Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2019, Vol. 28(8) 2538–2556

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

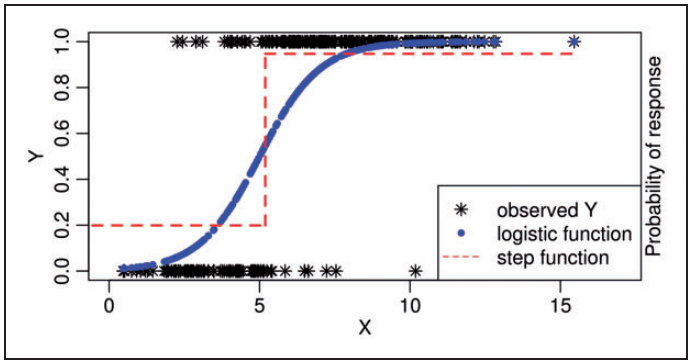

Figure 1. Plot of the observed biomarker X measurements for $y = 1$ and $y = 0$ (black stars). The blue dots depict the probability of response, p, when fitting a logistic model and the dashed red line shows p when it is modeled by a step function, and $p_1 = 0.17$, $p_2 = 0.95$ and $cp = 5.19$ are the 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	$\mu = 7$,	$\sigma^2 = 2$, $\beta_0 = -3$, $\beta_1 = 0.5$			(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.

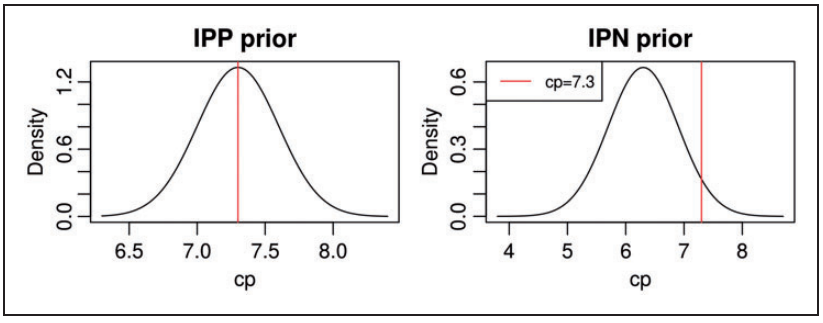

Figure 2. Density plots for the priors IPP and IPN. For the IPP prior, the true cutoff *cp* lies in a high probability region, while for the IPN prior, the true cutoff value lies on the tail of the distribution.

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

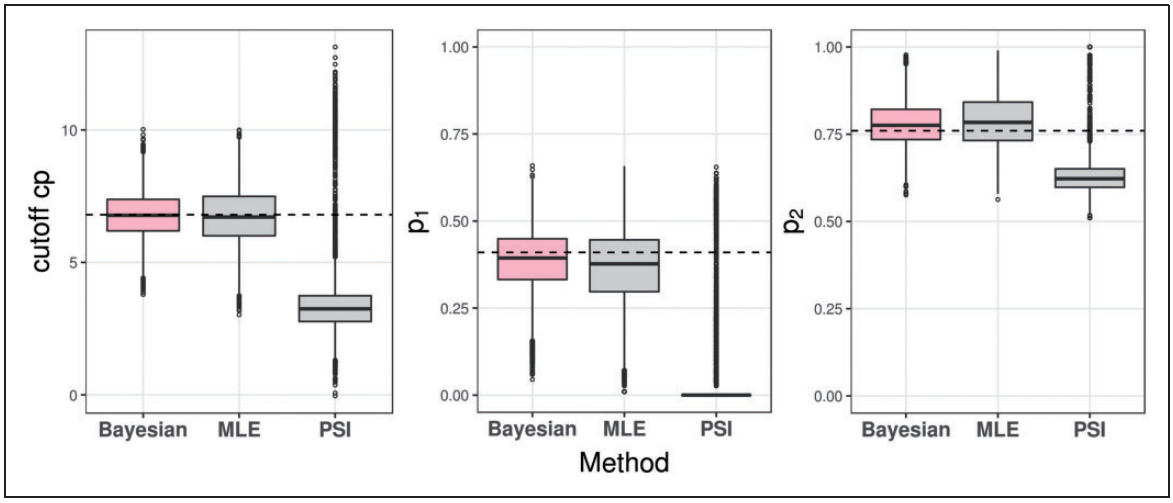

Figure 3. Bayesian posterior mean (left boxplots), MLE (middle boxplots) and PSI (right boxplots) estimators for the parameters cp (left panel), p_1 (middle panel), p_2 (right panel), over 10,000 simulation runs for scenario 5. The black horizontal dashed lines are the population parameters as calculated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

