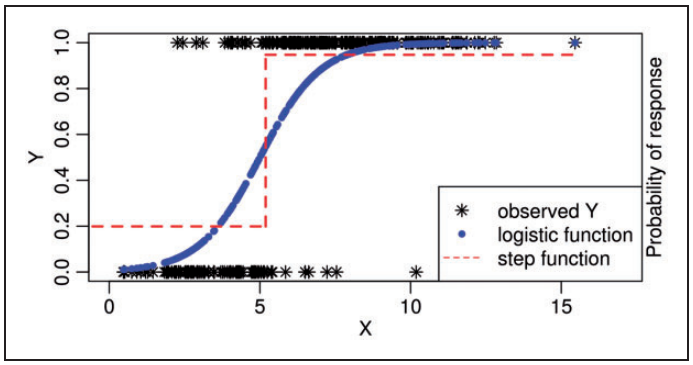

Figure 1. Plot of the observed biomarker X measurements for $y = 1$ and $y = 0$ (black stars). The blue dots depict the probability of response, p, when fitting a logistic model and the dashed red line shows p when it is modeled by a step function, and $p_1 = 0.17$, $p_2 = 0.95$ and $cp = 5.19$ are the posterior means as estimated by the Bayesian model.

conditional probability of response given a positive test result, i.e. $P(y = 1|T^+)$. Conventionally, for potential cutoff $cp \in \mathbb{R}$, the test is positive, T^+, if the biomarker exceeds the cutoff, $X \geq cp$, and is negative otherwise. Similar statements apply for the NPV which is defined as the conditional probability that an individual is a non-responder given a negative test result, i.e. $P(Y = 0|T^-) = P(Y = 0|X \leq cp)$. The model is specified in the following way

$$\begin{aligned} & Y|X \sim Bernoulli(p) \\ & p(x) = P(Y = 1|X = x) = \begin{cases} p_1 = P(Y = 1|X \leq cp), & \text{for } x \leq cp \\ p_2 = P(Y = 1|X > cp), & \text{for } x > cp \end{cases} \end{aligned} \tag{1}$$

The $p_1 = 1 - \text{NPV}$ expresses the probability of response given X is below the cutoff value cp and $p_2 = \text{PPV}$ expresses the probability of response given that X is greater than cp.

Logistic regression can be used for decision-making, i.e. to classify a subject as responder or not, only in conjunction to a probability threshold, i.e. $p = 0.5$.[6] However, the advantage of using the step function is that the cutoff is a parameter of the model and therefore a Bayesian approach can be applied. The strong assumption we make that the probability of response can be modeled by a step function is probably not always reflecting the reality. However, it may serve as an approximating model in cases where there are two populations that have a pronounced difference in the response rate. It follows from literature on misspecified models[7,8] that even if the model is misspecified, the estimates of the assumed step function are consistent for the parameter values for which the assumed model minimizes the distance from the true distribution in terms of Kullback–Leibler (KL) divergence.[9]

2.1.1 Prior specification

In a Bayesian setup, the idea is to represent the uncertainty about the parameters by a prior distribution. Prior information can take into account subjective beliefs about the values of the parameters of the model. This external information can be historical information from experiments, experts opinion or literature findings. A Bayesian approach can thus be useful as it allows flexibility combining the available prior knowledge on test characteristics with new data. Importantly, incorrect prior information can lead to unreliable posterior estimates, and therefore great attention should be paid to the choice of the prior. On the other hand, if good prior information is available then the gain is in the precision of the estimates.

Here, the parameters p_1, p_2 and the cutoff are assumed to have probability distributions reflecting the uncertainty in their parameters values. For the probabilities of response p_1 and p_2, we consider distributions that the support set is the interval (0, 1). Furthermore, we require that $p_2 > p_1$. The simplest case is to assign uniform priors, i.e.

$$p_1 \sim Unif(0, 1) \quad \text{and} \quad p_2 \sim Unif(p_1, 1) \tag{2}$$

Other options may include Truncated Normal or Beta distributions.

For the cutoff cp, we can consider an informative prior, if prior information is relevant and an uninformative prior, when there is no information available, usually expressed by a uniform distribution. Finally, a weighted sum of informative and non-informative priors can be considered to acknowledge potential prior-data conflict. We propose here a two-component mixture of priors, which allow for robustness. The first component of the mixture prior is the informative part which expresses the subjective belief we have and is derived from prior experiments, animal data or literature. Then, the second component is the weakly (or non-) informative part that ensures robustness against potential prior-data conflict. We characterize a prior distribution as weakly informative if the information that provides is intentionally weaker than whatever actual prior knowledge is available.

As discussed by Schmidli et al.,[10] since one of the mixture components is usually vague, mixture priors will often be heavy tailed and therefore robust. Let g_1 be the probability density function (pdf) of the uninformative component and g_2 the pdf for the informative part. The mixture prior can be expressed as

$$cp = w\, g_1 + (1 - w)\, g_2 \tag{3}$$

with

$$w \sim Beta(1, 1)$$

The weight parameter w will be updated at each iteration by the Bayesian model as described in section 3.

2.1.2 Prior specification for constrained PPV

In this section, we present the case where the objective is to estimate a cutoff associated with a targeted clinical utility value by controlling the PPV of the test. For example, we might be interested in the posterior distribution of the cutoff expected to yield a PPV between 70% and 100% or a 1-NPV to be between 0 and 20%. Whether a cutoff that yields a pre-specified predicted value exists would of course depend on the relationship between the biomarker and the response. The idea is then to incorporate the restriction on the predictive values via the prior information and require that only information on the pre-specified domain is acceptable. In that case, the constraints can be controlled through priors, e.g.

$$p_1 \sim Unif(0, p_2) \quad \text{and} \quad p_2 \sim Unif(0.7, 1)$$

It is worth noting that even if the parameter is constrained such that the actual desired range is not achievable, e.g. $p_2 \notin (0.7, 1)$, the method will result in the cutoff value that is as close as possible to achieve this constraint (i.e. the mode of the posterior density is on the lower bound of the constrained interval).

2.1.3 Posterior distribution

The posterior distribution of interest is formulated as

$$f(cp,\ p_1, p_2|x, y) \propto L(p_1, p_2, cp|x, y) \times f(p_1) \times f(p_2) \times f(cp) \tag{4}$$

where $L(p_1, p_2, cp|x, y)$ is the likelihood function of the data and $f(\cdot)$ denotes the density of the prior and $f(\cdot|x, y)$ the posterior density of the distribution of the parameters.

2.1.4 Maximum likelihood estimation

The log likelihood of the model described in section 2.1 is given by

$$\log L = L(p_1, p_2, cp|x, y) = \sum_{i=1}^{n} y_i \log(p) + (1 - y_i) \log(1 - p)$$

with p as stated in equation (1) and n denotes the total sample size. The log likelihood function becomes

$$logL = \sum_{i=1}^{n_1} y_i \log(p_1) + (1 - y_i) \log(1 - p_1) + \sum_{i=1}^{n_2} y_i \log(p_2) + (1 - y_i) \log(1 - p_2)$$

where n_1, n_2 denote the sample size for the population that has $X \leq cp$ and $X > cp$, respectively. The maximum likelihood estimates $\widehat{cp}$, $\widehat{p_1}$ and $\widehat{p_2}$ are obtained by first minimizing $-logL$ with respect to p_1 and p_2, for given cp and then maximizing the resulting profile likelihood with respect to cp. One can see that $\widehat{p_1}$ and $\widehat{p_2}$ are just the average response rates in the subsamples $\{x_i \leq \widehat{cp}\}$ and $\{x_i > \widehat{cp}\}$ where x_i is the observed value of X (see Appendix 3 for a similar argument for the population parameters).

3 Simulation study

In this section, we examine the bias of the estimated cutoff under different distributional assumptions for the biomarker X via simulations. We compared the proposed Bayesian method with two frequentist approaches; the maximum likelihood estimator (MLE) and the predictive summary index (PSI).[11] The PSI estimates the optimal cutoff by maximizing the difference in predictive values for all possible cutoffs c and is expressed as $PSI = \max_c \{PPV(c) + NPV(c) - 1\}$. The PSI is derived in the target (patient) population as a measure of the goodness of the predictability in a diagnostic test, and thus is a more comprehensive measure than the Youden index[12] in a clinical setting. For the latter approach, the confidence intervals are calculated by the bootstrap

Table 1. Six simulation scenarios assuming different distributions for the marker X, the true cp, p_1 and p_2, as well as different generating models, a step function and a logistic function.

Scenarios	Distribution of X	μ_1	σ_1^2	μ_2	σ_2^2	True cp	True p_1	True p_2	Generating Model
1	Normal	$\mu = 7$,	$\sigma^2 = 1$			7.30	0.10	0.80	Step function
2	Mixture Normal (unequal variances)	6.5	0.09	8	0.25	7.30	0.20	0.90	Step function
3	lognormal	$\mu = 0$,	$\sigma^2 = 1$			2	0.30	0.85	Step function
4	Ordinal with 4 levels	X = 1, 2, 3, 4				3	0.10	0.75	Step function
5	Normal	$\mu = 7$, $\beta_0 = -3$,	$\sigma^2 = 2$, $\beta_1 = 0.5$			(6.80)	(0.41)	(0.76)	Logistic function
6	Normal	5	1	9	1	$cp_1 = 6$ $cp_2 = 10$	$p_1 = 0.20$, $p_2 = 0.60$, $p_3 = 0.80$		Step function

Note: For the latter generating model, the true parameters that are in parenthesis are the population parameters as calculated by minimizing the Kullback–Leibler divergence.

method by resampling the data $B = 500$ times, calculating the $\widehat{PSI_j}$ per sample $j = 1, \ldots, B$ and then taking $\alpha/2$ and $1 - \alpha/2$ quantiles of the $\widehat{PSI_j}$ to construct a $(1 - \alpha)$ 100% CI. For the Bayesian approach, the credible intervals are obtained by using the empirical $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution (quantile method). A level of $\alpha = 0.05$ was used for both methods.

We include in our results the MLE of the parameters p_1, p_2, cp together with the 95% confidence intervals (CI) as a comparison. In general, maximum likelihood methods do not perform well when parameter estimates are on the boundary of the parameter space,[13] leading to some non-convergence issues. On the other hand, Bayesian inference via MCMC algorithms permits full posterior inference even in the absence of asymptotic normality[14] and have no issues with parameter estimates on the boundary. In our simulation, we did not anticipate any optimization issues regarding the optimization with the ML method.

We simulated 10,000 datasets on which we applied all methods. We also report the coverage probability and the width of the credible and confidence intervals over the simulation runs. The analysis for the MLE and PSI estimation was done in R version 3.3.3.[15] The 10,000 datasets were generated in R (for the MLE and PSI estimation) and then exported to SAS version 9.4 (SAS Institute Inc., Cary, NC, USA) (for the Bayesian estimation), such that the analysis was consistent for all the methods. For the PSI method the R-package "OptimalCutpoints"[16] was used and for the profile MLE the R-library "bbmle".[17]

The posterior computation was done by using Markov Chain Monte Carlo (MCMC). In our analysis we used the Metropolis-Hastings[18,19] iterative sampling method to approximate the posterior distribution and get posterior estimates for the parameters in equation (4). Posterior computation was conducted using PROC MCMC procedure in SAS. The burn-in consisted of 10,000 iterations, and 50,000 subsequent iterations were used for posterior summaries. Convergence of the MCMC chain was checked for randomly selected number of iterations, using diagnostic plots and the Gelman–Rubin convergence statistic as well as visually via trace plots, sample autocorrelations and kernel density plots. The SAS and R code can be found in Appendix 1.

3.1 Simulation setting

3.1.1 Generating data using a step function and a logistic function

The true model that was used to generate the binary outcome y has one biomarker X. We consider six different simulation scenarios, each with $n = 200$, and $n = 50$. Furthermore, we assumed that the biomarker X follows different distributions as shown in Table 1. Each component of the response vector y is viewed as a realization of a Bernoulli random variable with probability of success p, i.e. $y|X \sim Bermoulli(p)$. In scenarios 1–4 and 6, the generating model has response probability p expressed as a step function, with $p(X) = \begin{cases} p_1, & \text{if } X \leq cp \\ p_2, & \text{if } X > cp \end{cases}$, whereas in scenario 5 the generating model is a logistic model with probability of response $p = \frac{e^{X\beta}}{1+e^{X\beta}}$.

Table 2. Mean bias of the estimate of the **cutoff** $\widehat{cp}$ over 10,000 simulation runs for the Bayesian method, the MLE and PSI approach for scenarios 1–4 for $n = 50$ and $n = 200$.

cp		Bias						
Methods		Bayesian					PSI	MLE
Prior		UP	IPN	IPP	MixN	MixP		
Scenario 1	$n = 200$	3×10^{-4}	-1×10^{-3}	1×10^{-4}	2×10^{-4}	8×10^{-4}	0.288	-4×10^{-3}
	$n = 50$	4×10^{-2}	-5×10^{-2}	-2×10^{-2}	1×10^{-3}	6×10^{-3}	0.393	9×10^{-2}
Scenario 2	$n = 200$	-7×10^{-3}	-1×10^{-2}	-2×10^{-2}	-1×10^{-2}	-9×10^{-3}	1×10^{-2}	-2×10^{-2}
	$n = 50$	-1×10^{-2}	-1×10^{-1}	8×10^{-2}	-3×10^{-2}	-3×10^{-2}	-6×10^{-4}	0.173
Scenario 3	$n = 200$	1×10^{-2}	-4×10^{-2}	-5×10^{-4}	-4×10^{-2}	2×10^{-3}	3.447	-3×10^{-3}
	$n = 50$	2×10^{-1}	-4×10^{-1}	4×10^{-4}	-2×10^{-2}	9×10^{-2}	1.449	0.365
Scenario 4	$n = 200$	-2×10^{-2}	2×10^{-2}	4×10^{-4}	5×10^{-3}	4×10^{-4}	2×10^{-4}	-2×10^{-3}
	$n = 50$	-8×10^{-3}	4×10^{-2}	7×10^{-3}	2×10^{-2}	7×10^{-3}	3×10^{-2}	0.996

The primary purpose of including scenario 5 is to investigate the behavior of the Bayesian method (together with the MLE and the PSI method), when the fitted model is divergent from the true underlying model. For this scenario, the true cp, p_1 and p_2 are not defined by the data generating mechanism. In fact, it is known (see e.g.[7,8]) that the estimated parameters from the Bayesian and MLE method are consistent for the ones that minimize the Kullback–Leibler divergence between the fitted (step) model and the true (logistic) model. We give details on the limiting population parameter in Appendix 1.

In scenario 4, we explore the case that the biomarker X is ordinal. The data were generated in the following way; assuming $X \sim Normal(\mu = 7, \sigma^2 = 1)$ as in scenario 1, we calculate the quartiles of X that form the four levels of the ordinal variable (the lowest quartile corresponds to category $X = 1$ and the fourth quartile to $X = 4$). Each component of the response Y is a realization from a Bernoulli random variable with $p(X) = \begin{cases} p_1, & \text{if } X = 1, 2 \\ p_2, & \text{if } X \geq 3 \end{cases}$

Moreover, we are interested to address the case that the true generating model has two cutoffs and the fitted model assumes only one cutoff (scenario 6 in Table 1). To simulate data for this scenario, scenario 6, we assumed that $p(X) = \begin{cases} p_1, & \text{if } X \leq cp_1 \\ p_2, & \text{if } cp_1 < X \leq cp_2 \\ p_3, & \text{if } X > cp_2 \end{cases}$. If the data indicate the existence of two cut-off values, this might indicate the existence of two subgroups with different response probabilities. For the scenarios 2 and 6, we assumed that the biomarker X follows a mixture of two normal distributions expressed as $X \sim Normal(\mu = \mu_1, \sigma^2 = \sigma_1^2) + Normal(\mu = \mu_2, \sigma^2 = \sigma_2^2)$.

3.2 Simulation results

This section describes the simulation results regarding the finite sample properties of the estimators from the Bayesian method, the PSI index and the ML. In our results, we chose to report the Bayesian posterior mean, as we consider it an adequate measure to summarize the posterior density and we found that the cutoffs were generally similar whatever estimate kept from the posterior distribution among the mode, median or mean. In Tables 2 and 3, we report the Bias of estimators for cp (Table 2), p_1, p_2 (Table 3) for scenarios 1–4 based on 10,000 simulation runs. Coverage probability and interval width of the confidence and credible intervals are shown in Tables 4 and 5.

