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

Appendix

I. 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 A.1.

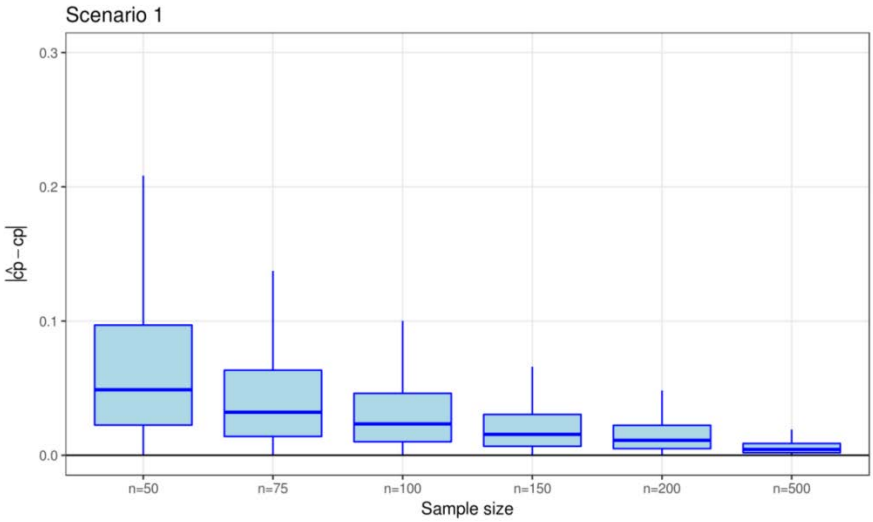

Figure A.1: Boxplots of the absolute difference between the estimate and the true value of the cutoff *cp* over 10 000 simulation runs for Scenario 1 for varying samples sizes *(n=50, 75, 100, 150, 200, 500)*. Results shown for the Bayesian method with a uniform prior. The posterior mean was used as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

In Table A.1 and A.2 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 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.

p_1, p_2	Bias						
Methods	**Bayesian**					**PSI**	**MLE**
Prior	UP	IPN	IPP	MixN	MixP		
Scenario 1							
p_1	$3x10^{-2}$	$3x10^{-2}$	$3x10^{-2}$	$3x10^{-2}$	$7x10^{-3}$	$4x10^{-2}$	$4x10^{-2}$
p_2	$-3x10^{-2}$	$-4x10^{-2}$	$3x10^{-2}$	$-3x10^{-2}$	$-8x10^{-3}$	$9x10^{-2}$	$8x10^{-2}$
Scenario 2							
p_1	$3x10^{-2}$	$2x10^{-2}$	$2x10^{-2}$	$3x10^{-2}$	$2x10^{-2}$	$-3x10^{-2}$	$6x10^{-2}$
p_2	$-4x10^{-2}$	$-4x10^{-2}$	$-3x10^{-2}$	$-3x10^{-2}$	$-3x10^{-2}$	$-4x10^{-3}$	$5x10^{-2}$
Scenario 3							
p_1	$2x10^{-2}$	$2x10^{-3}$	$2x10^{-2}$	$1x10^{-2}$	$2x10^{-2}$	$-7x10^{-3}$	$6x10^{-2}$
p_2	$-5x10^{-2}$	$-1x10^{-1}$	$-5x10^{-2}$	$-6x10^{-2}$	$-5x10^{-2}$	$8x10^{-2}$	$1x10^{-1}$
Scenario 4							
p_1	$4x10^{-2}$	$4x10^{-2}$	$4x10^{-2}$	$4x10^{-2}$	$4x10^{-2}$	$3x10^{-3}$	$1x10^{-1}$
p_2	$-2x10^{-2}$	$-2x10^{-2}$	$-2x10^{-2}$	$-2x10^{-2}$	$-2x10^{-2}$	$6x10^{-3}$	$5x10^{-2}$

Table A.1: 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	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	0.956	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

Table A.2: 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*. 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.

In Figure A.2 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.

