Supplementary Material to "Estimating the Design Operating Characteristics in Bayesian Adaptive Clinical Trials"

Shirin Golchi[1]

[1]Department of Epidemiology and Biostatistics, McGill University

A Cross-validated estimates

In Section 4.3 of the manuscript a leave-one-out cross-validation approach is described to assess the estimation performance of the proposed model on the ordinal scale endpoint/PO model application. Figure S.1 shows the cross-validated point estimates and 95% credible intervals for the 80 points that were used to train the proposed Beta/GP. As a reminder the estimates for each point in Figure S.1 is obtained from the other 79 points, i.e., by excluding the corresponding point from the training set.

B Simulation study

For the simulation study described in this section, we consider a simple but common scenario within the clinical trials framework where the goal is to evaluate the effect of a treatment in reducing (or increasing) the odds of an event. In such a setting, a Beta-binomial model may be used where the event risk is assigned beta priors under each arm. Together with a binomial likelihood the posterior distribution for the probability of event is obtained as

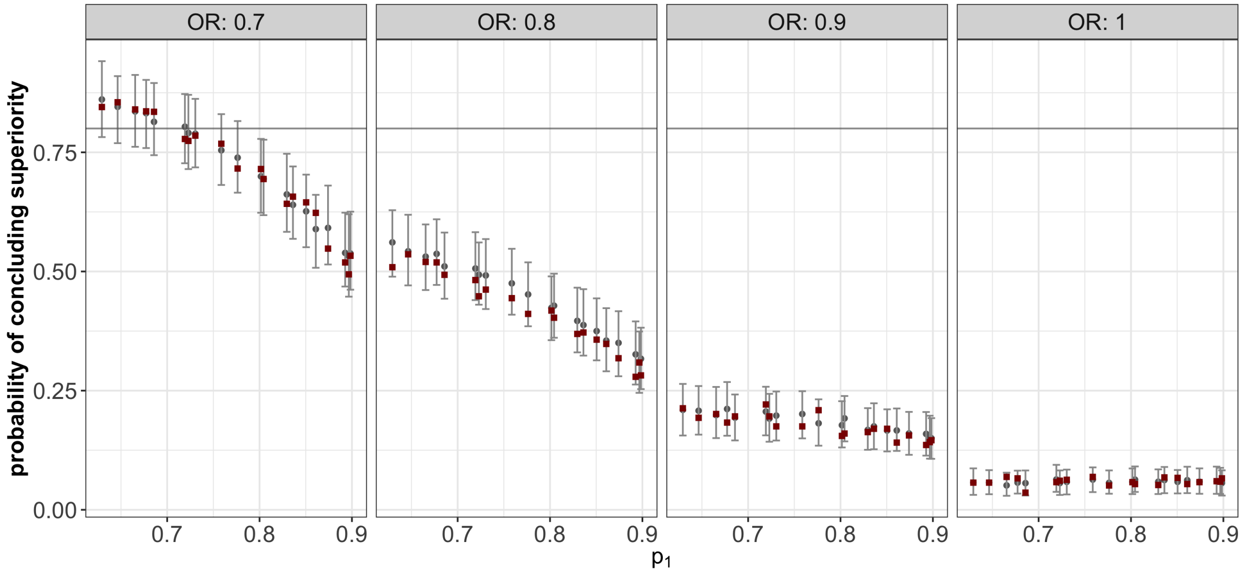


Figure S.1: Cross-validated point estimates (round dots) and 95% credible intervals together with the "true" probabilities of stopping early obtained from simulation.

a Beta distribution. The posterior distribution for OR is respectively obtained by drawing samples from the risk posterior distributions under each arm.

In the conjugate framework, obtaining the test statistic that is derived from the posterior does not require MCMC sampling for each simulated trial. The analytic posterior results in much more efficient simulations of the sampling distribution. Therefore, we can simulate a fixed sample trial design and estimate the probability of success (superiority) over a wide range of parameter values. As the test set, we define a fine grid of size 100 over the parameter space given by the Cartesian product of $(0.25, 0.7)$ (for the base event risk p_0) and $(0.65, 1)$ (for the OR). We then simulate the sampling distribution of the posterior probability of effectiveness by simple Monte Carlo for each of the (p_0, OR) pairs over this grid. The probability of success with a fixed sample is obtained for each set of parameter values as the upper 5% tail of the sampling distribution, estimated as the 95% sample quantile.