For the Bayesian method, we also report results for four different prior specifications. The first, the naïve case, corresponds to a uniform prior (UP) in the interval of the range of the biomarker measurements. Note here that with a uniform prior, it is well known[20] that the Bayesian posterior mode corresponds to the ML estimator. Other priors we considered are a perfect informative prior (denoted as IPP), an imperfect informative prior (denoted as IPN) and two mixture priors (MixP and MixN) each with two components; a weighted sum of a uniform and informative prior (UP + IPP) and a uniform and imperfect informative prior (UP + IPN), respectively. More specifically, for the IPP prior, we assume a distribution for which the true cutoff lies in an interval of high probability, whereas for the IPN prior the true cutoff lies in one of the tails of the distribution. An illustration of the IPP and IPN priors used for scenario 1 can be found in Figure 2. Obviously, when the prior does not include

Table 3. Mean bias of the estimates of predictive values $\widehat{p_1}$ and $\widehat{p_2}$ over 10,000 simulation runs for the Bayesian method, the MLE and PSI approach for scenarios $1-4$ and for $n=200$.

p_1, p_2	Bias						
Methods	Bayesian					PSI	MLE
Prior	UP	IPN	IPP	MixN	MixP		
Scenario 1							
p_1	7×10^{-3}	7×10^{-3}	7×10^{-3}	7×10^{-3}	7×10^{-3}	3×10^{-2}	-1×10^{-4}
p_2	-8×10^{-3}	-8×10^{-3}	-8×10^{-3}	-8×10^{-3}	-8×10^{-3}	4×10^{-2}	4×10^{-4}
Scenario 2							
p_1	7×10^{-3}	7×10^{-3}	6×10^{-3}	7×10^{-3}	7×10^{-3}	-3×10^{-3}	-1×10^{-3}
p_2	-9×10^{-3}	-1×10^{-2}	-1×10^{-2}	-1×10^{-2}	-9×10^{-3}	8×10^{-4}	2×10^{-4}
Scenario 3							
p_1	5×10^{-3}	2×10^{-3}	4×10^{-3}	4×10^{-3}	4×10^{-3}	5×10^{-2}	-1×10^{-3}
p_2	-1×10^{-2}	-2×10^{-2}	-2×10^{-2}	-2×10^{-2}	-1×10^{-2}	9×10^{-3}	3×10^{-2}
Scenario 4							
p_1	1×10^{-2}	1×10^{-2}	9×10^{-3}	1×10^{-2}	9×10^{-3}	2×10^{-4}	9×10^{-2}
p_2	-8×10^{-3}	-5×10^{-3}	-5×10^{-3}	-5×10^{-3}	-5×10^{-3}	-4×10^{-3}	-5×10^{-2}

Note: For the Bayesian method, we display the results for all different prior specifications.

Table 4. Mean coverage and width of the credible/confidence intervals of $\widehat{cp}$ over 10,000 simulation runs for scenarios 1–4 for $n=50$ and $n=200$.

cp		Coverage							Interval width						
Methods		Bayesian					PSI	MLE	Bayesian					PSI	MLE
Prior		UP	IPN	IPP	MixN	MixP			UP	IPN	IPP	MixN	MixP		
Scenario 1	$n=200$	0.968	0.969	0.969	0.950	0.969	0.677	0.919	0.088	0.088	0.088	0.088	0.088	1.547	0.085
	$n=50$	0.972	0.971	0.979	0.971	0.970	0.588	0.722	0.892	0.628	0.353	0.656	0.553	1.522	0.174
Scenario 2	$n=200$	0.962	0.962	0.962	0.964	0.967	0.858	0.797	0.183	0.184	0.179	0.185	0.178	0.649	0.136
	$n=50$	0.979	0.964	0.969	0.976	0.977	0.901	0.467	0.832	0.787	0.514	0.684	0.619	1.534	0.216
Scenario 3	$n=200$	0.959	0.955	0.997	0.939	0.995	0.782	0.486	0.431	0.382	0.138	0.407	0.205	8.669	0.131
	$n=50$	0.980	0.889	100	0.979	0.985	0.905	0.188	2.448	1.321	0.178	1.803	1.464	5.410	0.642
Scenario 4	$n=200$	0.984	0.976	0.999	0.995	0.999	100	0.993	4×10^{-4}	0	0	0.005	0	0.083	0.042
	$n=50$	0.967	0.948	0.989	0.992	0.998	0.999	0.002	0.035	0.018	0.030	0.184	0.105	0.967	0.039

Note: The credible intervals for all the different priors are computed by the quantile method. Bootstrapping was used to calculate the confidence interval for the PSI method and the profile CI are presented for the MLE method.

the true value of the cutoff, then the posterior estimates are expected to be biased for finite sample sizes. The priors for p_1, p_2 were taken as uniform distributions as given by equation (2).

Regarding the estimation of the cutoff cp, in scenarios 1–4, results in Table 2 show that estimators using all three methods behave similarly in terms of bias, resulting in nearly unbiased estimators. The Bayesian method gives a much better coverage than the MLE and PSI methods for the scenarios where the marker is continuous (Table 4). For the PSI method in scenarios 1 and 3, the bias of the estimate of cp is far too high in absolute terms (see Table 2). Additionally, the coverage of the bootstrapped confidence interval is far from the nominal level and the interval width is much wider compared to the other methods. The Bayesian method performs either the same or better compared to MLE and PSI in terms of bias and coverage both in case of the continuous and the ordinal biomarker.

For all priors that we considered, the resulting estimators are on average unbiased for both $n = 200$ and $n = 50$. As expected, with the robust mixture prior and the informative prior, estimates have the smallest bias on average. The IPP prior gives a smaller interval width with the mixture prior second. Moreover, with the IPP prior we get more precise estimates while obtaining the same or better coverage compared to the other prior specifications.

Table 5. Average coverage and width of the credible/confidence interval for the estimate of the predictive values p_1 and p_2 over 10,000 simulation runs for scenarios 1–4 and for $n = 200$.

p_1, p_2	Coverage							Interval width						
Methods	Bayesian					PSI	MLE	Bayesian					PSI	MLE
Prior	UP	IPN	IPP	MixN	MixP			UP	IPN	IPP	MixN	MixP		
Scenario 1														
p_1	0.949	0.949	0.951	0.942	0.949	0.972	0.932	0.107	0.107	0.107	0.106	0.106	0.247	0.106
p_2	0.949	0.946	0.949	0.939	0.943	0.879	0.946	0.177	0.178	0.178	0.175	0.178	0.233	0.182
Scenario 2														
p_1	0.946	0.946	0.948	0.945	0.944	0.959	0.938	0.151	0.150	0.150	0.151	0.150	0.192	0.151
p_2	0.949	0.948	0.949	0.949	0.948	0.959	0.979	0.123	0.124	0.124	0.123	0.123	0.178	0.134
Scenario 3														
p_1	0.949	0.951	0.949	0.949	0.949	0.985	0.936	0.147	0.146	0.144	0.146	0.144	0.283	0.145
p_2	0.955	0.941	0.954	0.953	0.956	0.558	0.980	0.206	0.211	0.197	0.206	0.199	0.244	0.204
Scenario 4														
p_1	0.949	0.927	0.948	0.946	0.948	0.938	0.994	0.120	0.118	0.117	0.118	0.118	0.122	0.345
p_2	0.937	0.950	0.951	0.949	0.951	0.955	0.991	0.165	0.167	0.165	0.166	0.165	0.172	0.191

Note: The credible intervals for all the different priors are computed by the quantile method. Bootstrapping was used to calculate the confidence interval for the PSI method and the CI are presented for the MLE method.

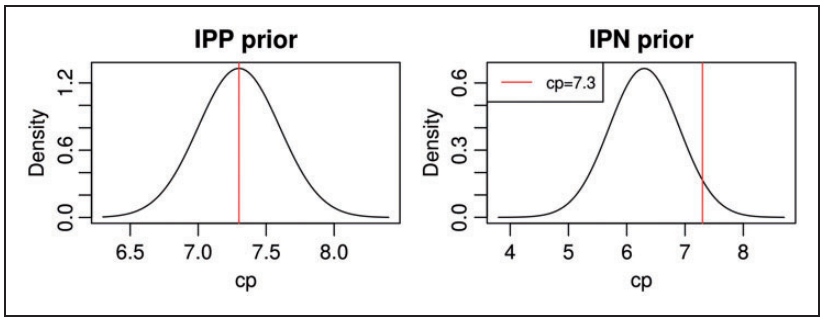

Figure 2. Density plots for the priors IPP and IPN. For the IPP prior, the true cutoff *cp* lies in a high probability region, while for the IPN prior, the true cutoff value lies on the tail of the distribution.

To see how the prior affects the estimation, we calculate the absolute difference between the estimated and true value of the cutoff over the simulation runs and we present the results for the Bayesian method for scenario 1 for all different prior specifications as shown in Figure 8 in Appendix 1. In Figure 8, we see that the absolute difference between the estimate and the true value of *cp* was on average below 10%. As for the predictive values, we discuss our findings for $n = 200$ and show the results for the estimate of the cutoff. Detailed figures for the predictive values for $n = 50$ can be found in Tables 6 and 7 in Appendix 1.

As shown in Tables 3 and 5, all methods performed well with good coverage and very small bias for both p_1 and p_2. The bias of the estimates for the predictive values p_1 and p_2 was always below 1% for all scenarios. Coverage probabilities for the credible intervals reach the nominal value for the Bayesian and the ML method but is not always the case for the estimation of p_2 when using the PSI index as seen, for example, in scenario 1 and scenario 3, where the coverage probability for the PSI method is far from the nominal (Table 5). The length of the credible interval (for the Bayesian method) was similar to the confidence interval for the MLE and always narrower compared to PSI.

For scenario 5 where the true model is generated assuming a logistic response curve, we estimated the cutoff and the corresponding probabilities of response by applying the Bayesian method as well as the MLE and the PSI approaches. In that case, the true cutoff is not directly defined by the data-generating mechanism. However, the population parameters are defined by minimizing the KL divergence between the true (logistic) and the assumed

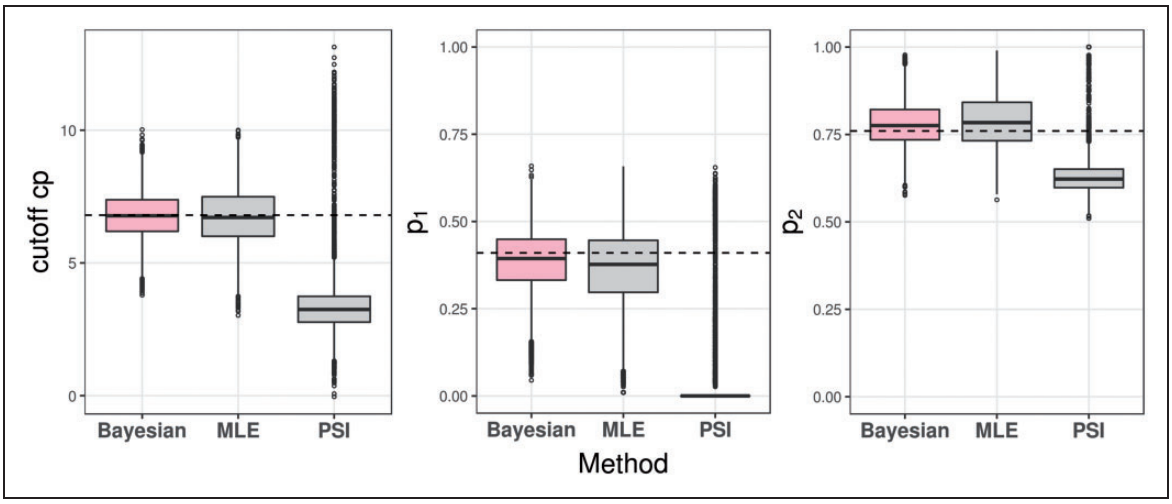

Figure 3. Bayesian posterior mean (left boxplots), MLE (middle boxplots) and PSI (right boxplots) estimators for the parameters cp (left panel), p_1 (middle panel), p_2 (right panel), over 10,000 simulation runs for scenario 5. The black horizontal dashed lines are the population parameters as calculated by minimizing the Kullback–Liebler divergence.

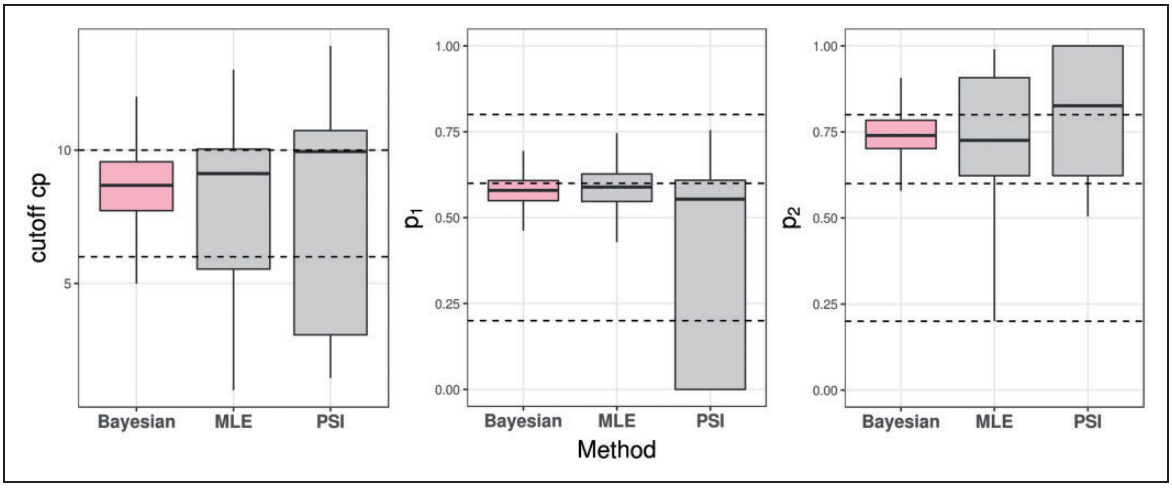

Figure 4. Boxplots of the Bayesian posterior mean (left boxplots), MLE (middle boxplots) and PSI (right boxplots) estimators for cp (left panel), p_1 (middle panel), p_2 (right panel), over 10,000 simulation runs for Scenario 6. The black horizontal dashed lines correspond to the true values of cp_1, cp_2, p_1, p_2, p_3.

(step) model as discussed in section 2.1 and more detailed in Appendix 1. The results of the distribution of the estimates of the parameters for scenario 5 for the three methods are shown in boxplots in Figure 3.

In this scenario, the Bayesian estimates are more consistent and have a smaller variability compared to the MLE and the PSI method. As can be seen from the boxplots, the ML and the PSI methods result in heavy tailed distributions for all the parameters and especially for the estimate of the cutoff. The estimates concerning the cutoff and the predicted values obtained with the PSI method, differ significantly as compared to the other two methods. This is partially due to the fact that the PSI optimizes a different utility function than the Bayesian and the ML approach. While the Bayesian and the ML methods use the likelihood as an objective function, the PSI method seeks to maximize the difference between PPV and 1-NPV.

