Utilizing Bayesian Predictive Power in Clinical Trial Design

Ofir Harari[1], Grace Hsu[1], Louis Dron[1], Jay J. H. Park[1,2], Kristian Thorlund[1,3], and Edward J. Mills[1,3]

[1]*Cytel Canada, Vancouver, BC, Canada*

[2]*Experimental Medicine, Department of Medicine, University of British Columbia, Vancouver, BC, Canada*

[3]*Department of Health Research Methodology, Evidence, and Impact, McMaster University, Hamilton, ON, Canada*

August 25, 2020

Abstract

The Bayesian paradigm provides an ideal platform to update uncertainties and carry them over into the future in the presence of data. Bayesian Predictive Power (BPP) reflects our belief in the eventual success of a clinical trial to meet its goals. In this paper we derive mathematical expressions for the most common types of outcomes, to make the BPP accessible to practitioners, facilitate fast computations in adaptive trial design simulations that use interim futility monitoring, and propose an organized BPP-based phase II-to-phase III design framework.

1 Introduction

The use of Bayesian predictive probability of success and predictive power as an alternative to the traditional, *conditional* power was first proposed by Spiegelhalter et al. (1986) for clinical trials that include interim analyses. The need to move away from conditional power calculations that make strong assumptions about the treatment effect towards *assurance* that averages over a distribution of potential treatment effects was then advocated by O'Hagan et al. (2005). More recently, the technical details of using predictive probability of success (PoS) in futility assessment at interim analysis involving time-to-event data and based on the logrank statistic, were discussed by Rufibach et al. (2016b), with further details regarding the choice of prior distributions provided by Rufibach et al. (2016a). Futility testing in Cox regression model using predictive probability of success (PPOS) is discussed in great detail by Tang (2015), who also proceeds to find optimal designs.

In this manuscript we derive closed-form expressions for the Bayesian predictive power (BPP) in non-inferiority testing for two-arm clinical trials that are either informed by external data, such as an independent, preceding trial, or include patient from interim analyses that go towards the final analysis. We also demonstrate how these expressions can be used to speed up computer-intensive simulations to help design Bayesian adaptive clinical trials.

The manuscript is organized as follows. In Section 2, we cover the basics of BPP. We derive expressions for the cross-trial and within-trial BPP concerning both numeric (Normal), dichotomous and time-to-event outcomes, the latter two being based on large sample normal approximations. We then provide the reader with an intermediate summary of results and a numeric example. The focus of Section 3 is the design of phase III trials with BPP-based rules for qualification of investigated therapies to the confirmatory trial. These rules are formulated in terms of quantiles of the BPP distribution, and we proceed to propose optimal rules that are derived from the Selectivity criterion. Section 4 covers the application of BPP to the design of Bayesian adaptive trials that require nontrivial computations. The paper

concludes with a discussion.

Investigators willing to employ some of the methods presented in this manuscript in their research are welcome to use the free web application created to accompany this paper and facilitate future analyses. For details see the Software section.

2 Bayesian Predictive Power in clinical trial design

Given a set of observations $y_1, \ldots, y_n \mid \boldsymbol{\theta} \overset{\text{i.i.d}}{\sim} p(y \mid \boldsymbol{\theta})$ and a prior distribution $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$, the *posterior predictive distribution* of a future observation y^* is

$$p(y^* \mid \boldsymbol{y}) = \int_{\Theta} p(y^* \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{y})\, \mathrm{d}\boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\left[p(y^* \mid \boldsymbol{\theta})\right],$$

where $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$ is the posterior distribution of $\boldsymbol{\theta}$. Let $T(\boldsymbol{y}^*)$ be a test statistic in a hypotheses test, evaluated at a vector of future observations $\boldsymbol{y}^*$, and let $\mathcal{C}$ denote the rejection region. The Bayesian Predictive Power is the posterior probability of rejecting the null hypothesis, namely

$$\mathbb{P}\left(T(\boldsymbol{y}^*) \in \mathcal{C} \;\middle|\; \boldsymbol{y}\right) = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\left[T(\boldsymbol{y}^*) \in \mathcal{C} \;\middle|\; \boldsymbol{y}; \boldsymbol{\theta}\right]. \tag{1}$$

Note that (1) is a simply Bayesian update of the Assurance proposed by O'Hagan et al. (2005).