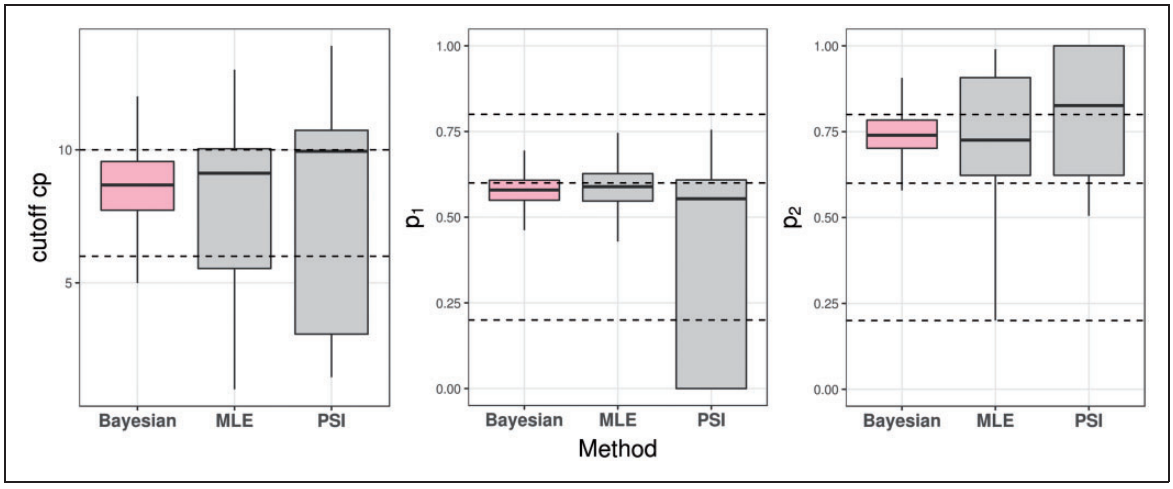

Figure 4. Boxplots of the Bayesian posterior mean (left boxplots), MLE (middle boxplots) and PSI (right boxplots) estimators for cp (left panel), p_1 (middle panel), p_2 (right panel), over 10,000 simulation runs for Scenario 6. The black horizontal dashed lines correspond to the true values of cp_1,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