For scenario 6, the generating model assumes that there exist two cutoff values and three response probabilities p_1, p_2, p_3 respectively. The Bayesian model we fit to estimate the cutoff and the corresponding predictive values, assumes that there is only one cutoff value. For simplicity, we used an UP prior for the Bayesian method. The results of the fitted model are shown in Figure 4. Focusing on the estimate of cp, we analyzed the results in more

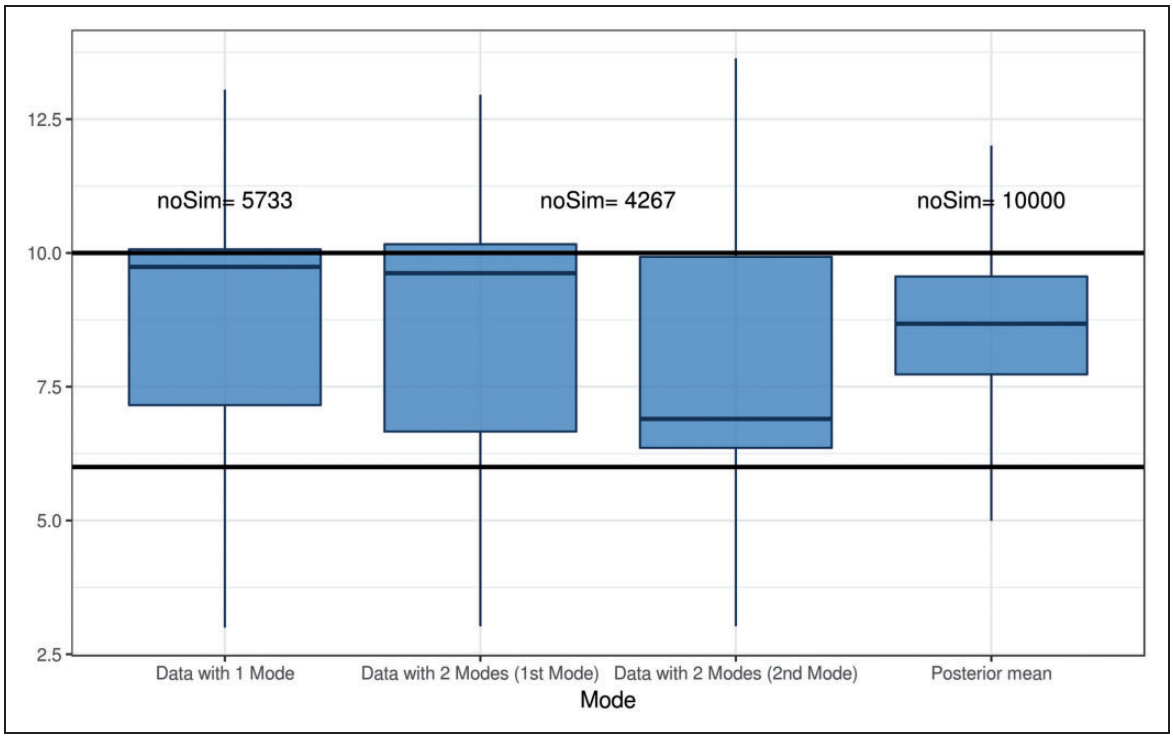

Figure 5. Distribution of the modes of the posterior distribution for the estimated cp, over 10,000 simulation runs for Scenario 6 estimated by the Bayesian model. If the posterior density is unimodal, then the only mode of the distribution is plotted (noSim = 5733) (left boxplot). In case the posterior distribution is bimodal (noSim = 4267), then the two modes are plotted (middle boxplots). In the right boxplot, the black lines correspond to the true values of $cp_1 = 6$, $cp_2 = 10$.

detail. We checked whether the obtained posterior distribution was bimodal, and if so, we reported the two modes. To check for bimodality, i.e. if the posterior density function has two peaks, we used the Hartigan's dip test for unimodality.[21] A p-value less than 0.05 is taken to indicate non-unimodality (it means at least bimodality).

Figure 5 shows the distribution of the estimated cutoffs when posterior density is judged to be unimodal (5733 out of 10,000 simulations) and when it is found to be a bimodal posterior distribution (4267 out of 10,000 simulations). Looking across all simulations, we see that the cutoff is somewhere between the two true cutoffs. When only a single mode is identified, there is a clear tendency to be close to the second true cutoff $cp_2 = 10$. When two modes are found, the underlying two true cutoffs are estimated reasonably well despite the model misspecification.

4 Application

4.1 The prostate cancer data

We consider the prostate specific antigen (PSA) study of 12,000 men aged 50–65, which was a randomized study with a beta-carotene group as the treatment group vs. a placebo group. A substudy reported by Etzioni et al.[22] analyzed serum levels of total PSA (on the log scale) for 683 subjects. The dataset is described in literature[2,23] where you can find additional details about the study, which was analyzed from a non-Bayesian perspective. The primary scientific question under investigation was whether PSA could be used to diagnose prostate cancer, and was found that the total PSA is a significant predictor of the occurrence of cancer with fairly good accuracy. Albeit the good diagnostic ability of the marker PSA, we are interested in estimating a cutoff that takes into account the clinical benefit of this marker.

In this paper, we considered response to a treatment as the outcome of interest but the method can be used also when we refer to diagnostic tests, where the outcome is presence of disease or not. We analyzed the data described above by applying our Bayesian method to estimate the cutoff related with disease rates. Probabilistic statements are derived for the optimal cutoff as well as the predictive values of the marker (logPSA). We assume a uniform

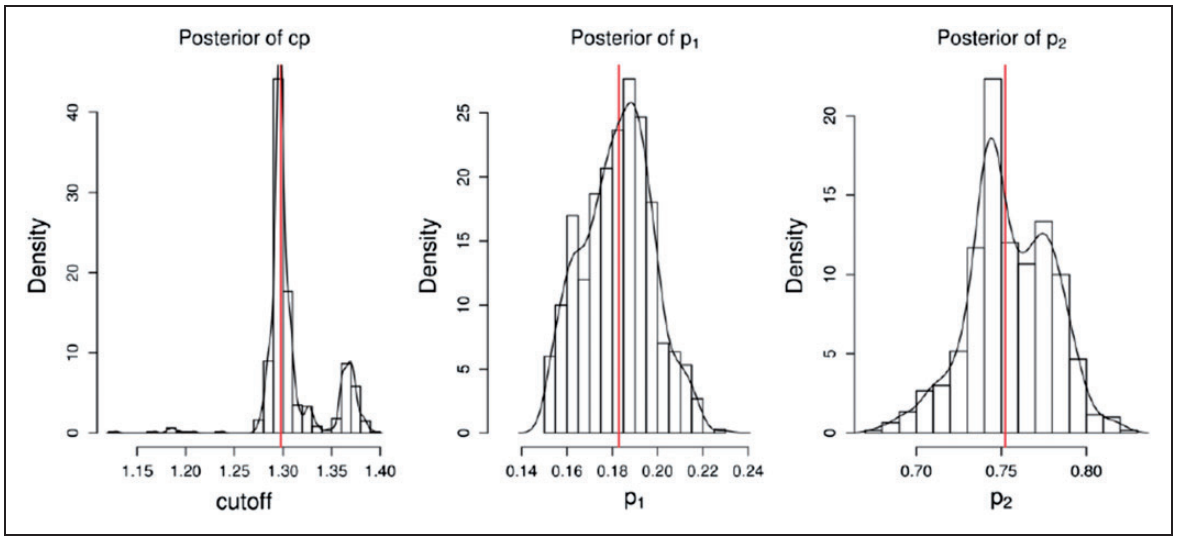

Figure 6. Plot of the posterior distribution for the parameter cp (left panel), p_1 (middle panel), p_2 (right panel) estimated by the Bayesian model. The red vertical lines denote the median of the distribution.

prior for the cutoff in the interval (−10,10) and priors for the predictive values defined as in equation (2). We also report the ML estimator and the PSI index.

Figure 6 shows the posterior distributions for the cutoff (left panel) and the predictive values p_1 and p_2 (middle and right panels respectively). The MLE of the cutoff was found equal to 1.29 with 95% CI (1.27−1.31), while the posterior mean of the cutoff was 1.30 with 95% credible interval (1.27−1.38). The PSI index, which we remind that maximizes a different objective function, estimates the optimal cutoff to be 3.63 with 95% bootstrapped CI (2.00−3.77). At that cut-off, the PPV and 1-NPV was equal to 1 and 0.32, respectively. The Bayesian posterior mean for p_1 and p_2 was found equal to 0.18 with 95% credible interval (0.15−0.21) and 0.75 with 95% credible interval (0.70−0.79) respectively. The MLE for p_1 was 0.18 with 95% confidence interval (0.15–0.21) and for p_2 was 0.75 with 95% confidence interval (0.68–0.81).

4.2 Application on survival data: Weibull model for melanoma data

To illustrate that the proposed approach is useful for more complex settings, we consider identifying the appropriate cutoff for a time to event endpoint. For the following applications on time to event data, we assume the following: let T_i denote the event time for subject i. Due to censoring, instead of observing T_i, we observe the bivariate vector $(min(T_i,\ C_i),\ \Delta_i)$ where $\Delta_i = I(T_i \leq C_i)$ with I the indicator function and C_i the censoring time.

The data used are the melanoma dataset available from the R package *timereg*.[24] The data consist of measurements made on patients with malignant melanoma and patients with a thick tumor are thought to have an increased chance of death from melanoma, thus the objective is to estimate a cutoff value on (the log scale of) the tumor size such that the patients below and above the cutoff have a pronounced difference in their hazard rates. We run the analysis using the R package *MHadaptive*[25] and we used uniform priors for all the parameters. The R-code is available upon request from the author.

To set up the model in the survival setting, the thickness of the tumor on the log scale is denoted by X, T denotes time to death and is assumed to have a Weibull distribution with shape parameter r and scale parameter λ. The assumption is that, based on the thickness of the tumor, we can estimate a cutoff cp such that the two groups defined by cp have different hazard functions. Therefore, the shape and scale parameter for the patients whose thickness of their tumor is below cp is r_1 and λ_1, respectively, and accordingly, r_2 and λ_2 for those patients with $X > cp$.

$$T|X \sim\ Weibull(r, \lambda)\ \text{ with } r = \begin{cases} r_1, & \text{if } X \leq cp \\ r_2, & \text{if } X > cp \end{cases} \quad \text{and} \quad \lambda = \begin{cases} \lambda_1, & \text{if } X \leq cp \\ \lambda_2, & \text{if } X > cp \end{cases}$$

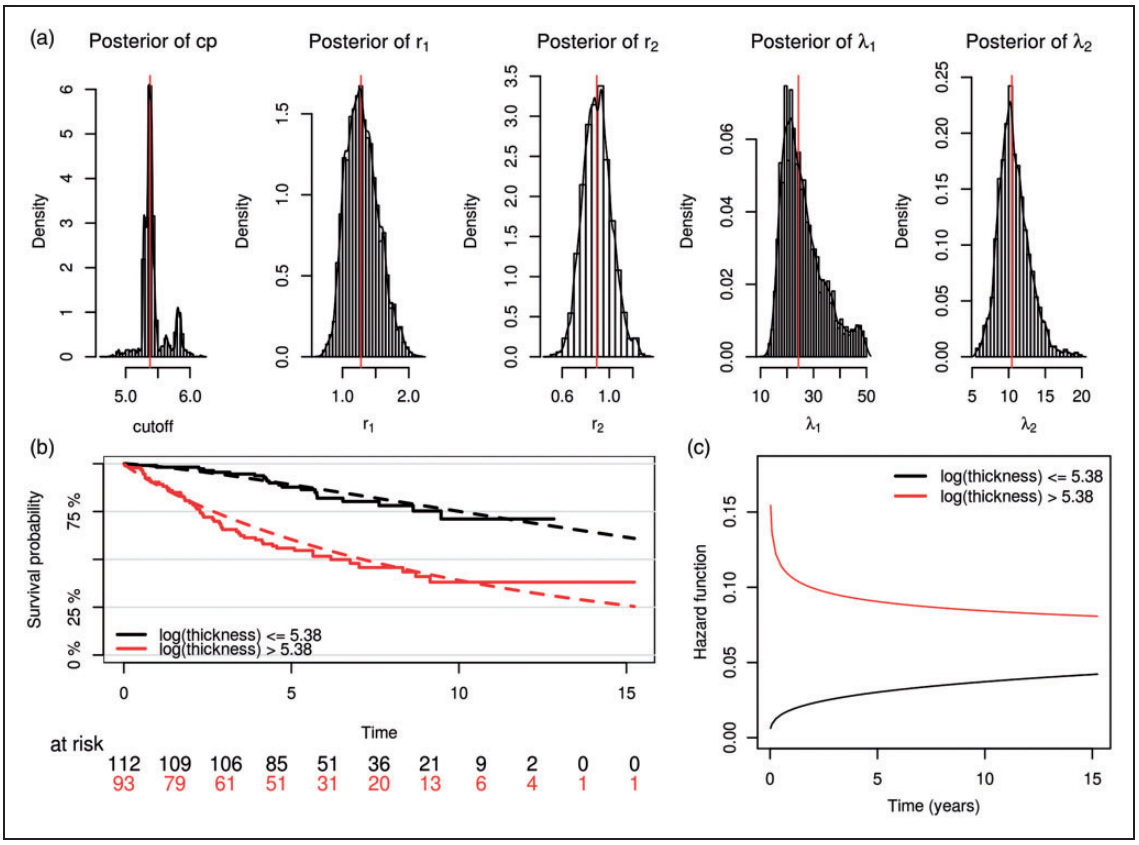

Figure 7. (a) Histograms of the posterior distributions for the cutoff (left panel), the shape parameters r_1 and r_2 (middle panels) and scale parameters λ_1 and λ_2 (right panels) for the Weibull model fitted to the melanoma data. The red vertical lines correspond to the posterior median of the distribution. (b) Survival curves for the patients above and below the estimated cutoff $cp = 5.38$, taken as the posterior median of the posterior density. The solid lines are the Kaplan–Meier curves and the dashed lines the Weibull survival curves. (c) Plot of the hazard function for the groups below and above the cutoff estimated by the Weibull model by plugging in the posterior means of the parameters r_1, λ_1, r_2, λ_2.

Figure 7(a) shows the posterior densities for the cutoff, the shape and scale parameters. We took the medians of the posterior densities as point estimates for each parameter. In Figure 7(b) we plot the survival curves, estimated with the Kaplan–Meier estimate, for the patients below and above the posterior cutoff estimate, which was taken as the posterior mean equal to $\widehat{cp} = 5.38$ with 95% credible interval (5.07–5.86). In the same figure, we plot the survival curves for the Weibull model with dashed lines. As seen from the plot, the survival probability decreases with higher tumor thickness value. To test whether the survival curves for the patients below and above the estimated cutoff value differ significantly, we applied the log-rank test which showed that there is a significant difference in survival ($p < <0.05$). Figure 7(c) shows the hazard function for the two groups by plugging in the estimated shape and scale parameters, i.e. the hazard function for the Weibull model becomes $h(t) = \begin{cases} \frac{r_1}{\lambda_1}(\frac{t}{\lambda_1})^{r_1-1}, & \text{if } X \leq cp \\ \frac{r_2}{\lambda_2}(\frac{t}{\lambda_2})^{r_2-1}, & \text{if } X > cp \end{cases}$, with r_1, λ_1, r_2, λ_2 taken as the means of the posterior densities.

5 Discussion

To enable targeted therapies and enhance medical decision-making, biomarkers are increasingly used in diagnostic tests. When using quantitative biomarkers for classification purposes, defining a reliable cutoff value for the biomarker is a critical step in the drug development process, as the patient selection process in the subsequent development steps may depend on this value. Although classification probabilities, sensitivity and specificity, are considered more relevant to quantify the inherent accuracy of the test, predictive values quantify the clinical utility of the test.

We have proposed a Bayesian method to estimate the cutoff value of a biomarker assay using the predictive values, and also determine the uncertainty around these estimates. We used a step function, which serves as an approximate model facilitating classification into two groups that have a pronounced difference in their response rates. The advantage of using the step function is that the cutoff and predictive values are parameters of the model. Even in the case that the assumption of a step function is strong and the model is misspecified, the estimates of the assumed step function are consistent for the parameter values for which the assumed model minimizes the distance from the true distribution in terms of Kullback–Leibler divergence.[7,8] A more careful investigation of this approach is worth further exploration.

Alternative approaches in classification problems using logistic regression are frequently employed in practice, for example using a probability threshold of $p = 0.5$ to classify patients, or choose p such that the Brier score,[26] a measure of accuracy of predictions, is minimized. However, these methods do not directly address the goal of population separation with regard to positive and negative predictive values. Moreover, they do not directly provide credible or confidence intervals for the parameters of interest which was one of the major goals of the proposed method. Nevertheless, we have compared the Bayesian approach with these methods and found that the estimated parameters of *cp* are more biased compared to the Bayesian estimates. Detailed figures can be found in Appendix 1.

The proposed Bayesian approach allows for the estimation of the distribution of the cutoff for continuous and ordinal biomarkers and permits probabilistic statements about the cutoff values and, say, the response rates in the two groups. Together with the potential incorporation of prior information, this is deemed useful especially in the earlier phases of drug development. Results suggest that the proposed Bayesian method is very tractable in estimating the parameters of interest, resulting in point estimators (e.g. posterior mean) that are practically unbiased in all scenarios, for all prior constellations and sample size assumptions.

In this article, we presented four different prior specifications, including uninformative, informative, and mixture priors. In all cases, estimation gave satisfying results. Especially when more accurate prior information is available, the estimated parameters are nearly unbiased with high precision and good coverage. We suggest a mixture prior that works well in practice, as it is robust towards potential prior-data conflict. For a dataset of $n = 200$ observations, the Bayesian approach takes 6.3 s to run on a windows machine with processor Intel Xeon CPU E7-8867 v3 @ 2.5 GHz, compared to frequentist approaches (MLE 0.15 s and for PSI 3.7 s together with the bootstrapped CI). Although the computational time for the proposed approach is increased, as is the case for Bayesian methods, is not prohibitive.