2.1 Normal outcomes

Let $y_1, \ldots, y_n \sim \mathcal{N}(\mu, \sigma^2)$ be a random sample from the distribution of some numeric clinical outcome. Assume that σ is known, and consider a future size m sample average $\overline{y}^*_m$. Since $\overline{y}_n \mid \mu \sim \mathcal{N}(\mu, \sigma^2/n)$ and $\overline{y}^*_m \mid \mu \sim \mathcal{N}(\mu, \sigma^2/m)$, it is a well documented result (see e.g. Gelman et al., 2013) that if we assign a flat, non-informative prior $\pi(\mu) \propto 1$, then the posterior

predictive distribution of $\overline{y}^*_m$ is

$$\overline{y}^*_m \mid \boldsymbol{y}_n \sim \mathcal{N}\left(\overline{y}_n\,,\, \sigma^2\left[\frac{1}{n}+\frac{1}{m}\right]\right). \tag{2}$$

The overall, size $n+m$ sample mean can be decomposed

$$\overline{y}_{n+m} = \frac{n\overline{y}_n}{n+m} + \frac{m\overline{y}^*_m}{n+m}.$$

leading to

$$\overline{y}_{n+m} \mid \boldsymbol{y}_n \sim \mathcal{N}\left(\overline{y}_n\,,\, \frac{m\sigma^2}{n(n+m)}\right), \tag{3}$$

seeing as

$$\mathbb{E}\left\{\overline{y}_{n+m} \mid \boldsymbol{y}_n\right\} = \frac{n\overline{y}_n}{n+m} + \frac{m}{n+m}\mathbb{E}\left\{\overline{y}^*_m \mid \boldsymbol{y}_n\right\}$$

and

$$\mathrm{Var}\left\{\overline{y}_{n+m} \mid \boldsymbol{y}_n\right\} = \frac{m^2}{(n+m)^2}\mathrm{Var}\left\{\overline{y}^*_m \mid \boldsymbol{y}_n\right\}.$$

2.1.1 Cross-Trial Predictive Power

In this section and for the rest of this paper, we assume that the eventual efficacy testing will be conducted in a "frequentist" manner, that is: the statistical test is chosen under type I error rate constraints. Alternative formulations of the rejection region of the test, in the form of exceedance sets of Bayesian posterior probabilities such as the one used by Saville et al. (2014), are a viable alternative, but the testing framework is not one of Bayesian model comparison via Bayes factors as in Gelman et al. (2013). Bayesian criteria are thus used as mere decision rules for the purpose of interim monitoring within the standard hypothesis

testing framework.

Suppose now that we wish to perform a *non-inferiority* test of the form

$$\begin{cases} \mathcal{H}_0 & : \mu^t - \mu^c < -\gamma \quad \text{vs.} \\ \mathcal{H}_1 & : \mu^t - \mu^c \geqslant -\gamma \end{cases}$$

(here, for convenience, the underlying assumption is that large values are desirable. $\gamma > 0$ is the MCID: *Minimum Clinically Important Difference*) at significance level α, based on control observations $y_1^{c*}, \ldots, y_{m_c}^{c*} \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_c, \sigma_c^2)$ and treated patient data $y_1^{t*}, \ldots, y_{m_t}^{t*} \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_t, \sigma_t^2)$, where rejection of the null hypothesis is obtained whenever

$$Z := \frac{\overline{y}_{m_t}^{t*} - \overline{y}_{m_c}^{c*} + \gamma}{\sqrt{\frac{\sigma_t^2}{m_t} + \frac{\sigma_c^2}{m_c}}} > z_{1-\alpha}.$$

Here we wish to inform our distributions through Bayesian updating that is informed by past control and treatment data $y_1^c, \ldots, y_{n_c}^c \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_c, \sigma_c^2)$ and $y_1^t, \ldots, y_{n_t}^t \overset{\text{i.i.d}}{\sim} \mathcal{N}(\mu_t, \sigma_t^2)$, respectively. However, the eventual statistical testing will be conducted based on the most recent trial data only.