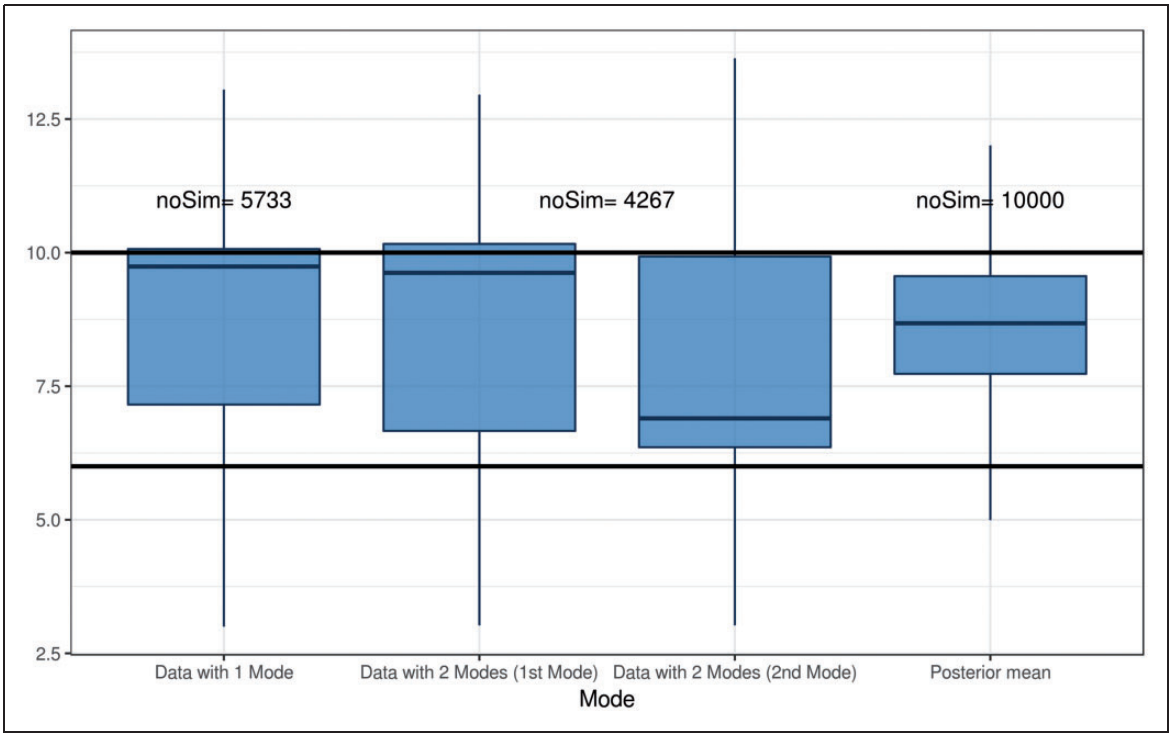

Figure 5. Distribution of the modes of the posterior distribution for the estimated cp, over 10,000 simulation runs for Scenario 6 estimated by the Bayesian model. If the posterior density is unimodal, then the only mode of the distribution is plotted (noSim = 5733) (left boxplot). In case the posterior 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

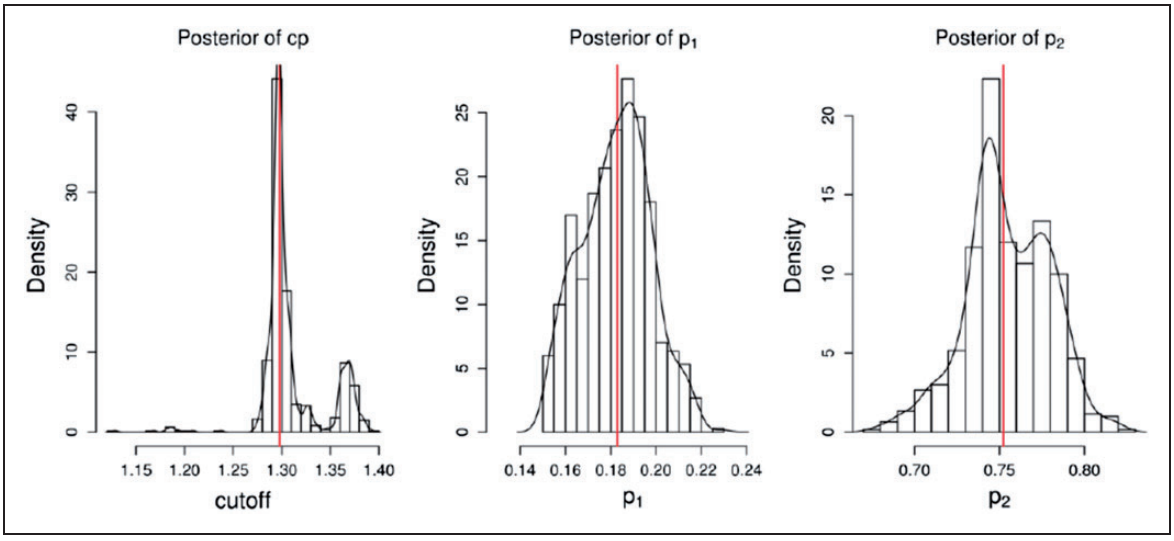

Figure 6. Plot of the posterior distribution for the parameter cp (left panel), p_1 (middle panel), p_2 (right panel) estimated by the Bayesian model. The red vertical lines denote the median of the distribution.

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 $X > cp$.