The approach described in this article can be used as a basis for further investigation. The suggested method was applied to a single biological marker, but it can be generalized to multiple markers. One way to deal with multiple markers is to estimate a composite score for each patient using a combination of markers (under some working model, for example, under the logistic model), and then consider this score as the new marker. Furthermore, it would be of great interest to consider the generalization of the method to estimate multiple cutoffs that can be used potentially for subgroup identification. In that case, model selection can be used to decide how many cutoffs (indicating the number of subgroups) the model can have according to the data.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 633567 and is part of the IDEAS European training network (http://www.ideas-itn.eu/). This report is in part independent research arising from Prof Jaki's Senior Research Fellowship (NIHR-SRF-2015-08-001) supported by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

ORCID iD

Eleni Vradi http://orcid.org/0000-0003-0330-3309

Supplementary Material

Supplementary material is available for this article online.

References

1. Colburn WA. Biomarkers in drug discovery and development: from target identification through drug marketing. *J Clin Pharmacol* 2003; **43**: 329–341.
2. Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press, 2003.
3. Perkins NJ and Schisterman EF. The inconsistency of ''optimal'' cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am J Epidemiol* 2006; **163**: 670–675.
4. Schisterman EF and Perkins N. Confidence intervals for the Youden index and corresponding optimal cut-point. *Commun Stat Part B: Simul Computat* 2007; **36**: 549–563.
5. Lunceford JK. Clinical utility estimation for assay cutoffs in early phase oncology enrichment trials. *Pharmaceut Stat* 2015; **14**: 233–341.
6. Lever J, Krzywinski M and Altman N. Points of significance: logistic regression. *Nat Methods* 2016; 13: 541–542.
7. Huber PJ. The behaviour of maximum likelihood estimates under non?standard conditions. In: LeCam LM and Neyman J (eds) *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability* vol. 1, 1967, pp. 221–233. Berkeley: University of California Press.
8. Bunke O and Milhaud X. Asymptotic behavior of Bayes estimates under possibly incorrect models. *Ann Stat* 1998; **26**: 617–644.
9. Kullback S and Leibler RA. On information and sufficiency. *Ann Math Stat* 1951; **22**: 79–86.
10. Schmidli H, Gsteiger S, Roychoudhury S, et al. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* 2014; **70**: 1023–1032.
11. Linn S and Grunau PD. New patient-oriented summary measure of net total gain in certainty for dichotomous diagnostic tests. *Epidemiol Perspect Innovat* 2006; **3**: 11.
12. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3**: 32–35.
13. Chu H and Cole SR. Estimation of risk ratios in cohort studies with common outcomes: a Bayesian approach. *Epidemiology* 2010; **21**: 855–862.
14. Wagenmakers EJ, Lee M, Lodewyckx T, et al. Bayesian versus frequentist inference. In: Hoijtink H, Klugkist I and Boelen PA (eds) *Bayesian evaluation of informative hypotheses*. New York, NY: Springer, 2008, pp.181-207.
15. Team RC. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. R version 3.3.3 Vienna, Austria.
16. Lopez-Raton M, Rodriguez-Alvarez MX, Cadarso-Suarez C, et al. Optimal cutpoints: an R package for selecting optimal cut-points in diagnostic tests. *J Stat Software* 2014; **61**: 1–35. http//www.jstatsoft.org.
17. Bolker B. R Development Core Team. 2014. bbmle: Tools for general maximum likelihood estimation. R package version 1.0. 17. Computer program. 2011.
18. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970; **57**: 97–109.
19. Metropolis N, Rosenbluth AW, Rosenbluth MN, et al. Equation of state calculations by fast computing machines. *J Chem Phys* 1953; **21**: 1087–1092.
20. Ibrahim JG, Chen MH and Sinha D. *Bayesian survival analysis*. New York: Springer, 2001.
21. Hartigan JA and Hartigan PM. The dip test of unimodality. *Ann Stat* 1985; **3**: 70–84.
22. Etzioni R, Pepe M, Longton G, et al. Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Med Decis Making* 1999; **19**: 242–251.
23. Broemeling LD. *Advanced bayesian methods for medical test accuracy*. Boca Raton, FL: Chapman & Hall/CRC Press, 2011.
24. Martinussen T and Scheike T. *Dynamic regression models for survival analysis. Statistics for biology and health*. NY: Springer, 2006.
25. Chivers C. *MHadaptive: general Markov Chain Monte Carlo for Bayesian inference using adaptive Metropolis-Hastings sampling*. Retrieved from http://cran.r-project.org/web/packages/MHadaptive/MHadaptive.pdf (2012).
26. Brier G. Verification of forecasts expressed in terms of probability. *Mon Wea Rev* 1950; **78**: 1–3.

Appendix 1. Bias, sample size and prior specification

We explored in a simulation study the performance of the Bayesian method in terms of the (absolute) difference of the estimated *cp* from the true value of the cutoff for different sample sizes ($n = 50$, 75, 100, 150, 200, 500). As expected, when the sample size increases, the bias is shrinking towards zero as we can see in Figure 8.

In Tables 6 and 7, we present simulation results concerning the predictive values for a sample size of $n = 50$. These results are complementary for the simulations described in section 3.2. We report the Bias, Coverage and interval width for scenarios 1–4 and for all methods. For the Bayesian method, even with a small sample size, the

Table 6. Mean bias of the estimate of the predictive values p_1 and p_2 over 10,000 simulation runs for the Bayesian method, the MLE and PSI approach and scenarios 1–4 and for $n = 50$.

p_1, p_2	Bias						
Methods	Bayesian						
Prior	UP	IPN	IPP	MixN	MixP	PSI	MLE
Scenario 1							
p_1	3×10^{-2}	3×10^{-2}	3×10^{-2}	3×10^{-2}	7×10^{-3}	4×10^{-2}	4×10^{-2}
p_2	-3×10^{-2}	-4×10^{-2}	3×10^{-2}	-3×10^{-2}	-8×10^{-3}	9×10^{-2}	8×10^{-2}
Scenario 2							
p_1	3×10^{-2}	2×10^{-2}	2×10^{-2}	3×10^{-2}	2×10^{-2}	-3×10^{-2}	6×10^{-2}
p_2	-4×10^{-2}	-4×10^{-2}	-3×10^{-2}	-3×10^{-2}	-3×10^{-2}	-4×10^{-3}	5×10^{-2}
Scenario 3							
p_1	2×10^{-2}	2×10^{-3}	2×10^{-2}	1×10^{-2}	2×10^{-2}	-7×10^{-3}	6×10^{-2}
p_2	-5×10^{-2}	-1×10^{-1}	-5×10^{-2}	-6×10^{-2}	-5×10^{-2}	8×10^{-2}	1×10^{-1}
Scenario 4							
p_1	4×10^{-2}	4×10^{-2}	4×10^{-2}	4×10^{-2}	4×10^{-2}	3×10^{-3}	1×10^{-1}
p_2	-2×10^{-2}	-2×10^{-2}	-2×10^{-2}	-2×10^{-2}	-2×10^{-2}	6×10^{-3}	5×10^{-2}

Table 7. Average coverage and width of the credible/confidence interval for the estimates of the predictive values p_1 and p_2 over 10,000 simulation runs for scenarios 1–4 for $n = 50$.

p_1, p_2	Coverage							Interval width						
Methods	Bayesian							Bayesian						
Prior	UP	IPN	IPP	MixN	MixP	PSI	MLE	UP	IPN	IPP	MixN	MixP	PSI	MLE
Scenario 1														
p_1	0.956	0.969	0.957	0.961	0.955	0. 986	0.914	0.235	0.230	0.223	0.232	0.231	0. 287	0.217
p_2	0.975	0.949	0.951	0.969	0.969	0.772	0.976	0.365	0.369	0.352	0.364	0.359	0.292	0.372
Scenario 2														
p_1	0.952	0.968	0.951	0.952	0.949	0.943	0.882	0.308	0.309	0.298	0.306	0.303	0.333	0.292
p_2	0.971	0.947	0.951	0.966	0.964	0.957	0.969	0.258	0.269	0.256	0.258	0.256	0.300	0.290
Scenario 3														
p_1	0.960	0.971	0.946	0.962	0.951	0.969	0.897	0.309	0.314	0.282	0.302	0.294	0.395	0.291
p_2	0.982	0.885	0.965	0.968	0.981	0.814	0.902	0.416	0.427	0.368	0.411	0.390	0.356	0.443
Scenario 4														
p_1	0.956	0.927	0.956	0.954	0.960	0.954	0.991	0.243	0.243	0.242	0.248	0.244	0.279	0.443
p_2	0.950	0.949	0.951	0.951	0.955	0.949	0.987	0.315	0.317	0.314	0.320	0.317	0.407	0.487

Note: The credible intervals are computed by the quantile method. Bootstrapping was used to calculate the confidence interval for the PSI method and the profile CI are presented for the MLE method.

bias of the parameters (on absolute scale) is always less than 4% on average. For the PSI and ML method, the bias of the estimates is small, whereas the coverage does not always reach the nominal level and the interval widths are always slightly bigger than the Bayesian method.

In Figure 9, we see the distribution of the absolute difference of the estimated *cp* from the true value of the cutoff over the 10,000 simulation runs, for the Bayesian method when we consider different priors. The results are presented for data generated as in Scenario 1 with a sample size of $n = 50$. Even with a small sample size, the bias is always smaller than 10% on average. When the prior is informative precise then we achieve the smallest bias, whereas when we consider a robust mixture of precise and uniform prior the bias is slightly higher but still very small.

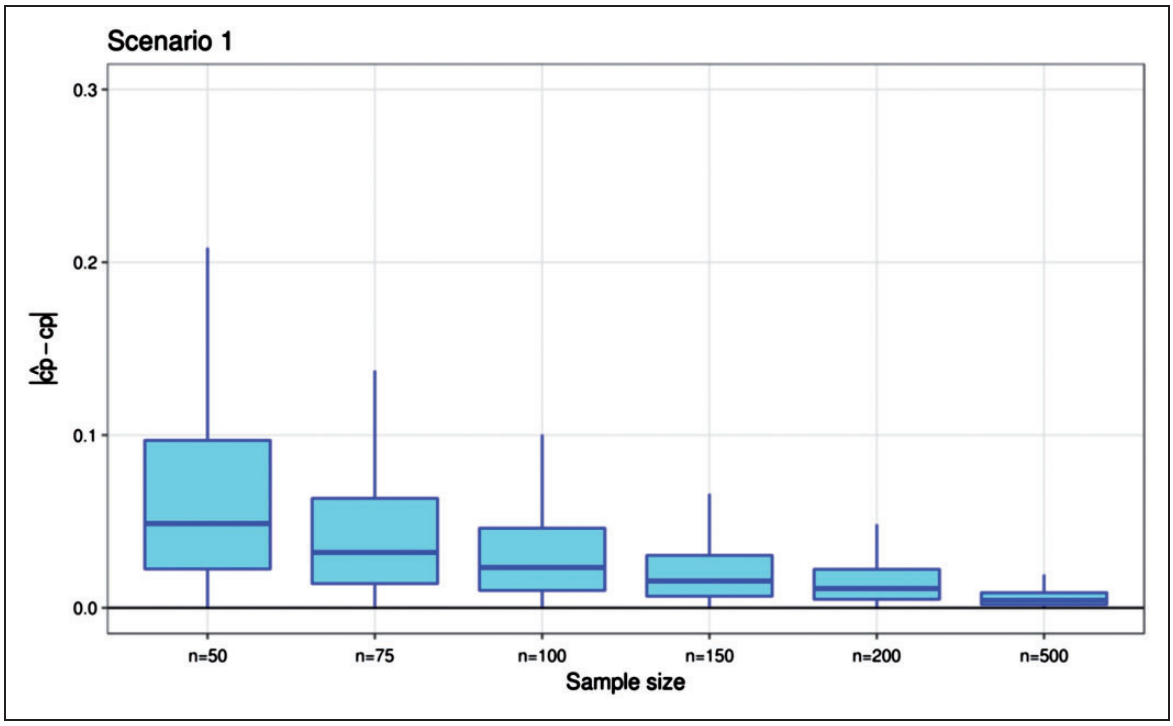

Figure 8. Boxplots of the absolute difference between the estimate and the true value of the cutoff *cp* over 10,000 simulation runs for Scenario 1 for varying samples sizes ($n = 50, 75, 100, 150, 200, 500$). Results shown for the Bayesian method with a uniform prior. The posterior mean was used as an estimate for the cutoff.

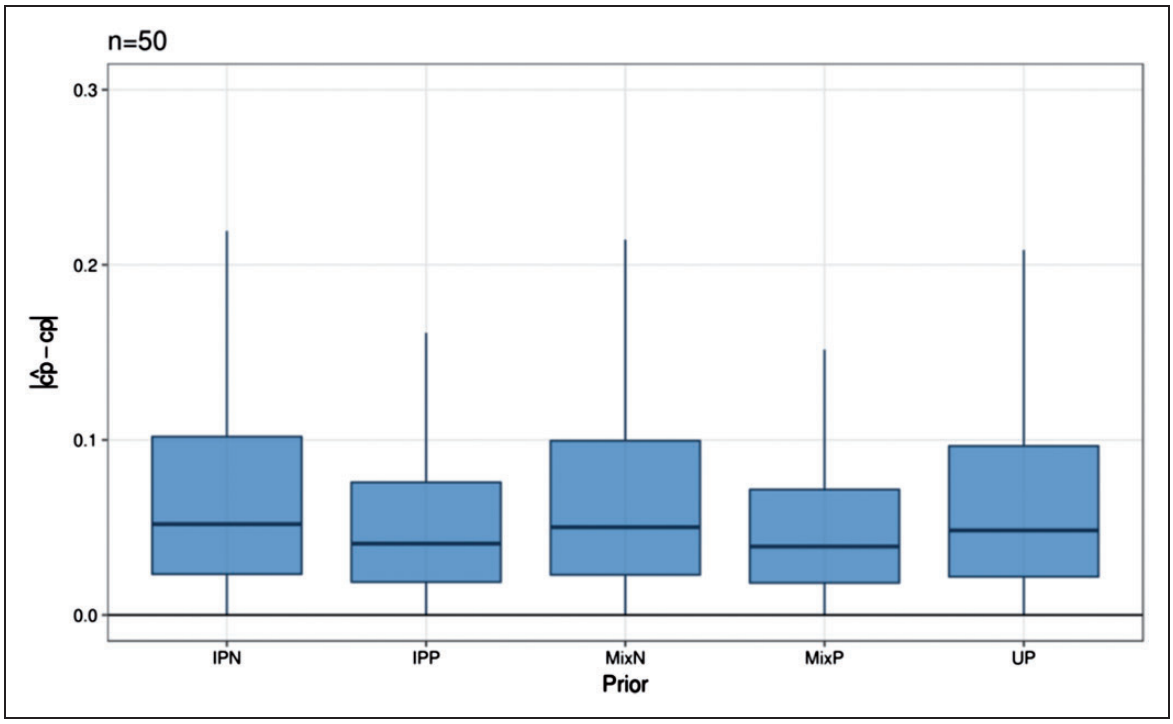

Figure 9. Boxplots for the absolute difference between the estimate $\widehat{cp}$ and the true value of *cp* estimated with the Bayesian model over 10,000 simulation runs for Scenario 1. In this simulation, we used $n = 50$ samples for the case of (from left to right) an Informative Prior Non-precise (IPN), an Informative Prior Precise (IPP), a Mixture Prior Non-precise (UP + IPN), a Mixture Prior Precise (UP + IPP) and a Uniform Prior (UP).

