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

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 potential 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 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 $X = (X_1, X_2, \dots, X_n) \in \mathbb{R}$ denote the continuous biomarker measurements for *n* individuals and assume that *X* is available to be measured on all patients. Let $Y = (Y_1, Y_2, \dots, Y_n)$ denote the binary response variable, where $Y_i \in \{0,1\}$ for all $i = 1, \dots, n$. 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 Positive Predictive Value (PPV) is defined as the 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 Negative Predictive Value (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:

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

$$p(x) = P(Y = 1|X = x) = \begin{cases} p_1 = P(Y = 1|X \leq cp), & \text{for } x \leq cp \\ p_2 = P(Y = 1|X > cp), & \text{for } x > cp \end{cases} \quad (2.1)$$

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

[Figure 1 about here]

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.

$$p_1 \sim Unif(0,1) \quad \text{and} \quad p_2 \sim Unif(p_1, 1) \tag{2.2}$$

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 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:

$$cp = w\, g_1 + (1 - w)\, g_2 \qquad (2.3)$$

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 positive predictive value

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 are acceptable. In that case, the constraints can be controlled through priors, e.g.

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

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 cut-point that is as close as possible to achieve this constraint (i.e. the mass 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

$$f(cp, p_1, p_2|x, y) \propto L(p_1, p_2, cp|x, y) \times f(p_1) \times f(p_2) \times f(cp) \qquad (2.4)$$

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 (2.1) 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 are the observed values of X (see the appendix 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, 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 method by resampling the data $B = 500$ times, calculating the $\widehat{PSI_j}$ per sample $j = 1, \dots, B$. 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 Maximum Likelihood Estimator (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 (2.4). 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 the appendix.

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 $p(X) = \begin{cases} p_1, & if\ X \leq cp \\ p_2, & if\ X > cp \end{cases}$, whereas in scenario 5 the generating model is a logistic model with probability of response $p = \frac{e^{X\beta}}{1+e^{X\beta}}$.

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 the Appendix.

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 4th quartile to $X = 4$). Each component of the response Y is a realization from a Bernoulli random variable with $p(X) = \begin{cases} p_1, & if\ X = 1,2 \\ p_2, & if\ X \geq 3 \end{cases}$.

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 $p(X) = \begin{cases} p_1, & if\ X \leq cp_1 \\ p_2, & if\ cp_1 < X \leq cp_2 \\ p_3, & if\ X > cp_2 \end{cases}$. 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 $X \sim Normal(\mu = \mu_1, \sigma^2 = {\sigma^2}_1) + Normal(\mu = \mu_2, \sigma^2 = {\sigma^2}_2)$.

[Table 1 about here]

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 Table 2 and Table 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 Table 4 and Table 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 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 (2.2).

[Figure 2 about here]

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 2 about here]

[Table 3 about here]

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 A.1 in the Appendix. In Figure A.1, 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 Table A.1 and Table A.2 in the Appendix.

As shown in Table 3 and Table 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

304 method is far from the nominal (Table 5). The length of the credible interval (for the Bayesian
305 method) was similar to the confidence interval for the MLE and always narrower compared to PSI.
306
307 **[Table 4 about here]**
308 **[Table 5 about here]**
309
310 For scenario 5 where the true model is generated assuming a logistic response curve, we estimated the
311 cutoff and the corresponding probabilities of response by applying the Bayesian method as well as the
312 MLE and the PSI approaches. In that case, the true cutoff is not directly defined by the data generating
313 mechanism. However, the population parameters are defined by minimizing the KL divergence
314 between the true (logistic) and the assumed (step) model as discussed in section 2.1 and more detailed
315 in the Appendix. The results of the distribution of the estimates of the parameters for scenario 5 for the
316 three methods are shown in boxplots in Figure 3.
317
318 In this scenario, the Bayesian estimates are more consistent and have a smaller variability compared to
319 the MLE and the PSI method. As can be seen from the boxplots, the ML and the PSI methods result in
320 heavy tailed distributions for all the parameters and especially for the estimate of the cutoff. The
321 estimates concerning the cutoff and the predicted values obtained with the PSI method, differ
322 significantly as compared to the other two methods. This is partially due to the fact that the PSI
323 optimizes a different utility function than the Bayesian and the ML approach. While the Bayesian and
324 the ML methods use the likelihood as an objective function, the PSI method seeks to maximize the
325 difference between predictive values (PPV- (1-NPV))
326
327 **[Figure 3 about here]**

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 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 4 about here]

Figure 5 shows the distribution of the estimated cutoffs when posterior density is judged to be unimodal (5 733 out of 10 000 simulations) and when it is found to be a bimodal posterior distribution (4 267 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.