For the simulation design, one hundred training sets are generated using the methods described in Section 4.1. Specifically, at every iteration, 100 points are generated over the two dimensional parameter space $\Theta = (0.25, 0.7) \times (0.6, 1)$ and a training set of size 20 is constructed using the clustering method of Lekivetz and Jones (2014). The trial is simulated over these 20 points and GP models are trained over the outputs obtained for the 20 sets of parameter values.

The simulation outputs are therefore defined to capture the performance of the proposed method in predicting power over the test set. We compute root mean squared error (RMSE), bias and posterior standard deviation (PSD) for any point θ in the test set and for each of the 100 training sets, as follows:

$$\text{RMSE} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\phi_k - \phi_t)^2}$$

where $\phi_k = P(\pi > 0.95; a_k(\theta), b_k(\theta))$ is the estimate of power obtained as the upper tail of a beta distribution with parameters given as the k^{th} posterior samples $a_k(\theta)$ and $b_k(\theta)$; and ϕ_t is the "true" power obtained by trial simulation at θ;

$$\text{bias} = \hat{\phi} - \phi_t,$$

where the posterior mean is used as the point estimate of power, $\hat{\phi} = \frac{1}{K}\sum_{k=1}^{K}(\phi_k)$;

$$\text{PSD} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\phi_k - \hat{\phi})^2}.$$

Figures S.2-S.4 show the above measures averaged over the 100 training sets throughout the parameter space. The RMSE is on average below 4.5% and only increases for extreme values of the base risk, p_0 (small effect size and low power). Similarly, averaged over the design, absolute bias does not exceed 0.04, the largest bias occurs at the two ends of the range of p_0 values. As for posterior standard error, the highest precision is achieved where the effect size is relatively large $OR < 0.8$ with the posterior variance increasing under the null hypotheses and for small/large values of p_0.

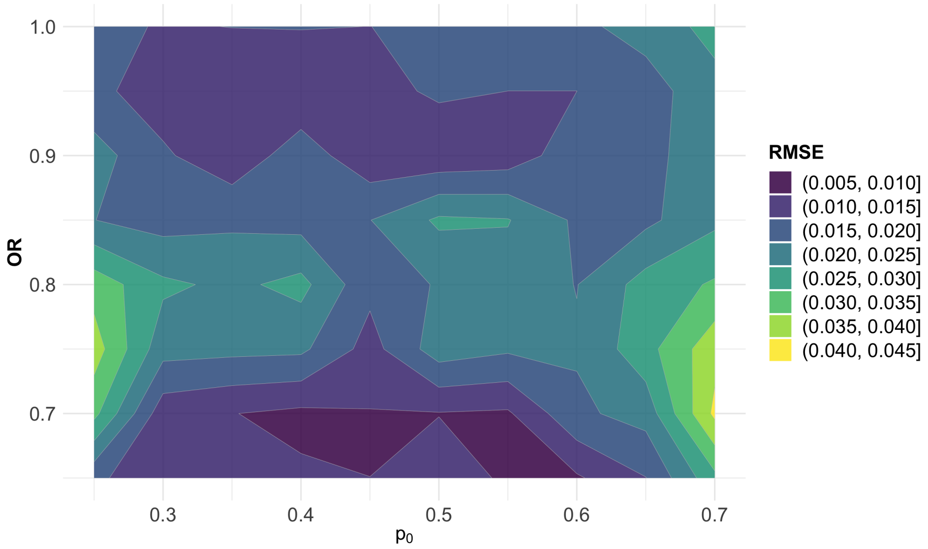


Figure S.2: Root mean squared error averaged over 100 training sets over the input space.