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

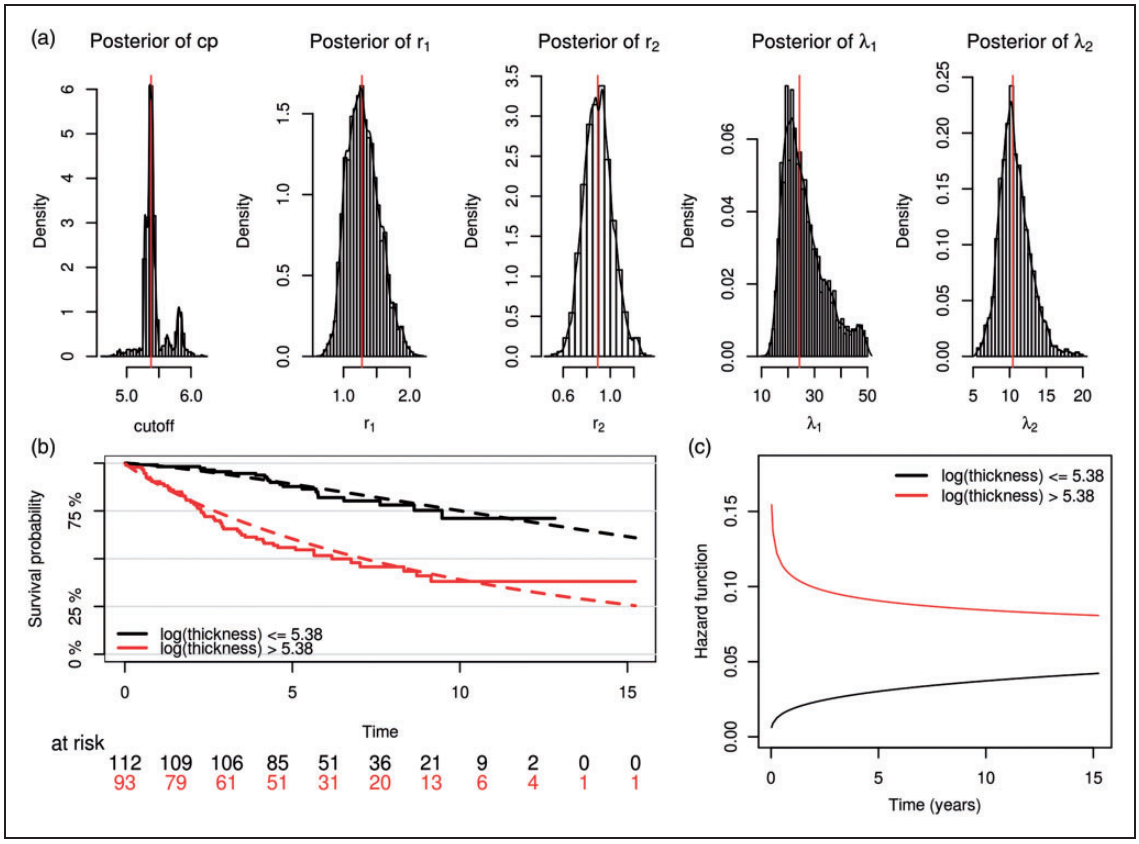

Figure 7. (a) Histograms of the posterior distributions for the cutoff (left panel), the shape parameters r_1 and r_2 (middle panels) and scale parameters λ_1 and λ_2 (right panels) for the Weibull model fitted to the melanoma data. The red vertical lines correspond to the posterior median of the distribution. (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 $h(t) = \begin{cases} \frac{r_1}{\lambda_1}(\frac{t}{\lambda_1})^{r_1-1}, & \text{if } X \le cp \\ \frac{r_2}{\lambda_2}(\frac{t}{\lambda_2})^{r_2-1}, & \text{if } X > cp \end{cases}$, 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

Supplemental material

Supplemental material for this article is available 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.