The Bayesian Predictive Power is then, from (2),

$$\begin{aligned} &\text{BPP}(\boldsymbol{y}^t, \boldsymbol{y}^c, \sigma_t, \sigma_c, n_t, n_c, m_t, m_c, \gamma, \alpha) \\ &\quad = \mathbb{P}\left(Z > z_{1-\alpha} \;\middle|\; \boldsymbol{y}\right) \\ &\quad = \mathbb{P}\left(\frac{\overline{y}_{m_t}^{t*} - \overline{y}_{m_c}^{c*} - (\overline{y}_{n_t}^t - \overline{y}_{n_c}^c)}{\sigma_{\hat{\Delta}^*|\boldsymbol{y}}} > \frac{z_{1-\alpha}\sigma_{\hat{\Delta}^*} - (\overline{y}_{n_t}^t - \overline{y}_{n_c}^c) - \gamma}{\sigma_{\hat{\Delta}^*|\boldsymbol{y}}} \;\middle|\; \boldsymbol{y}\right) \\ &\quad = \Phi\left(\frac{\overline{y}_{n_t}^t - \overline{y}_{n_c}^c + \gamma - z_{1-\alpha}\sigma_{\hat{\Delta}^*}}{\sigma_{\hat{\Delta}^*|\boldsymbol{y}}}\right), \end{aligned} \tag{4}$$

where $\Phi(\cdot)$ is the standard normal CDF,

$$\sigma^2_{\hat{\Delta}*} = \frac{\sigma_t^2}{m_t} + \frac{\sigma_c^2}{m_c} \tag{5}$$

and

$$\sigma^2_{\hat{\Delta}*|\boldsymbol{y}} = \sigma_t^2 \left[\frac{1}{n_t} + \frac{1}{m_t}\right] + \sigma_c^2 \left[\frac{1}{n_c} + \frac{1}{m_c}\right]. \tag{6}$$

This is what Rufibach et al. (2016b) refer to as "Updating PoS based on external information".

2.1.2 Within-trial Predictive Power

If the trial is conducted in a seamless manner (as in the case considered Spiegelhalter et al., 1986, where the goal is responding to the results of an interim analysis), incorporating data from multiple phases in the final analysis, the test statistic will then be based on the entire data set, i.e.

$$Z := \frac{\overline{y}^t_{n_t+m_t} - \overline{y}^c_{n_c+m_c} + \gamma}{\sqrt{\dfrac{\sigma_t^2}{n_t+m_t} + \dfrac{\sigma_c^2}{n_c+m_c}}}.$$

Following the exact same steps, this time taking advantage of (3), we obtain an identical expression to (4), this time having

$$\sigma^2_{\hat{\Delta}*} = \frac{\sigma_t^2}{n_t+m_t} + \frac{\sigma_c^2}{n_c+m_c} \tag{7}$$

and

$$\sigma^2_{\hat{\Delta}*|\boldsymbol{y}} = \frac{m_t\sigma_t^2}{n_t(n_t+m_t)} + \frac{m_c\sigma_c^2}{n_c(n_c+m_c)^2}. \tag{8}$$

2.2 Dichotomous outcomes

Suppose now that $y_n \sim \text{Binom}(n, p)$ is the total number of events observed in a group of n patients checked for a dichotomous outcome. Berry et al. (2010) provide an example of a single arm adaptive trial with dichotomous outcomes that includes Bayesian decision rules based on predictive probabilities.