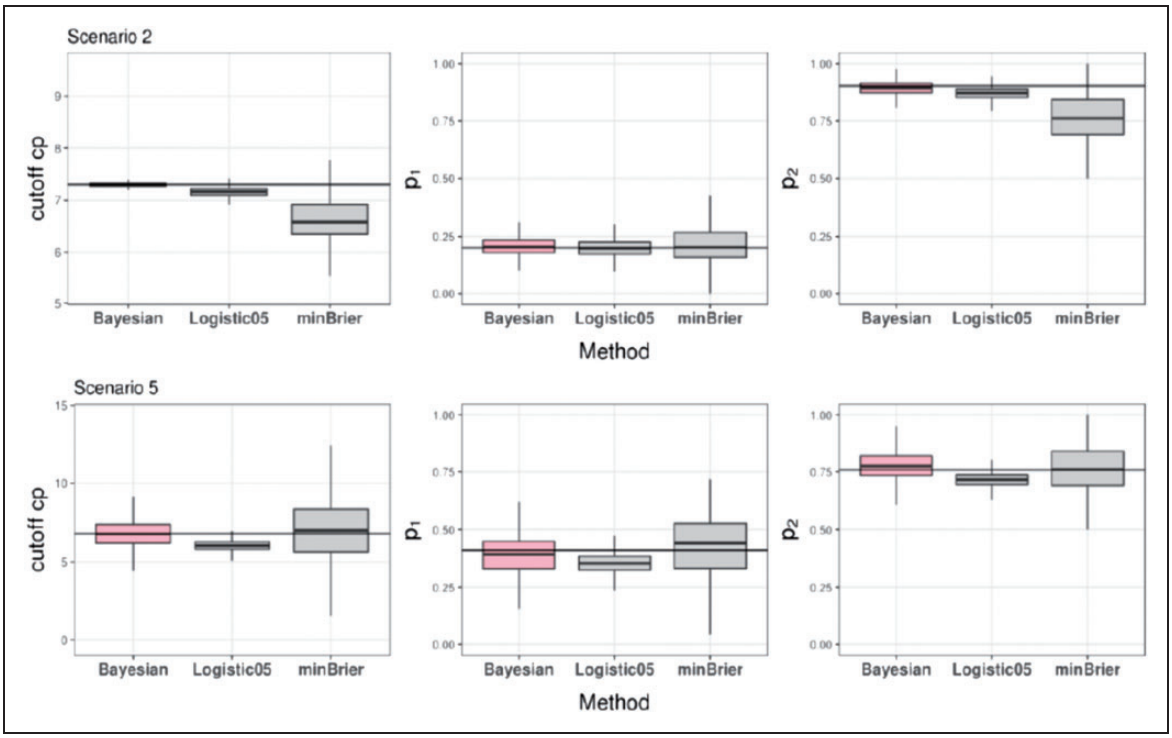

Figure 10. Boxplots of the estimated parameters cp, p_1, p_2 (left, middle and right plots, respectively) by the Bayesian method, the Logistic regression with a cutoff at $p = 0.5$ and by minimizing the Brier score. Results shown for 10,000 simulation runs for scenario 2 where the generating model is a step function (upper panel) and scenario 5 where the generating model is logistic (lower panel). The black horizontal lines correspond to the true values of the parameters.

2 Comparison with other methods

We considered the simulated data from scenario 2 (generating model step function) and scenario 5 (generating model logistic function) as examples to show the results regarding the fit of the logistic with the choice of $p = 0.5$ and the method that estimates p as the value minimizes the Brier score. Results are shown in Figure 10, where we see that the estimated parameters by the logistic model with the choice of $p = 0.5$ are more biased compared with the Bayesian approach. For scenario 5, the posterior means by the proposed approach are similar to the method that estimates p as the value that minimizes the Brier score but the latter approach results in much higher variability. However, results differ from the method that used the probability cutoff of $p = 0.5$, where we see that it underestimate the true parameters.

3 Conditional Kullback–Leibler divergence between the theoretical and fitted model

3.1 Estimation of the predictive values

Let us assume that the data generating function of the true model is a logistic function, i.e. $Y|X \sim Bernoulli(p)$, with link function $logit(p) = X\beta$, $p(x) = \frac{e^{X\beta}}{1+e^{X\beta}}$ and joint probability distribution function $g(x, y)$. The conditional distribution of $Y|X$ is G and $g(y|x)$ the conditional density. Let us now consider that the fitted model assumes a step function for the probability of response with $Y|X \sim Bernoulli(q)$, $q(x) = \begin{cases} q_1, & \text{if } x \leq cp \\ q_2, & \text{if } x > cp \end{cases}$ and corresponding conditional probability distribution F. The joint probability distribution function is $f(x, y)$ and $f(y|x)$ the conditional density. We would like to show that the estimates of the parameters in the step model are the ones that minimize the Kullbaclk–Leibler (KL) divergence between the two probability distributions F and G. That is, the expectation of the log difference between the conditional probability of data in the original distribution with the approximate distribution.

The conditional Kullbaclk–Leibler divergence between the two probability distributions F and G is defined as

$$D_{KL}(G||F) = \int_{X\in A} g(x) \int_{Y\in B} g(y|x) log \frac{g(y|x)}{f(y|x)}\, dy\, dx$$

where $g(x)$ is the pdf of X, where $X \in A$ and $Y \in B$.

We first calculate the inner integral $\int_{Y\in B} g(y|x) log \frac{g(y|x)}{f(y|x)}\, dy =$

$$E_G\left[y \log \frac{p(x)}{q(x)} + (1-y)\log \frac{1-p(x)}{1-q(x)}\right] == \begin{cases} E_G\left[y \log \frac{p(x)}{q_1} + (1-y)\log \frac{1-p(x)}{1-q_1}\right], & \text{for } X \le cp \\ E_G\left[y \log \frac{p(x)}{q_2} + (1-y)\log \frac{1-p(x)}{1-q_2}\right], & \text{for } X > cp \end{cases}$$

$$= \begin{cases} p(x)\log \frac{p(x)}{q_1} + (1-p(x))\log \frac{1-p(x)}{1-q_1}, & \text{for} \quad X \le cp \quad (I) \\ p(x)\log \frac{p(x)}{q_2} + (1-p(x))\log \frac{1-p(x)}{1-q_2}, & \text{for} \quad cp < X \; (II) \end{cases}$$

We need to minimize $D_{KL}(g(y|x)||f(y|x))$ over X, assuming that X has pdf $g(x)$ and $X \in [0, cp] \cup (cp, \infty]$. For a given cp, we estimate q_1 and q_2 by minimizing

$$D_{KL}^{(I)}(g(y|x)||f(y|x)) = \int_0^{cp} g(x) \left[p(x)\log \frac{p(x)}{q_1} + (1-p(x))\log \frac{1-p(x)}{1-q_1}\right] dx \quad \text{and}$$

$$D_{KL}^{(II)}(g(y|x)||f(y|x)) = \int_{cp}^{\infty} g(x) \left[p(x)\log \frac{p(x)}{q_1} + (1-p(x))\log \frac{1-p(x)}{1-q_1}\right] dx \quad \text{respectively}$$

$$D_{KL}^{(I)}((g(y|x)||f(y|x)) = \int_0^{cp} g(x) \left[p(x)\log \frac{p(x)}{q_1} + (1-p(x))\log \frac{1-p(x)}{1-q_1}\right] dx$$

$$= \int_0^{cp} g(x)\, p(x) \log p(x) dx - \int_0^{cp} g(x)\, p(x) \log q_1 dx$$

$$+ \int_0^{cp} g(x)(1-p(x))\log(1-p(x)) dx - \int_0^{cp} g(x)(1-p(x))\log(1-q_1) dx$$

Calculate $\frac{d}{dq_1} D_{KL}^{(I)}(g(y|x)||f(y|x)) = -\frac{1}{q_1} \int_0^{cp} g(x)\, p(x) dx + \frac{1}{1-q_1} \int_0^{cp} g(x)(1-p(x))\, dx$

Setting equal to zero and solving with respect to q_1, we then obtain

$$q_1 = \frac{\int_0^{cp} g(x)\, p(x)\, dx}{\int_0^{cp} g(x)\, dx}$$

Following the same calculations for $D_{KL}^{(II)}(g(y|x)||f(y|x))$ and solving with respect to q_2, we get $q_2 = \frac{\int_{cp}^{\infty} g(x)\, p(x)\, dx}{\int_{cp}^{\infty} g(x)\, dx}$

3.2 Estimation of the cutoff

The estimation of the cut-off cp is not straightforward and can be done by using numerical minimization. To do this we need to repeat the calculations above for all possible values of cp and to find the step model that minimizes $D_{KL}(g(y|x)||f(y|x))$.

4 R and SAS code

The R code is not included here due to the extent of the code and the R scripts are available upon request from the corresponding author. The following is the SAS code that was used for fitting the Bayesian model for Scenario 1 using a mixture prior with imprecise part (MixN). The code can be modified to include other prior specifications.

```
PROC MCMC
  data=Data outpost=Dataoutput
    nbi=10000
    nmc=30000
    thin=50
    seed=seed
  monitor=(p1 p2 cp I w);
by dataID; # this is used for the simulated data; otherwise is omitted if a single dataset is used.
PARMS cp1 cp2 p1 p2 w I;
  prior cp1 ~ uniform(1,15);
  prior cp2 ~ normal(5,sd=1);
  hyperprior I~ beta(1,1);
  prior w ~ binary(I);
  cp = w*cp1 + (1-w)*cp2;
  prior p1 ~ uniform(0, 1);
  prior p2 ~ uniform(p1, 1);
  p= (X<=cp)*p1 + (X>cp)*p2;
  model y~ binary(p);
RUN;
```

Bayesian variable selection and classification with control of predictive values

Journal Title
XX(X):2–??

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Eleni Vradi[1,4], Thomas Jaki[2], Richardus Vonk[3], Werner Brannath[4]

Abstract

In clinical development, the selection of novel prognostic biomarkers has to be paired with the evaluation of the clinical utility of the risk score defined by these biomarkers. Before we apply the selected markers into routine standard care, for classification or patient selection, a cutoff value must be assessed based on the positive (PPV) and negative (NPV) predictive value. In this paper, we propose a Bayesian variable selection method which incorporates information about the predictive values into the biomarker selection process and simultaneously estimates the cutoff value on the risk score of the selected markers. A step function is used to model the probability of response, such that the cutoff and predictive values are parameters of the model. This model allows for a pre-specification of a minimum PPV (or NPV) in the variable selection algorithm. The choice of different prior distributions is compared and discussed via simulation studies and a real data application.

Keywords

Bayesian variable selection, classification, predictive values, shrinkage priors, step function, risk score, biomarkers, cutoff estimation

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

1 Introduction

In disease screening and prognosis, markers that predict the risk of a disease are needed. The selection of novel prognostic markers and the evaluation of their predictive accuracy can be used to stratify patients according to future risk of an outcome. Therefore, identifying the most promising biomarkers is of major importance in a clinical setting. From a Bayesian perspective, biomarker (variable) selection is done by imposing a prior distribution on the regression coefficients. There are two main alternative choices regarding priors: discrete mixture priors or shrinkage priors. However, in a Bayesian setting, the selection of the non-zero coefficients is done *ad hoc*, since the probability that the value of an effect is exactly zero is always zero. Among many methods, a popular choice is setting an appropriate thresholding criterion on the posterior inclusion probability of each variable[1] or by examining if zero is included in the credible interval for each of the effects[2,3]. Other methods for choosing variables or models can be constructed using decision theoretic arguments[4].

After a set of markers is selected, a prognostic factor or risk score must be defined and its predictive accuracy must be evaluated before it is adopted for clinical screening or patient selection. The most popular accuracy measure used in clinical literature is the area under the ROC curve (AUC)[5,6], which is a summary index of two accuracy metrics: the true positive rate (TPR) or sensitivity and false positive rate (FPR) (1-specificity). However, neither the AUC nor sensitivity and specificity reflect the ability of predicting the future outcome conditional on the risk score. In contrast, measures such as positive and negative predictive values (PPV and NPV) are clinically relevant quantities that we can directly interpret in terms of the probability of disease given a positive or negative test result. Because predictive

[1] Research and Early Development Statistics, Bayer AG, Muellerstrasse 178, 13342 Berlin, Germany
[2] Medical and Pharmaceutical Statistics Research Unit Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, United Kingdom
[3] Medical Writing and Statistics Oncology, Bayer AG, 13342 Berlin, Germany
[4] FB 03 Mathematik/Informatik Institute for Statistics and Competence Center for Clinical Trials, University of Bremen, 28359 Bremen, Germany

Corresponding author:
Eleni Vradi, Bayer AG, 13342 Berlin, Germany
Email: eleni.vradi@bayer.com

Prepared using sagej.cls

values are functions of sensitivity, specificity and the prevalence of the disease, a risk score with high sensitivity and specificity, and thus a high AUC, can have poor PPV when applied to low prevalence populations[7].

Another key factor to a successful drug development project is the identification of the patient population most likely to respond to a treatment or predict a person's disease state. Because drugs are only clinically useful and effective for patients below or above a certain risk score, a reliable cutoff value (on the risk score) is needed in order to classify patients according to their probability of response. Commonly used strategies for cutoff selection are based on summary indexes mentioned above (i.e. desired sensitivity or PPV) as for example, using the Youden index[8] or the PSI index[9]. These measures are used to select a cutoff on a pre-defined risk score (of pre-selected variables). In other words, classification and cutoff selection are implemented using a two stage procedure; first perform variable selection and compute a risk score out of the selected features and then at a second stage, select a cutoff. However, with this approach the selected features are not necessarily the ones that would guarantee a minimum classification performance, as the optimization is done under different objective functions, i.e. at the first stage, variable selection is done using a logistic model and at the second stage, the estimation of the cutoff is done by fitting a step function[10]. Instead, information regarding target classification measures should be included also in the objective function of the first step in order to optimize the balance between i) the desired accuracy measures and ii) the sparsity of the resulted model. In most parts of this paper, we consider the PPV over the NPV as the most important measure. However, this assumption is not crucial for the methodology explained below.

In this paper, we propose a Bayesian method that combines the properties of variable selection, with the advantages of using a step function to model the probability of response for classification purposes. By using the step function the predictive values and the cutoff are now parameters of the model. If a target response rate is required by researchers, e.g. PPV should only take values above a pre-determined lower value, it can be efficiently incorporated in the variable selection algorithm. By restricting either the PPV or the NPV values of the test in a pre-specified interval through the selection of the prior, we are able to select those features that satisfy the restriction.

In section 2, we introduce the hierarchical Bayesian model for variable selection that take into account the predictive values of the risk score which is calculated from the selected variables and present the model that controls one of the predictive values. The model that assumes a step function to model the probability of response is a very useful way to classify individuals based on their response rates. A particular interest lies in the posterior distribution of the cut-off and predictive values. On the one hand, it gives guidance on the choice of the cut-off value, and, on the other hand, provides information on whether the assumed constraints on the predicitve values can be put in practice. We already note at this point that the posterior estimates for the cut-off and predictive values fundamentally deviate from the true model parameters, because the first has to be conditional on the values of the selected biomarker coefficients in the final risk score.

In section 3, we conducted a simulation study to investigate the behavior of different shrinkage priors on the regression coefficients as well to compare the behavior of the proposed model under mispecification. We use MCMC methods to estimate the posterior parameter distributions and the posterior inclusion probabilities for the predictors. A real data example is presented in section 4 before we conclude with some discussion.

2 Methods

2.1 Bayesian hierarchical model for variable selection

In this section we present a Bayesian model for simultaneous variable selection and cutoff estimation of a risk score, as well as its predictive values. Suppose we have data (X, y), where $y = (y_1, y_2, ..., y_n)$ is the binary response variable, with $y_i \in \{0, 1\}$ for all $i = 1, ..., n$ and let $X = (X_{1j}, X_{2j}, ..., X_{nj})$ denote the $n \times d$ matrix of $j = 1, ..., d$ predictors (biomarker measurements) for n individuals and assume that X is measured on all patients. We will assume that the n observations are independent and the predictors are standardized to have mean zero and standard deviation one.

We consider the risk score Z to be a linear predictor, i.e. $\eta = Z = X^T\beta$, where β are the regression coefficients. As in[10], we assume that the probability of response p can be modeled by a step function that assumes constant positive and negative predictive values when the risk score is above or below a specific cut-off. With this assumption, the predictive

values and the cutoff are now parameters of the model, respectively. The positive predictive value is defined as the conditional probability of response given a positive test result, i.e. $P(Y = 1|T^+)$. Conventionally, the test is positive, T^+, if the biomarker exceeds a certain cutoff cp, and is negative otherwise. Similar statements apply for the negative predictive value which is defined as the conditional probability that an individual is a non-responder given a negative test result, i.e.$P(Y = 0|T^-)$. For further details about the model see[10].

For the risk score $Z = X^T\beta$, the positive predictive value is defined as $PPV(cp) = P(Y = 1|Z > cp)$ and the negative predictive value is defined as $P(Y = 0|Z \leq cp)$. As model parameter we will use the complementary of the NPV, namely $1 - NPV(cp) = P(Y = 1|Z \leq cp)$ will be used. The Bayesian hierarchical model for variable selection and classification is specified as

$$Y|X \sim Bernoulli(p)$$

$$p = P(Y = 1|Z = X^T\beta) = \begin{cases} p_1 = P(Y = 1|Z \leq cp) & \text{if} \quad X^T\beta \leq cp \\ p_2 = P(Y = 1|Z > cp) & \text{if} \quad X^T\beta > cp \end{cases} \tag{1}$$

$$\beta \sim F$$

$$cp \sim Uniform(a, b)$$

$$p_1 \sim Uniform(0, 1) \qquad \text{and} \qquad p_2 \sim Uniform(p_1, 1)$$

where F is a prior distribution for the coefficients. The different choices of prior will be discussed in the following subsection 2.2. The cutoff cp is a parameter lying in a bounded interval $[a, b]$ strictly included in the support of the risk score Z. The parameter $p_1 = 1 - NPV$ expresses the probability of response given that Z is below the cutoff value cp and $p_2 = PPV$ expresses the probability of response given that the value of $X^T\beta$ is greater than cp.