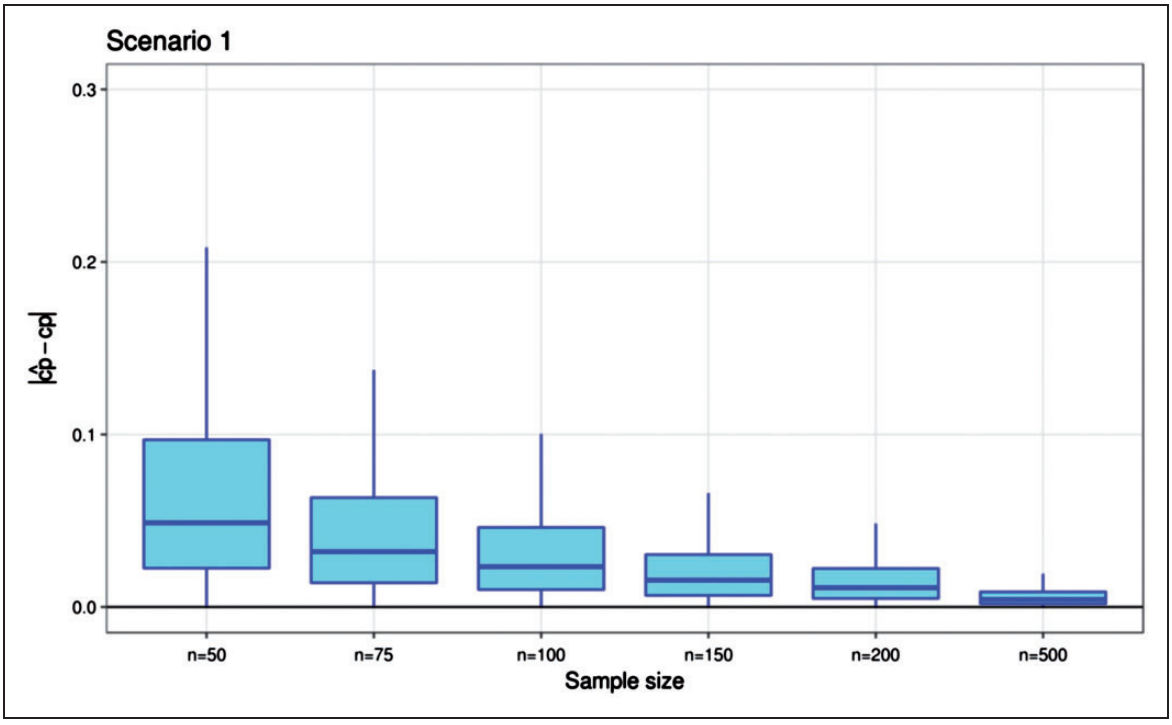

Figure 8. Boxplots of the absolute difference between the estimate and the true value of the cutoff *cp* over 10,000 simulation runs for Scenario 1 for varying samples sizes ($n = 50$, 75, 100, 150, 200, 500). Results shown for the Bayesian method with a uniform prior. The posterior mean was used as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