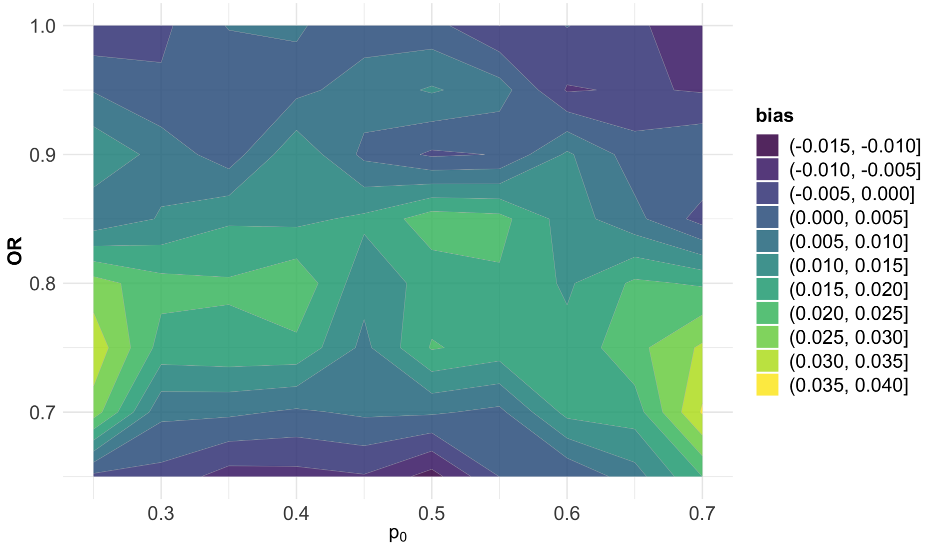


Figure S.3: Bias averaged over 100 training sets over the input space.