The primary question that arises in a clinical setting is the selection of a set of biomarkers that fulfill a specific target, concerning the clinical utility of the test. Therefore, variable

selection is done subject to pre-determined lower bound on positive predictive value (PPV) of the biomarker-based test. For example, the selection of biomarkers shall satisfy the restriction of a PPV above a threshold c_2 or an $1 - NPV$ below c_1. This can be achieved by restricting the priors of the variables on a subset of our choice. For example, non-informative priors can be set as

$$p_1 \sim Unif(0, c_1) \qquad \text{or} \qquad p_2 \sim Unif(c_2, 1)$$

2.2 Prior specification of the regression coefficients

Many prior specifications have been proposed that try to shrink small coefficients towards zero and retain in the model only coefficients with relevant larger effect sizes. Carvalho et al.[11,12] proposed the horseshoe prior (HS) for sparse signal detection, a hierarchical-shrinkage prior for the regression coefficients where the standard deviation is the product of a local (λ_j) and a global (τ) scaling parameter. It is given by

$$\begin{gathered} \beta_j|\lambda_j, \tau \sim Normal(0, \lambda_j^2\tau^2) \\ \lambda_j \sim Cauchy^+(0,1) \qquad \text{and} \qquad \tau \sim Cauchy^+(0,1) \end{gathered} \tag{2}$$

where the $Cauchy^+(0,1)$ is a standard half Cauchy distribution on the positive real numbers. The global shrinkage parameter tries to estimate the overall sparsity level, while the local shrinkage parameter is able to flag the non-zero elements of β.

The spike-and-slab (SpSl)[13,14] is a popular prior for sparse Bayesian estimation and is often written as a two-component mixture of Gaussians. Here we consider that the spike component concentrates its mass at zero whereas the slab component has its mass spread over a wide range of plausible values for the regression coefficients. Hence, we specify a spike and slab prior for β_j as $f(\beta_j) = (1 - \gamma_j) f_{spike}(\beta_j) + \gamma_j f_{slab}(\beta_j)$ and the prior inclusion probability $\gamma_j \sim Bernoulli(\pi)$ and $\pi \sim Unif(0,1)$. Regarding the choices for the f_{spike}, usually is taken to be a delta spike (Dirac spike) at the origin δ_0 and the f_{slab} a

normal density centered at zero. The prior for β_j can then be written as

$$\begin{aligned} \beta_j|\gamma_j, \sigma &\sim (1-\gamma_j)\delta_0 + \gamma_j Normal(0, \sigma^2) \\ \sigma^2 &\sim Inv - Gamma(a_s, b_s) \end{aligned} \tag{3}$$

The choice of the prior for the variance parameter is discussed in [15]. Here we considered $\sigma^2 \sim Inv - Gamma(1.5, 1.5)$ as mentioned in [16]. Instead of giving a continuous prior for the slab component (i.e. as we do in the horseshoe for the local shrinkage parameter λ_j), here the only values allowed are $\gamma_j = 0, 1$. The spike and slab prior is modeling the inclusion probability directly and therefore variable selection is based on the marginal posterior of γ_j as we discuss in more details in section 2.3.

Park and Casella [17] proposed the Bayesian Lasso by imposing the double exponential prior (DE) on the regression coefficients β with the density given by $f(\beta) = \prod_{j=1}^{d} \frac{\lambda}{2} e^{-\lambda|\beta_j|}$. When the shrinkage parameter λ has a gamma prior, the posterior mode of the double exponential prior has a straightforward interpretation since it corresponds to the frequentist Lasso estimates [18]. Several other choices have been proposed for the shrinkage parameter λ, as mentioned e.g. in Lykou et al. [19]. As discussed by Hans [20] and adopted in this paper, the choice for the hyperprior on the shrinkage parameter λ, belongs to the class of gamma priors. The model is specified as

$$\beta \sim DE(0, 1/\lambda) \qquad \text{and} \qquad \lambda \sim Gamma(c, d)$$

Under such continuous priors, like Laplace and HS the posterior probability of hitting the exact value zero is always zero, so some appropriate thresholding procedure needs to accompany the Bayesian procedure. The choice of the threshold is discussed in the following section 2.3.

2.3 Bayesian Variable selection and thresholding

To set up the Bayesian lasso variable selection, we consider the model given in (2) incorporating also the binary variable inclusion indicators $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_d)$ where $\gamma_j = 1$ indicates presence, and $\gamma_j = 0$ absence of covariate j, $j = 1, \ldots, d$) as suggested in [21] and [19].

The formulation for the risk score is now $Z^* = XD_\gamma\beta$, with $D_\gamma = diag(\gamma_1, \gamma_2, \ldots, \gamma_d)$ and variable selection will be based on the marginal posterior variable inclusion probabilities $f(\gamma_j = 1|y)$ for $j = 1, ..., d$, quantifying the posterior importance of each covariate. For the spike and slab prior, inclusion probabilities are calculated based on the marginal posterior $f(\gamma_j = 1|y)$ where $\gamma_j = 1$ when β_j is allocated to the slab component as formulated in (3).

For the horseshoe prior, which does not directly implement variable selection but shrinkage, one suggested threshold for variable selection is, to call β_j a signal (i.e. $\beta_j \neq 0$) if $\gamma_j \coloneqq 1 - \frac{1}{1+\lambda_j^2\tau^2} \geq 0.5$, and to call it noise otherwise[12]. For all the prior specifications, we estimate the posterior inclusion probability by the proportion of draws in the overall MCMC chain (excluding burn-in) that include the predictor β_j. Covariates are included in the model if the posterior inclusion probabilities are greater than 0.5 as suggested by Barbieri and Berger[1]. The authors provide arguments for the selection of the median probability model, defined to be the model containing all variables with $f(\gamma_j = 1|y) \geq \frac{1}{2}$, which was found to be the best model with regard to prediction performance.

2.4 Cutoff estimation and selection

After selection of the variables and estimation of the selected regression coefficients $\hat{\beta}$ we are interested in the posterior distribution of the biomarker cutoff cp in order make a good selection $\hat{cp}$, e.g. by computing the posterior mean. For this we could simply use the marginal of the joint posterior distribution. However, this distribution does not account for the specific choice of variables and regression coefficients that build the risk score to be used in future. Therefore, a better approach is to consider the conditional posterior distribution of cp given the now fixed set of variables and $\hat{\beta}$. This can be done using MCMC with Gibbs sampling considering $\hat{\beta}$ (with zeros for the de-selected variables) as given just like the data Y and X.

To summarize, we estimate the risk score and its cutoff in two steps: in the first step we apply the model (1) and we perform variable selection based on the marginal posterior inclusion probabilities $f(\gamma_j = 1|y) \geq \frac{1}{2}$. Then we derive an estimate for the coefficients from their marginal posterior distribution, e.g. the posterior means $\hat{\beta}$. At a second step, we apply the Bayesian procedure with the step model in (1) and same prior for p_1, p_2 and cp but

now for the fixed $\hat{\beta}$ (see Vradi et al.[10]). The resulting (marginal) posterior distribution of cp can now be used for the final choice of $\hat{cp}$, for instance, by taking the posterior mean.

3 Results

3.1 Simulation setting

In order to examine the performance of the proposed method, we conducted a simulation study, where interest lies in correct selection of regressors as well as incorrect selection of the unimportant variables. For each of the seven scenarios that we consider, we generated 1,000 data sets each with $n = 200$ individuals and varying number of predictors where we applied all the methods. We consider a classification problem with standard normally distributed predictors $X \sim MVN(0, \Sigma)$, where Σ is the covariance matrix. Let ρ denote the correlation between variables X_l, X_r where $l, r \in 1, \ldots, k, l \neq r$. The response y is a Bernoulli realization with probability p as it is defined in (2.1) for given p_1,p_2 and cut-off cp when a step model is used to generate the data, otherwise a logistic function is used with $logit(p) = X^T\beta$. The true model that was used to generate the outcome has k informative covariates $X_j, j = 1, ..., k$, $k \in \mathbb{N}$, $1 < k < d$. The noise variables are denoted as $m_1, \ldots, m_{d-k}$.

In scenario 1, we assume that there is only one informative predictor $(k = 1)$ in the dataset and $m = d - k = 5$ uninformative variables with $\beta = (2, \underbrace{0, ..., 0}_{5})$. In scenario 2, the data are generated using a logistic function for the probability of response and we assume that there are $m = 10$ uninformative predictors in the model and no important variables, i.e. the null model. In scenario 3, the total number of predictors is $d = 20$ with $m = 15$ uninformative and $\beta = (0.7, 1.5, 1, -2, -0.5, \underbrace{0, ..., 0}_{15})$. The correlation ρ between the informative predictors is 0.7. In scenario 4, there are $d = 15$ predictors with $k = 5$ informative ones and $\beta = (1.5, \underbrace{0.7, ..}_{2}, \underbrace{-1, ..}_{2}, \underbrace{0, ..., 0}_{10})$. The correlation ρ between the informative predictors is 0.5. For scenario 5, the data are generated

Table 1. Seven simulation scenarios for varying number of predictors, informative and non-informative, and varying correlation among the informative predictors. The true cut-off cp, and predictive values p_1 and p_2 that were used to generate the data are shown for scenarios 1, 3, 4 and 6 where we consider a step function as generating model. For scenario 2 and scenario 5, the true generating model assumes a logistic function.

Scenarios	Generating model	True cutoff cp	True p_1	True p_2	Number of Informative predictors (k)	Number of Non-Informative predictors (m=d-k)	correlation $\rho = corr(X_l, X_r)$, $l \neq r$
1	step	1.5	0.20	0.85	1	5	0
2 (Null model)	logistic	-	-	-	0	10	0
3	step	1.5	0.30	0.80	5	15	0.7
4	step	1	0.20	0.85	5	10	0.5
5	logistic	-	-	-	5	10	0.5
6 (2 stage)	step	1	0.20	0.85	5	10	0.5
7 (2 stage)	logistic	-	-	-	5	10	0.5

assuming a logistic function and the $k = 5$ informative predictors have effect sizes $\beta = (1.5, \underbrace{0.7, .., -2}_{2}, -0.5, \underbrace{0, ..., 0}_{10})$ and correlation $\rho = 0.5$.

An additional method that we consider for a comparison, is the two stage (2-stage) procedure described as follows; the algorithm firstly does variable selection assuming a logistic model and at the second stage, considering the selected variables from stage one, estimates the cut-off and predictive values by fitting the step function model. For this scenario, scenario 6, we consider the data in scenario 4 and fit the 2-stage algorithm. Similarly, in scenario 7, we implement the 2-stage approach on the data generated in scenario 5, where we assumed a logistic function to model the probability of response. The different simulation scenarios are summarized in Table 1.

Implementation of the methods was done in R version 3.4.3[22]. The analysis was done using Gibbs sampling with the library "R2jags"[23] together with the JAGS software[24]. For the MCMC, we used a burn-in of 2,000 iterations and keep the remaining 7,000 runs. We compared results for the three different priors, the Laplace, the horseshoe and the spike-and-slab prior. The 2-stage approach that we consider in scenario 6 and scenario 7, differs from the proposed method at the following important point: at the first stage the fitted model assumes a logistic function instead of the step function.

3.2 Simulation results

In this section we present results of the simulation study for all scenarios regarding i) posterior inclusion probabilities for the variables and ii) average number of True Positive (TP) and False Positive (FP) over the simulation runs of the covariates that are correctly and falsely included in the model. The TP and FP are defined as the number of important covariates correctly included in the model (i.e. variables with $\beta_j \neq 0$) and the number of covariates included in the model although their true effects are zero, respectively. The results presented here are for the constrained PPV in the interval $[0.8, 1]$ for all scenarios but 1 and 3. For scenario 1, we examined the behaviour of our method in the case that the lower bound of the constraint is equal to the true generating p_2, i.e the lower bound for the fitted model on the prior for p_2 was 0.85. For scenario 3, where the difference in the response rates is smaller than the other scenarios (i.e a higher p_1 and lower p_2), we took a lower bound for p_2 equal to 0.75 (such that it does not coincide with the true value of p_2).

Figure 1 shows the posterior inclusion probabilities over the simulation runs for all the variables in the model and for the three prior specifications (Laplace, HS and SpSl). We report the median together with the 1st and 3rd quartile. In Table 2, we present the average number of TP and FP over the simulation runs for the correctly and falsely included variables in the model for all scenarios and for all the different priors we considered. We expect the TP to be as close as possible to the number of informative predictors k. Only for the null model in scenario 2 (that there are no important variables in the model), the TP is the average number that no variable was selected (i.e. the average of the correctly selected null models) and the FP is the average that at least one variable was included in the model.

As we see in Figure 1, when the covariates are not correlated (scenario 1) the important predictors have always a very high posterior inclusion probability for all prior specifications and the non-informative variables have always a low inclusion probability (< 0.5), except the HS prior for which the median of the inclusion probabilities for the noisy predictors is close to the inclusion threshold. Specifically, in all scenarios, the Laplace prior results in higher posterior inclusion probabilities for the non-informative predictors, resulting in larger models. On the other hand, the SpSl prior has low inclusion probabilities for the noisy predictors, thus favoring sparser models, and further including the informative predictors

with high probability. As we see in Table 2, all the priors tend to behave similarly regarding TP in all scenarios. The Laplace prior results in higher number of FP, especially when the predictors are i) highly correlated (scenario 3) and ii) the data are generated from a logistic model and fitted with the step function (scenario 5).

In scenario 2, the HS has higher inclusion probabilities than the Laplace and SpSl, but all the priors have inclusion probabilities below the threshold value, i.e. correctly select the null model. All priors behave well in the case of the null model and under misspecification (i.e. data are generated by a logistic model and fit with a step model), resulting on average in high TP and low FP (see Table 2). For scenario 3 where there is high correlation among the informative predictors, some noisy variables are more frequently included. This seems to be especially true for the Laplace prior that tend to select more often uninformative variables. Overall, we see that the proposed method selects the correct model with high probability, i.e. high TP and low FP, for all the priors, however, some priors tend to include some more noisy variables than others.

[Insert Figure 1]

For scenario 5, where the fitted model (step function) is divergent from the true generating model (logistic function), we see that the important variables have a high probability of inclusion for the Laplace and SpSl priors, but the inclusion probability (for the important variables) is lower for the HS prior. We need to mention that with the Laplace prior we more often include noisy variables as with the other priors. When we compare it with the corresponding 2-stage approach in scenario 7 (where the fitted model (at the first stage) is the same as the true data generating model, i.e. a logistic function), we see that all the priors behave similarly for excluding the noisy predictors. However, for the 2-stage approach the variables X_2, X_3 and X_5 which have a weak effect, are not always included in the final model. Therefore the proposed method assuming the step function results in higher TP (i.e. including on average more important covariates in the model).

The performance of the model in terms of predictive values, was assessed on a validation dataset, where data were generated as described in section 3.1 under the different scenarios for a sample size of $\tilde{n} = 10,000$. Taking the estimated $\hat{\beta}$ and estimated cutoff $\hat{cp}$ (by applying the proposed method), we define as $\tilde{p_2} = P(\tilde{y} = 1|\tilde{X}^T\hat{\beta} > \hat{cp})$ and $\tilde{p_1} = P(\tilde{y} =$

Table 2. True Positives (TP) and False Positives (FP) over the 1,000 simulation runs for models we considered with different prior specifications, Laplace, Spike and Slab (SpSl) and Horseshoe prior (HS). TP is the average number of variables included in the model with $\beta_j \neq 0$, and FP is the average number of non-important variables selected. The second column in the table is the number of informative predictors (k) and non-informative predictors $(m = d - k)$ that were used to generate the data. For the null model in scenario 2, the TP are the average number that no variable was selected (the null model).