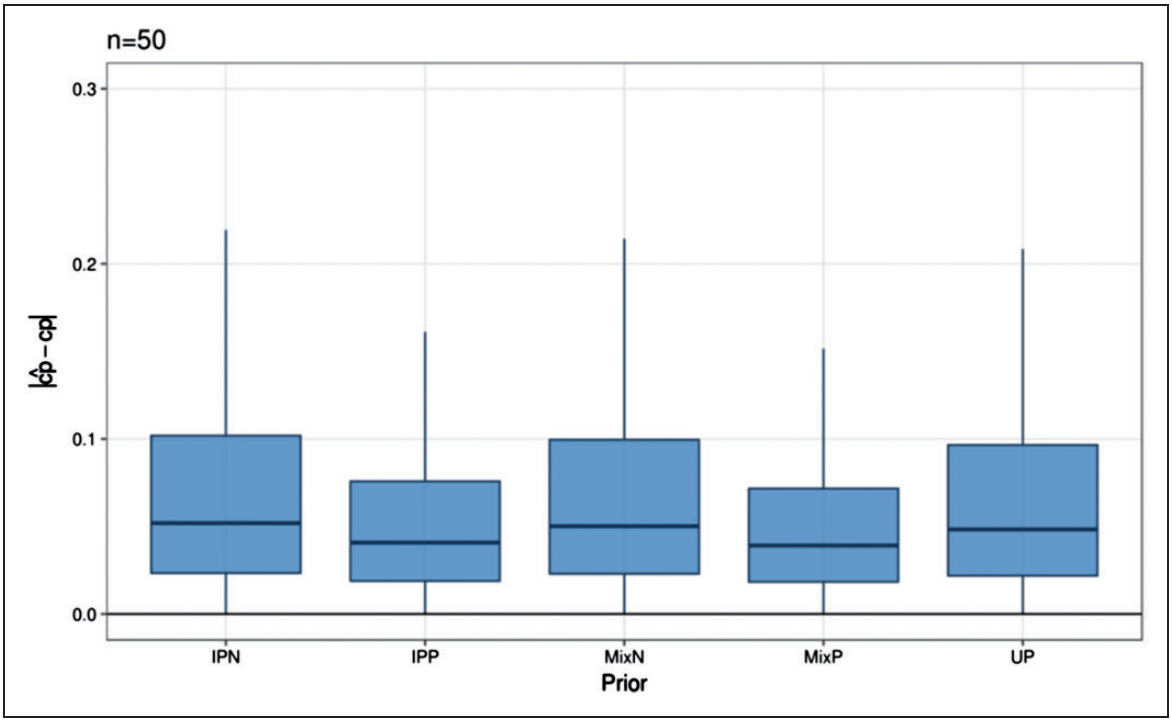

Figure 9. Boxplots for the absolute difference between the estimate $\widehat{cp}$ and the true value of *cp* estimated with the Bayesian model over 10,000 simulation runs for Scenario 1. In this simulation, we used $n = 50$ samples for the case of (from left to right) an Informative Prior Non-precise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

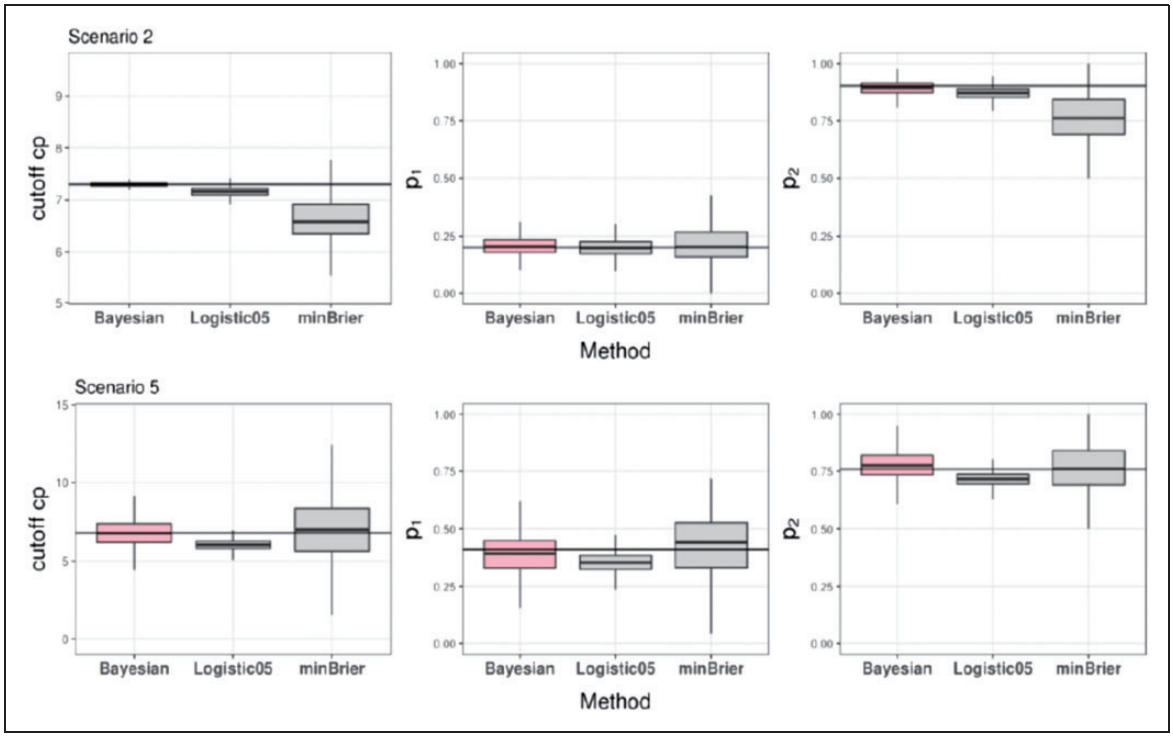

Figure 10. Boxplots of the estimated parameters cp, p_1, p_2 (left, middle and right plots, respectively) by the Bayesian method, the Logistic regression with a cutoff at $p = 0.5$ and by minimizing the Brier score. Results shown for 10,000 simulation runs for scenario 2 where the generating model is 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

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

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

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

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

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

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

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 $D^{(II)}_{KL}(g(y|x)||f(y|x))$ 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;
```