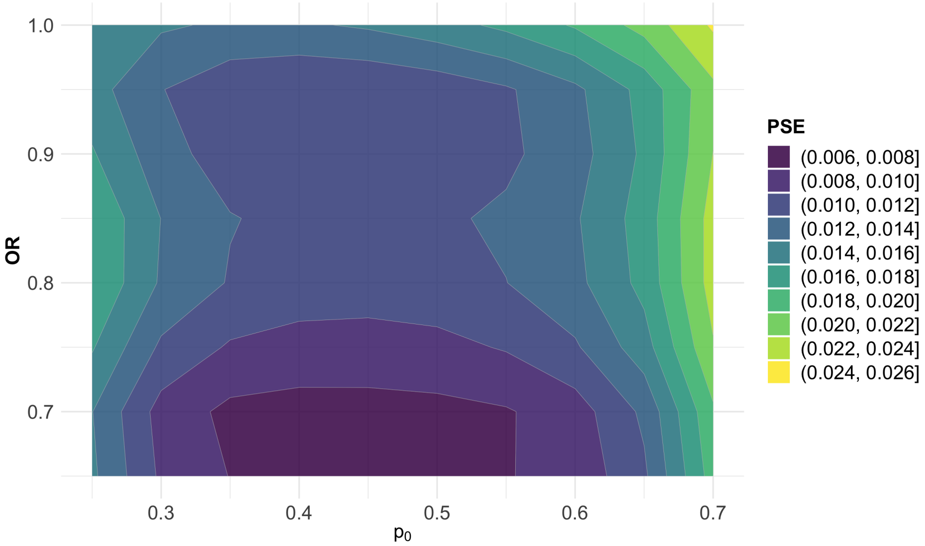


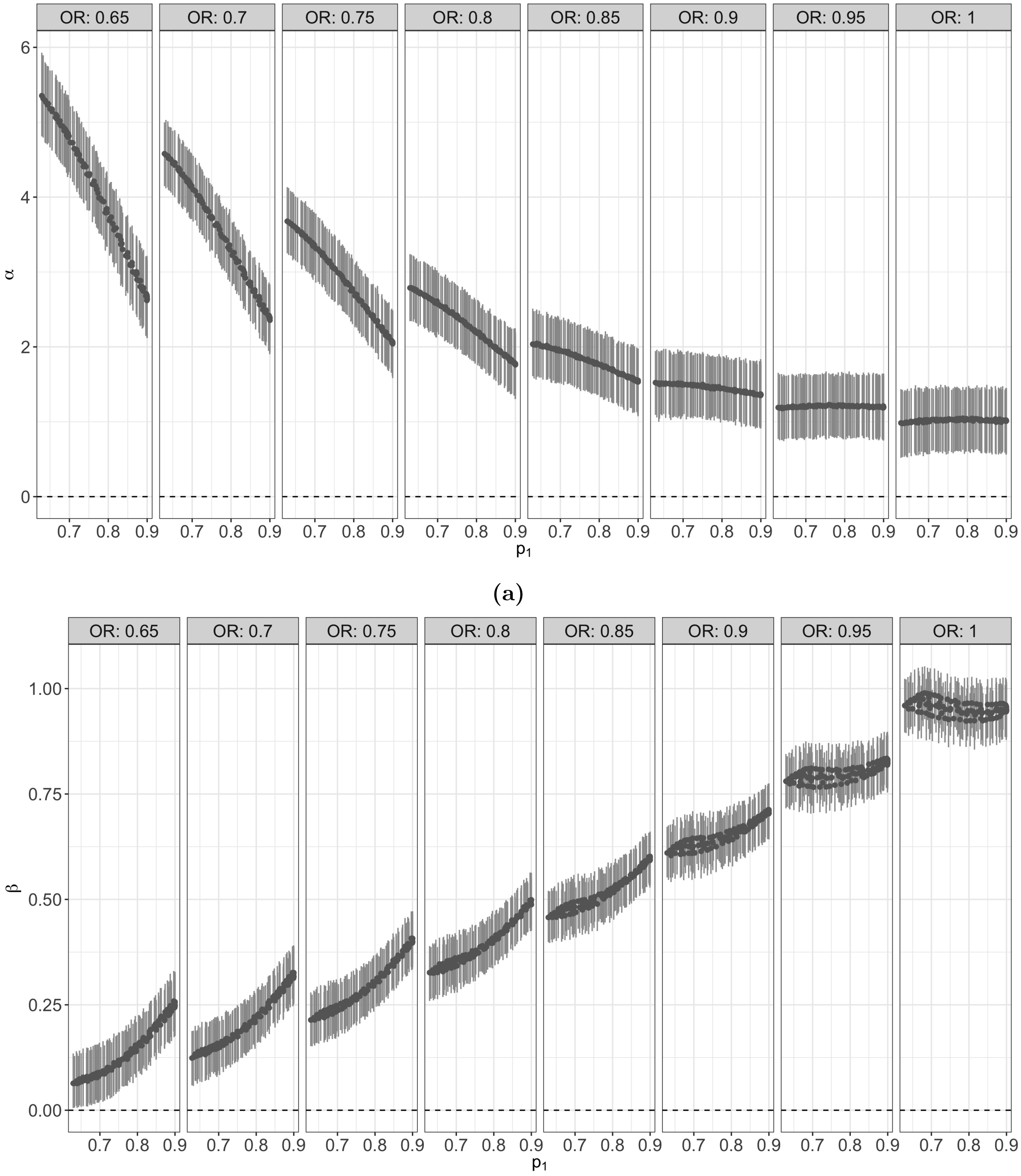
Figure S.4: posterior standard deviation averaged over 100 training sets over the input space.

C Post-hoc rejection sampling from the posterior GP's

As discussed within the manuscript, the post-hoc rejection sampling scheme to assure that the $a(\theta)$ and $b(\theta)$ estimates obtained from the GP posteriors in (5) are in fact positive is in fact very efficient for the present application. The rejection rate is in fact zero for most parameter configurations meaning that negative values are assigned practically zero probability under the posterior. This is illustrated in Figure S.5 that shows the posterior mean and 95% credible intervals for $a(\theta)$ and $b(\theta)$ across Θ.

References

Lekivetz, R. and Jones, B. (2014). Fast flexible space-filling designs for non-rectangular regions. *Quality and Reliability Engineering*, 31.



(a)

(b)

Figure S.5: Point estimates and 95% credible intervals for the (a) shape and (b) scale parameters of the Beta distribution in (3)