			TP			FP		
	k	**m=d-k**	**Laplace**	**SpSl**	**HS**	**Laplace**	**SpSl**	**HS**
Scenario 1	1	5	1	1	0.926	0.155	0.124	0.599
Scenario 2	0	10	0.879	0.943	0.849	0.121	0.057	0.151
Scenario 3	5	15	3.834	3.965	3.533	2.951	1.878	1.679
Scenario 4	5	10	4.951	4.933	4.641	1.449	0.718	0.376
Scenario 5	5	10	4.141	4.033	3.423	2.797	1.562	0.71
Scenario 6 (2-stage)	5	10	4.411	4.314	4.343	1.351	0.66	0.299
Scenario 7 (2-stage)	5	10	3.85	3.654	3.835	0.943	0.39	0.296

$1|\tilde{X}^T\hat{\beta} \leq \hat{cp})$, respectively, where $\tilde{X}$ is the matrix of the biomarker measurements of the validating set. In Figure 2, we present the boxplots of $\tilde{p_1}$, $\tilde{p_2}$ over the simulation runs. For scenario 2, we did not report any results, because in the case there are no biomarkers selected, then no cutoff is estimated and thus no classification is taking place.

[Insert Figure 2]

The bias of the predictive values calculated as $Bias(p_1) = \mathbb{E}(\hat{p_1} - \tilde{p_1})$ and respectively for p_2, is shown in Figure 3. As we see from the plot, the estimators of p_1 are nearly unbiased in all scenarios and for all priors. For p_2, we observe slightly higher bias especially with the HS prior. In Figure 3 we observe a tendency to slightly overestimate p_2 and underestimate p_1.

The classification ability of the model was assessed on the validation set described above and we use the Brier score[25] a measure of prediction accuracy that is calculated as

$$Brier = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} (\hat{p_i} - \tilde{y_i})^2$$

where $\hat{p} = \hat{p_2}$ if $\tilde{X}^T\hat{\beta} > \hat{cp}$ and $\hat{p} = \hat{p_1}$ otherwise, where $\tilde{X}$ is the matrix of the biomarker measurements of the validating set. In Figure 4 below we show the average Brier score over the simulation runs. As we see, in all scenarios the average Brier score is close to 0.18 for

the Laplace and SpSl and slightly higher for the HS prior. In the case of high correlation, the Brier score is close to 0.25 indicating a not desirable classification performance.

[Insert Figure 3]

[Insert Figure 4]

4 Real data application

We apply the proposed Bayesian method to a real dataset taken from a study as described in Vradi et al.[26]. The data consist of $d = 187$ protein measurements at baseline for $n = 53$ patients. The outcome is response to a treatment, $y \in [0, 1]$ andthe covariates are renamed as $X_1, \ldots, X_{187}$ to keep confidentiality. The aim of the analysis is to classify patients based on their biomarker measurements. The selection of the important variables satisfy the constrain that $p_2 \in (0.8, 1)$. We apply our proposed Bayesian method and compare the different priors, Laplace, SpSl and HS with the corresponding 2-stage approaches. For the cutoff and predictive values we used Uniform priors. We run the MCMC algorithm with Gibbs sampling for 100,000 iterations after a burn-in of 40,000 draws.

The Laplace prior selected 11 variables in the final model, whereas the SpSl prior resulted in the biggest model by including 78 proteins. The Horseshoe prior included 63 proteins in the final model and the 2-stage HS selected 72. In Figure 5 we present the posterior inclusion probabilities for the variables selected. We order the proteins selected by the SpSl prior according to their inclusion probabilities in decreasing order and keep the first top 10. Further we matched these 10 proteins to the variables selected by the other priors. We observed a significant overlap between the selected proteins with the four methods.

The lightest color on the heatmap represents lower inclusion probabilities and the darkest colors depict higher inclusion probabilities. In case that a variable was selected by the SpSl prior but was not selected by another prior, the heatmap has a white color, i.e. this is the case for the HS approach where only six variables where matched. In Figure 6 we plot the posterior median for the predictive values p_1, p_2 and the cutoff cp together with the 95% credible intervals. The different priors result in deviating number of selected variables (and estimated coefficients $\hat{\beta}$), and therefore the estimated cutoffs cp cannot be directly compared. The SpSl prior, which showed overall good performance in the simulation setting,

is the only case where the lower bound of the 95% credible interval of p_2 is different from the imposed lower bound of the constraint and thus p_2 is more informative.

[Insert Figure 5]

[Insert Figure 6]

5 Discussion

In clinical development, drugs are only clinically useful insofar as clinicians know for which patients the treatment should be used. The utility of novel biomarkers and clinical information for predicting patient prognosis and future outcome has great potential for advancing treatment decisions in personalized medicine. In most clinical applications, we face the problem of performing variable selection in order to reduce the complexity of the resulted model. However, the selection of biomarkers must be based on criteria of selecting those features that maximize either the positive or the negative predictive value. Together with the selection of the important variables in the model, a risk score and a cutoff value must be determined that will be used in practice for clinical decision making.

The Bayesian approach that we suggested for variable selection and cutoff estimation has the advantage that we can control the predictive values of the risk score. This is done by restricting the sample space to a range of desired values via the prior. The models presented in this article are based on the use of shrinkage or mixture priors and selection of the important variables is done *ad hoc* based on the posterior inclusion probabilities γ_j. In order to estimate the cutoff (and the predictive values), we fit the model assuming the step function but now at the second stage by fixing the estimated effects $\hat{\beta}$. We compared the proposed method with an alternative 2-stage approach, where at the first stage we fit a logistic model to select the informative variables and at a second stage we fit the step model for cutoff estimation.

Simulation results showed very good performance of the proposed method in terms of correct variables included in the model. Even when the fitted model (step model) is divergent from the true generating model (logistic model) the proposed method performs well, in terms of true positive rates of the variables. We found that when the true effect size is low (i.e. < 0.5), and especially in case of high correlation, predictors tend to not be selected in

the final model. The Laplace and SpSl prior behaved well in the scenario of the null model where very rarely selected any variable. Overall, across all scenarios and for the proposed method, we found that the SpSl prior behaved better that the other priors, in terms of high TP, low FP, low bias and good classification performance. We observed a tendency to slightly underestimate p_2 and overestimate p_1, but the bias was found overall small.

Obviously, the correlation structure of the covariates has an effect on the posterior inclusion probabilities. When regressors are correlated, all priors we compared tend to not identify all the important variables in the model, and some noisy variables are included more often. Thus, the classification error is also higher, compared to the situation when predictors are independent or there is a low correlation dependency. On the other hand, with correlated biomarkers, the selection of noisy biomarkers is not as severe as with independent biomarker, because the noisy ones will then mediate some effect of the correlated informative variables.

In this paper, we used as a risk score the linear predictor $Z = \eta = X^T\beta$, but of course, any transformation of Z can be considered. To avoid dependency on the scale upon which the continuous biomarker is measured, instead of using the raw values of the factor, we could standardize the scale by using the empirical distribution function (ecdf) of the risk score Z. For a vector $Z = (z_1, z_2, \ldots, z_n)$ the empirical distribution function is $F_n(Z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{z_i \leq z}$, with $\mathbb{1}$ the indicator function. For any $u \in (0, 1)$ we can define a biomarker positive result if $F_n^{-1}(u)$. Therefore, the assumption of a monotonic function for PPV and NPV is imposed to reflect the rationale for classifying subjects based on a binary classification rule. Instead of the ecdf, another proposal is to use the probit model with $P(y = 1|X) = \Phi(X^T\beta)$, Φ is the cumulative distribution function of the standard normal distribution.

Strong prior scientific knowledge is typically rare in biomarker identification problems, but when it is present (i.e. historical data or literature) it can be very easily incorporated. The new classification strategy is designed to guarantee a pre-determined predictive value (e.g. PPV of at least 80%). Attention should be paid in the case that the selection of a set of variables with pre-specified predictive values is not achievable due to the relationship between the risk score and the response. Our empirical results indicate that the method will

result in a posterior distribution of the restricted parameter that has a mode on the bound of the interval (here the lower bound). Hence, we expect that a data conflict with respect to the pre-specified minimal predictive value can be realized from the behaviour of the posterior.

Moreover, the choice of the shrinkage prior plays a crucial role in Bayesian variable selection, where many suggestions have been proposed to achieve sparsity, as for example, the horseshoe+[27], the R2-D2 prior[28], Dirichlet-Laplace priors[29] which belong within the class of global-local shrinkage priors. The behavior of the model under other choices of prior distributions as well as different choices on the hyperprior parameters is worth further exploration.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 633567 and is part of the IDEAS European training network (http://www.ideas-itn.eu/). This report is in part independent research arising from Prof Jaki's Senior Research Fellowship (NIHR-SRF-2015-08-001) supported by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

References

1. Barbieri MM, Berger JO et al. Optimal predictive model selection. *The Annals of Statistics* 2004; 32(3): 870–897.
2. van der Pas S, Szabó B, van der Vaart A et al. Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis* 2017; 12(4): 1221–1274.
3. Li Q, Lin N et al. The bayesian elastic net. *Bayesian Analysis* 2010; 5(1): 151–170.
4. Bernardo JM and Smith AF. *Bayesian theory*. IOP Publishing, 2001.
5. Swets JA. Roc analysis applied to the evaluation of medical imaging techniques. *Invest Radiol* 1979; 14(2): 09–121.
6. Hanley JA and McNeil BJ. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 1982; 143(1): 29–36.

7. Altman DG and Bland JM. Statistics notes: Diagnostic tests 2: predictive values. *BMJ* 1994; 309(6947): 102.
8. Fluss R, Faraggi D and Reiser B. Estimation of the youden index and its associated cutoff point. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 2005; 47(4): 458–472.
9. Linn S and Grunau PD. New patient-oriented summary measure of net total gain in certainty for dichotomous diagnostic tests. *Epidemiologic Perspectives & Innovations* 2006; 3(1): 11.
10. Vradi E, Jaki T, Vonk R et al. A bayesian model to estimate the cutoff and the clinical utility of a biomarker assay. *Statistical Methods in Medical Research* 0; 0(0): 0. DOI:10.1177/0962280218784778. PMID: 29966502.
11. Carvalho CM, Polson NG and Scott JG. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*. pp. 73–80.
12. Carvalho CM, Polson NG and Scott JG. The horseshoe estimator for sparse signals. *Biometrika* 2010; 97(2): 465–480.
13. Mitchell TJ and Beauchamp JJ. Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 1988; 83(404): 1023–1032.
14. George EI and McCulloch RE. Variable selection via gibbs sampling. *Journal of the American Statistical Association* 1993; 88(423): 881–889.
15. Gelman A et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis* 2006; 1(3): 515–534.
16. Li H and Pati D. Variable selection using shrinkage priors. *Computational Statistics & Data Analysis* 2017; 107: 107–119.
17. Park T and Casella G. The bayesian lasso. *Journal of the American Statistical Association* 2008; 103(482): 681–686.
18. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)* 1996; : 267–288.
19. Lykou A and Ntzoufras I. On bayesian lasso variable selection and the specification of the shrinkage parameter. *Statistics and Computing* 2013; 23(3): 361–390.
20. Hans C. Model uncertainty and variable selection in bayesian lasso regression. *Statistics and Computing* 2010; 20(2): 221–229.
21. Kuo L and Mallick B. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B* 1998; : 65–81.
22. Team RC. R: A language and environment for statistical computing [internet]. vienna, austria: R foundation for statistical computing; 2015, 2015.
23. Su YS and Yajima M. R2jags: A package for running jags from r. *R package version 003-08, URL http://CRAN R-project org/package= R2jags* 2012; .
24. Plummer M et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124. Vienna, Austria.
25. Brier GW. Verification of forecasts expressed in terms of probability. *Monthey Weather Review* 1950; 78(1): 1–3.

26. Vradi E, Brannath W, Jaki T et al. Model selection based on combined penalties for biomarker identification. *Journal of biopharmaceutical statistics* 2018; 28(4): 735–749.
27. Bhadra A, Datta J, Polson NG et al. The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis* 2017; 12(4): 1105–1131.
28. Zhang Y, Reich BJ and Bondell HD. High dimensional linear regression via the r2-d2 shrinkage prior. *arXiv preprint arXiv:160900046* 2016; .
29. Bhattacharya A, Pati D, Pillai NS et al. Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association* 2015; 110(512): 1479–1490.

Appendix

In Figure A1, we present the average brier score for all scenarios and the different priors for varying sample sizes of $n = 50, 100, 200, 500$. The classification ability of the model was assessed on a validation dataset, where data were generated as described in section 3.1 under the different scenarios for a sample size of $\tilde{n} = 10,000$. Taking the estimated $\hat{\beta}$ and estimated cutoff $\hat{cp}$ (by applying the proposed method), the Brier score is calculated as: $Brier = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} (\hat{p_i} - \tilde{y_i})^2$, where $\hat{p} = \hat{p_2}$ if $\tilde{X}^T \hat{\beta} > \hat{cp}$ and $\hat{p} = \hat{p_1}$ otherwise, where $\tilde{X}$ is the matrix of the biomarker measurements of the validating set. In the plot below we show the mean Brier score over the simulation runs.

[Insert Figure A1]

For scenario 2, we did not report the Brier score, because if there are no biomarkers, then no cutoff is estimated and thus no classification is taking place. For Scenario 3, where the correlation among informative predictors is high, we observe the highest error for all the priors. The HS prior resulted in the highest error in all scenarios, but the 2-stage approach. This result could be expected, as the HS prior did not perform well in selecting the true variables in the model (see Figure 1 and Table 2).

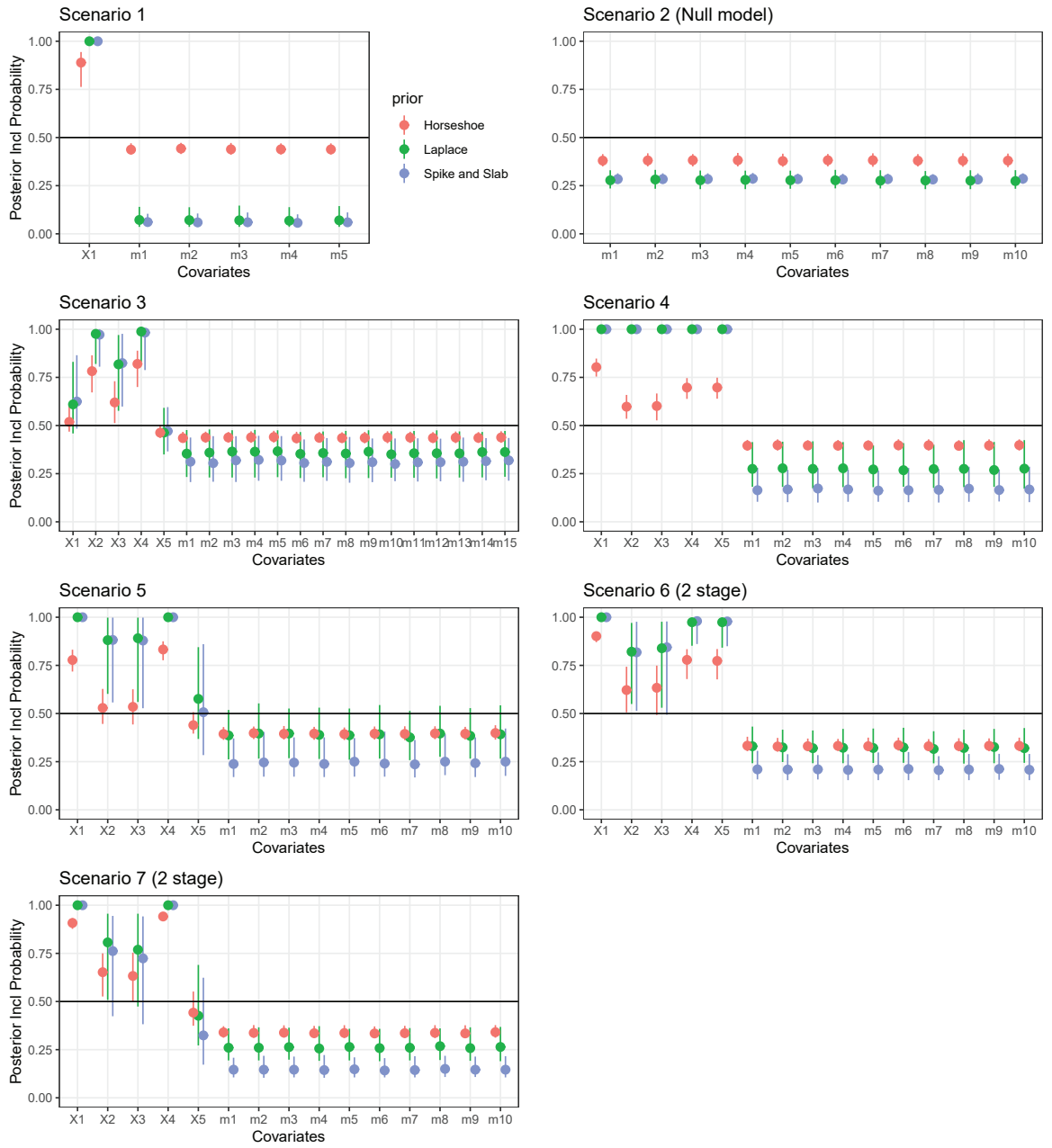

Figure 1. Posterior inclusion probabilities $f(\gamma_j|y)$ for all scenarios 1-7 over the 1,000 simulation runs. The dots are the medians of the posterior distribution together with the 1st and 3rd quantiles. In blue is the spike and slab prior, the Laplace prior in green and the horseshoe prior in red. On the horizontal axis are all the variables in the model, informative (X) and non-informative (m). The black horizontal line corresponds to the value 0.5 that is considered as threshold for including a variable in the model.