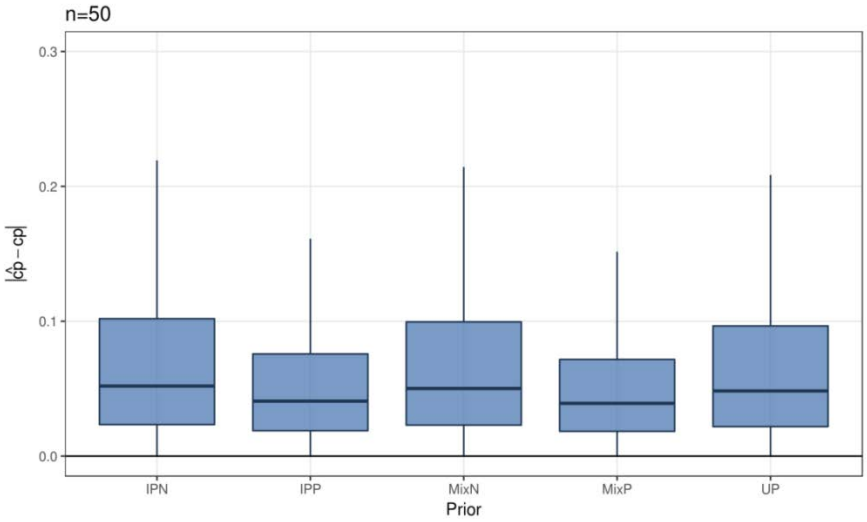

Figure A.2: Boxplots for the absolute difference between the estimate $\widehat{cp}$ and the true value of cp estimated with the Bayesian model over 10 000 simulation runs for Scenario 1. In this simulation we used $n=50$ samples for the case of (from left to right) an Informative Prior Non-precise (IPN), an Informative Prior Precise (IPP), a Mixture Prior Non-precise (UP+IPN), a Mixture Prior Precise (UP+IPP) and a Uniform Prior (UP).

II. 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 the Figure A.3 below, where we see that the estimated parameters by the logistic model with the choice of $p = 0.5$ are more biases 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 underestimate the true parameters.

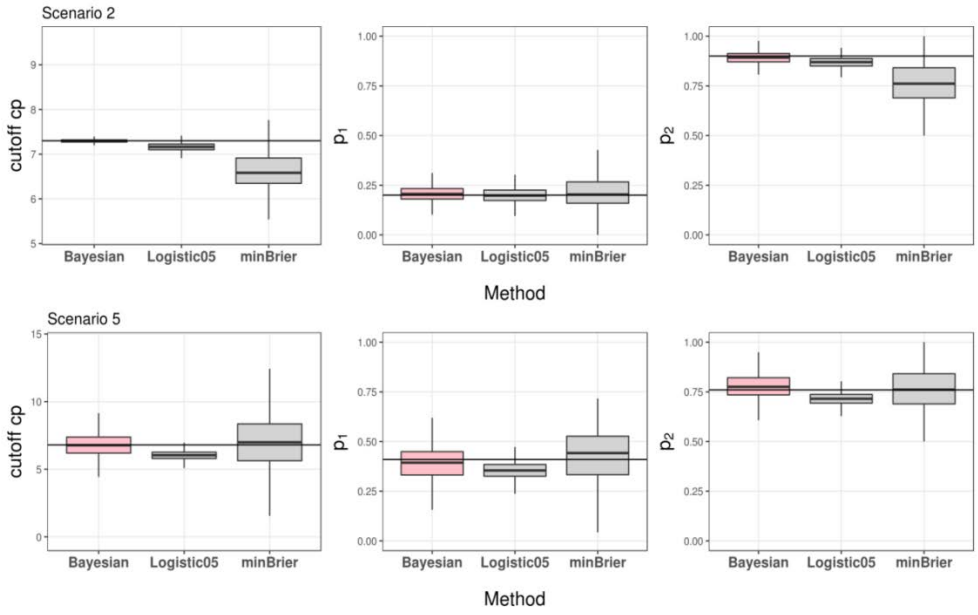

Figure A.3: Boxplots of the estimated parameters cp, p_1, p_2 (left , middle and right plots respectively) by the Bayesian method, the Logistic regression with a cutoff at $p = 0.5$ and by minimizing the Brier score. Results shown for 10,000 simulation runs for scenario 2 where the generating model is 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.

III. Conditional Kullback-Leibler divergence between the theoretical and fitted model

A. Estimation of the predictive values

Let's 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 $Y|X \sim Bernoulli(q), q(x) = \begin{cases} q_1, & if\ x \leq cp \\ q_2, & if\ x > cp \end{cases}$ 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 $\int_{Y\in B} g(y|x) log \frac{g(y|x)}{f(y|x)}\, dy =$

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

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

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

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

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

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

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

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

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

set equal to zero and solve with respect to q_1 we then obtain

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

Following the same calculations for $D_{KL}^{(II)}(g(y|x)||f(y|x))$ and solve with respect to q_2 we get

$$q_2 = \frac{\int_{cp}^{\infty} g(x)p(x)\ dx}{\int_{cp}^{\infty} g(x)\ dx}$$

B. 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 the $D_{KL}(g(y|x)||f(y|x))$.

IV. 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;
```