[Figure 5 about here]

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 [2] and [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 prior for the cutoff in the interval (0,100) and priors for the predictive values defined as in (2.2). 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 3.65 with 95% CI (3.62-3.69), while the posterior median was 3.66 with 95% credible interval (2.44-3.95). The PSI index which, we remind that maximizes a different objective function, estimates the optimal cutoff to be 37.66 with 95% bootstrapped CI (7.90-43.30). 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 were found equal to 0.17 with 95% credible interval (0.13-0.22) and 0.73 with 95% credible interval (0.61-0.79) 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).

[Figure 6 about here]

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 is 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 cut-off 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 that 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 $X > cp$.

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

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 bellow 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). At the same figure we plot the survival curves for the Weibull model in 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 $h(t) =$

$$\begin{cases} \frac{r_1}{\lambda_1}\left(\frac{t}{\lambda_1}\right)^{r_1 - 1}, & if\ X \le cp \\ \frac{r_2}{\lambda_2}\left(\frac{t}{\lambda_2}\right)^{r_2 - 1}, & if\ X > cp \end{cases}$$

, with $r_1, \lambda_1, r_2, \lambda_2$ taken as the means of the posterior densities.

[Figure 7 about here]

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.

As mentioned by a referee, one could alternatively use a standard classification algorithm, like for example logistic regression with a probability threshold of $p = 0.5$. One could also choose p such that the Brier score [26], a measure of accuracy of predictions, is minimized. 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 the Appendix.

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 6.3sec to run on a windows machine with processor Intel Xeon CPU E7-8867 v3 @ 2.5GHz, compared to frequentist approaches (MLE 0.15sec and for PSI 3.7sec 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 cut-offs (indicating the number of subgroups) the model can have according to the data.

Declaration of Conflicting Interests

The Authors declare that there is no conflict of interest.

References

1. Colburn WA. Biomarkers in drug discovery and development: from target identification through drug marketing. *The Journal of Clinical Pharmacology*. 2003 Apr 1;43(4):329-41

2. Pepe MS. The statistical evaluation of medical tests for classification and prediction. *Medicine*; 2003.

3. Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American journal of epidemiology*. 2006 Jan 12;163(7):670-5.

4. Schisterman EF, Perkins N. Confidence intervals for the Youden index and corresponding optimal cut-point. *Communications in Statistics Part B: Simulation and Computation*. 2007 May;36(3):549-563

5. Lunceford JK. Clinical utility estimation for assay cutoffs in early phase oncology enrichment trials. *Pharmaceutical statistics*. 2015 May 1;14(3):233-341.

6. Lever, J., Krzywinski, M. and Altman, N., 2016. Points of significance: Logistic regression

7. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* 1967 Jun (Vol. 1, No. 1, pp. 221-233).

8. Bunke O, Milhaud X. Asymptotic behavior of Bayes estimates under possibly incorrect models. *The Annals of Statistics*. 1998;26(2):617-44.

9. Kullback S, Leibler RA. On information and sufficiency. *The annals of mathematical statistics*. 1951 Mar 1;22(1):79-86.

10. Schmidli H, Gsteiger S, Roychoudhury S, O'Hagan A, Spiegelhalter D, Neuenschwander B. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*. 2014 Dec 1;70(4):1023-32.

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

12. Youden WJ. Index for rating diagnostic tests. Cancer. 1950 Jan 1;3(1):32-5.

13. Chu H, Cole SR. Estimation of risk ratios in cohort studies with common outcomes: a Bayesian approach. *Epidemiology*. 2010 Nov 1;21(6):855-62.

14. Wagenmakers EJ, Lee M, Lodewyckx T, Iverson GJ. Bayesian versus frequentist inference. InBayesian evaluation of informative hypotheses 2008 (pp. 181-207). Springer, New York, NY.

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 Apr;57(1):97-109.

19. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. The journal of chemical physics. 1953 Jun;21(6):1087-92

20. Ibrahim JG, Chen MH, Sinha D. *Bayesian survival analysis*. John Wiley & Sons, Ltd; 2005 Jul.

21. Hartigan JA, Hartigan PM. The dip test of unimodality. *The annals of Statistics*. 1985 Mar 1:70-84.

22. Etzioni R, Pepe M, Longton G, Hu C, Goodman G. Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Medical Decision Making*. 1999 Aug;19(3):242-51.

23. Broemeling LD. *Advanced Bayesian methods for medical test accuracy*. CRC Press; 2016 Apr 19.

24. Martinussen T, Scheike T. Dynamic regression models for survival analysis. Statistics for biology and health. Springer. 2006

25. Chivers C. MHadaptive: general Markov Chain Monte Carlo for Bayesian inference using adaptive Metropolis-Hastings sampling.

26. Brier, G., 1950. Verification of forecasts expressed in terms of probability. *Mon Wea Rev.*, pp. 78;1-3.