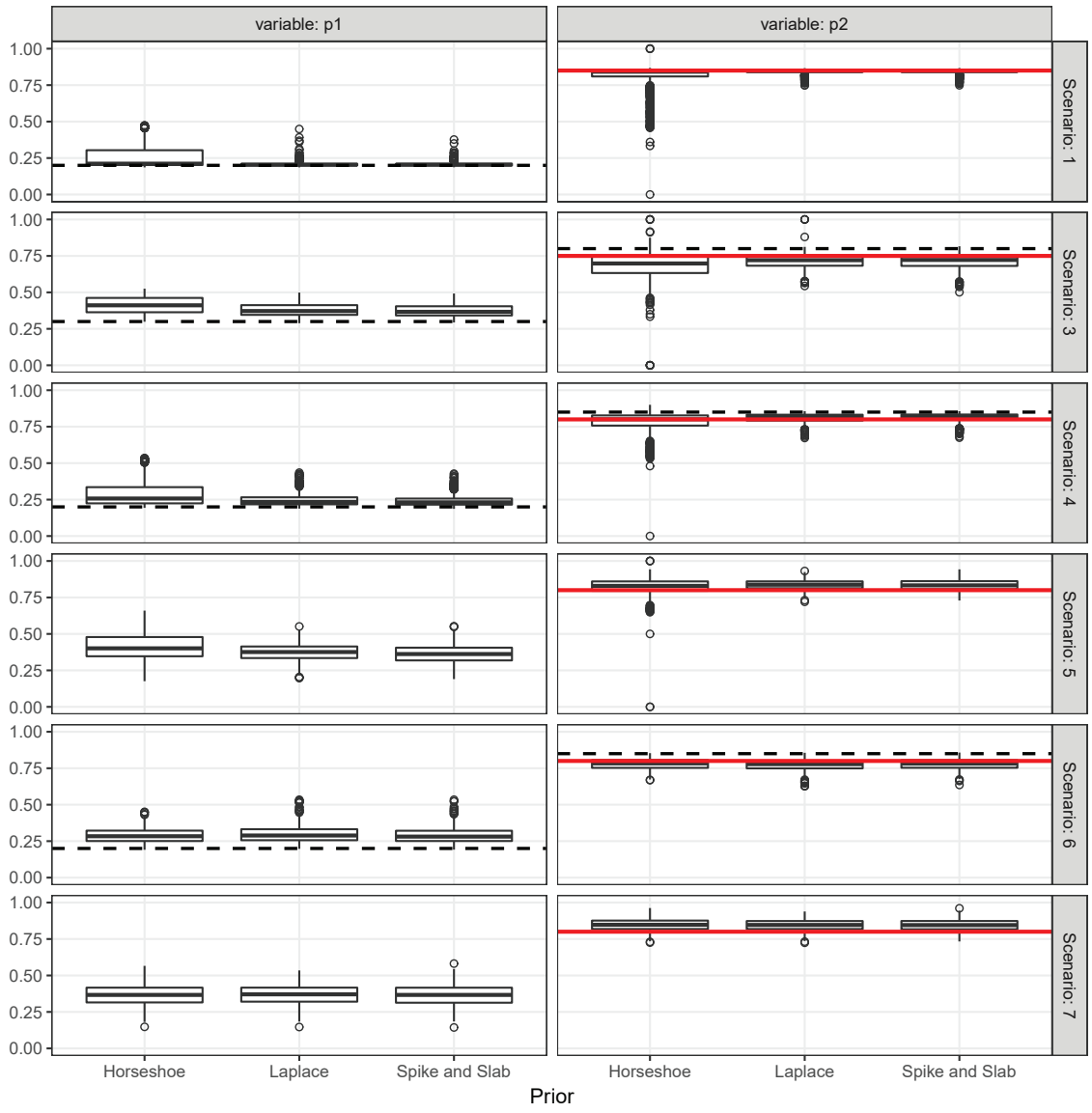

Figure 2. Boxplots of the $\tilde{p_1}$,$\tilde{p_2}$ (true values for the selected biomarkers scores) for all priors, HS (left boxplots) Laplace (middle boxplots), SpSl (right boxplots) and for all scenarios except the null model (scenario 2). The horizontal solid red line is the lower bound on p_2 that was used in the estimation procedure. The horizontal dashed black line is the value of p_1, p_2 that was used to generate the data. For scenario 5 and 7 the generating model uses a logistic function and therefore there is no generating p_1, p_2 are defined.

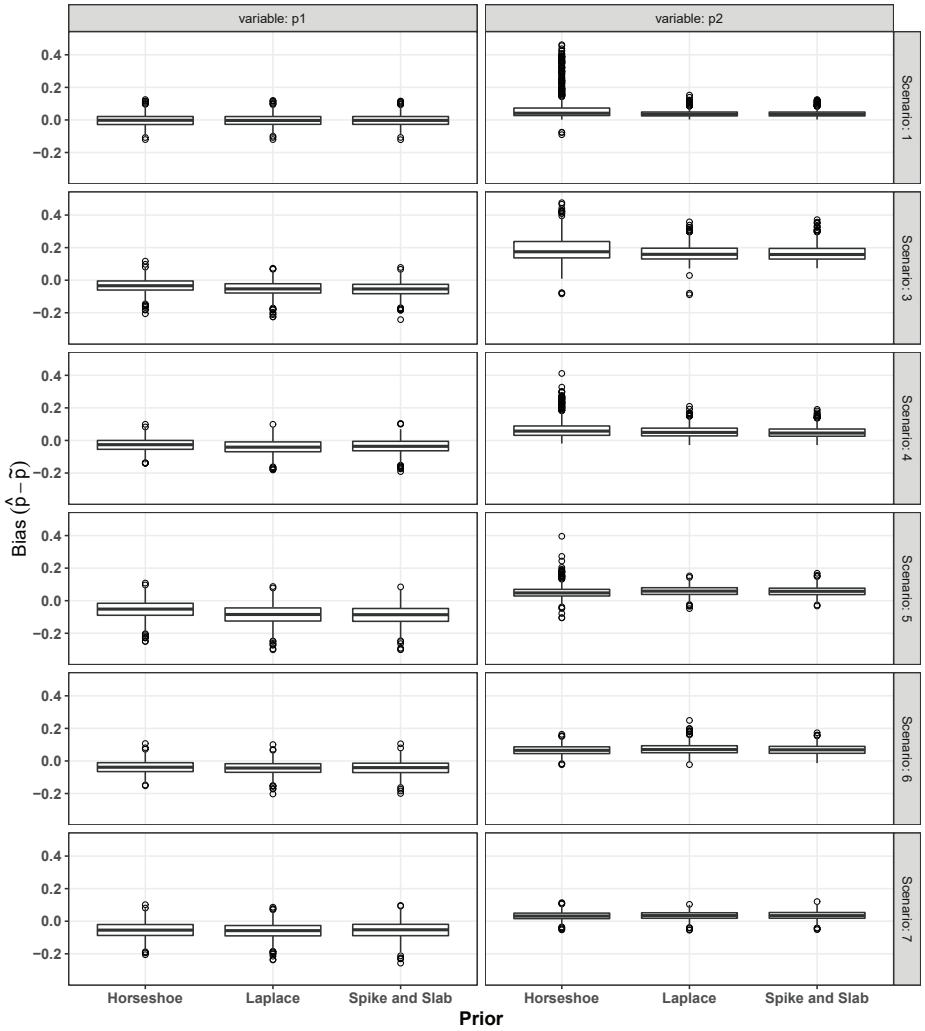

Figure 3. Boxplots of the $\hat{p_1} - \tilde{p_1}$ (left panel) and $\hat{p_2} - \tilde{p_2}$ (right panel) over the 1000 simulation runs for all priors, HS (left boxplots) Laplace (middle boxplots), SpSl (right boxplots) and for all priors and for all scenarios except the Null model (scenario 2).

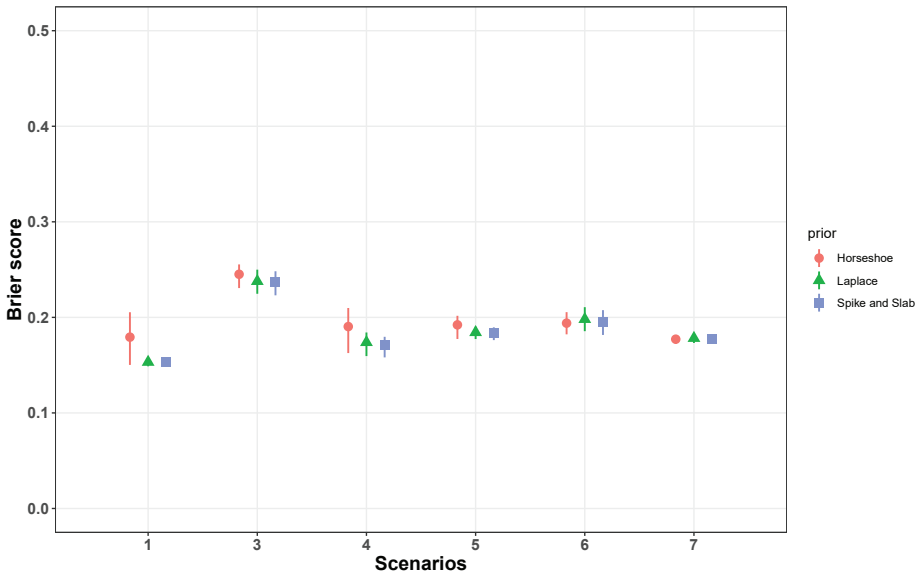

Figure 4. Plot of the average Brier score together with the 1st and 3rd quantile, over the simulation runs for all scenarios and for all priors, Laplace (green triangle), SpSl (blue square) and HS (red dot).

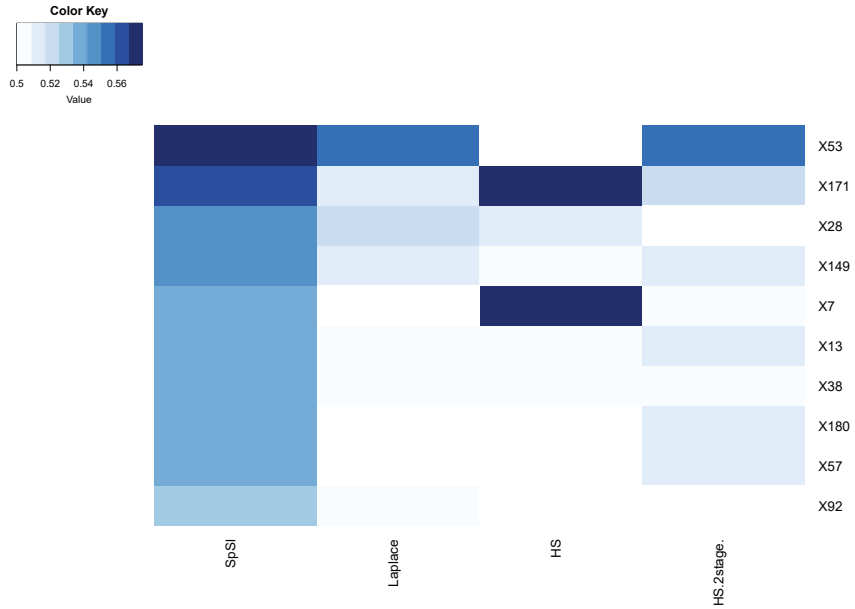

Figure 5. Heatmap of the inclusion probabilities of the top 10 variables selected. The selection of the top variables was done by ordering the inclusion probabilities of the variables selected by the SpSl in decreasing order and then matching these variables with the selected ones by Laplace, HS and HS (2-stage) prior. The Laplace (2-stage) and SpSl (2-stage) approaches resulted in selecting the null model. The white color on the heatmap indicates that some variables selected by the SpSl prior were not selected by the other priors.

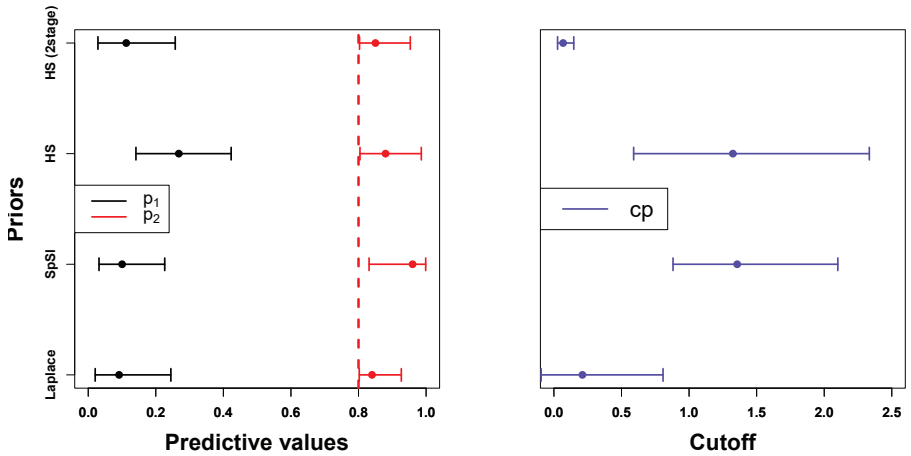

Figure 6. Plot of the posterior medians together with the 95% credible intervals for the predictive values p_1, and p_2 (left panel) and the cutoff cp (right panel). On the vertical axis are the different prior, Laplace, SpSl and HS. Results for the Laplace (2-stage) and SpSl (2-stage) approaches are not reported as the selected model was the null model. The vertical dashed red line is the lower constraint used on p_2.

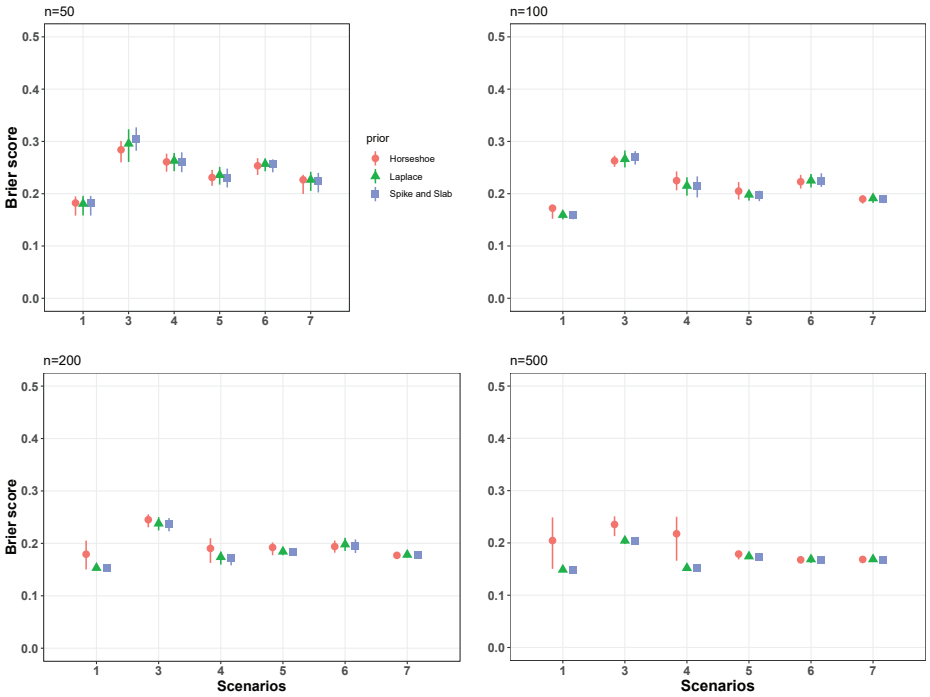

Figure 7. Plot of the average brier score together with the 1st and 3rd quantile, over the simulation runs for all scenarios and for all priors, Laplace (green triangle), SpSl (blue square) and HS (red dot) and for sample sizes of $n = 50, 100, 200, 500$.