Assigning $p \sim \text{U}(0, 1)$, we have

$$y_m^* \mid y_n \sim \text{BetaBinom}\left(m, y_n + 1, n - y_n + 1\right), \tag{9}$$

where y_m^* is the number of events to be observed in a future size m sample from the same population. When both n and m are "large enough", we can approximate this distribution by the Normal distribution with the original mean and variance, namely

$$y_m^* \mid y_n \overset{\bullet}{\sim} \mathcal{N}\left(\frac{m(y_n+1)}{n+2}, \frac{(y_n+1)(n-y_n+1)(m+n+2)}{(n+2)^2(n+3)}\right)$$

and subsequently, after neglecting some constants,

$$\widehat{p}_m^* \mid y_n \overset{\bullet}{\sim} \mathcal{N}\left(\widehat{p}_n, \widehat{p}_n(1-\widehat{p}_n)\left[\frac{1}{n}+\frac{1}{m}\right]\right), \tag{10}$$

where $\widehat{p}$ denotes the sample proportion. As in the Normal case, we can proceed to derive the posterior predictive distribution of the overall sample proportion, noting that

$$\widehat{p}_{n+m} = \frac{n\widehat{p}_n}{n+m} + \frac{m\widehat{p}_m^*}{n+m}$$

and using the same arguments as before we obtain

$$\widehat{p}_m^* \mid y_n \overset{\bullet}{\sim} \mathcal{N}\left(\widehat{p}_n, \frac{m\widehat{p}_n(1-\widehat{p}_n)}{n(n+m)}\right). \tag{11}$$

We may now put the above to use in a non-inferiority hypotheses test (assuming events are undesirable)

$$\begin{cases} \mathcal{H}_0 & : p^c - p^t < -\gamma \quad \text{vs.} \\ \mathcal{H}_1 & : p^c - p^t \geqslant -\gamma \end{cases}$$

where rejection of the null hypothesis occurs if

$$Z := \frac{\widehat{p}^{c*}_{m_c} - \widehat{p}^{t*}_{m_t} + \gamma}{\sqrt{\dfrac{\widehat{p}^{t*}_{m_t}(1-\widehat{p}^{t*}_{m_t})}{m_t} + \dfrac{\widehat{p}^{c*}_{m_c}(1-\widehat{p}^{c*}_{m_c})}{m_c}}} > z_{1-\alpha}$$

In light of (10) and (11), the Bayesian Predictive Power in both cases can be approximated by

$$\text{BPP}(y^t, y^c, n_t, n_c, m_t, m_c, \gamma, \alpha) = \Phi\left(\frac{\widehat{p}^c_{n_c} - \widehat{p}^t_{n_t} + \gamma - z_{1-\alpha}\sigma_{\hat{\Delta}*}}{\sigma_{\hat{\Delta}*|\boldsymbol{y}}}\right), \tag{12}$$

where in the case of independent tests

$$\sigma^2_{\hat{\Delta}*} = \frac{\widehat{p}^t_{n_t}(1-\widehat{p}^t_{n_t})}{m_t} + \frac{\widehat{p}^c_{n_c}(1-\widehat{p}^c_{n_c})}{m_c} \tag{13}$$

and

$$\sigma^2_{\hat{\Delta}*|\boldsymbol{y}} = \widehat{p}^t_{n_t}(1-\widehat{p}^t_{n_t})\left[\frac{1}{n_t} + \frac{1}{m_t}\right] + \widehat{p}^c_{n_c}(1-\widehat{p}^c_{n_c})\left[\frac{1}{n_c} + \frac{1}{m_c}\right] \tag{14}$$

whereas for a seamless test

$$\sigma^2_{\hat{\Delta}*} = \frac{\widehat{p}^t_{n_t}(1-\widehat{p}^t_{n_t})}{n_t + m_t} + \frac{\widehat{p}^c_{n_c}(1-\widehat{p}^c_{n_c})}{n_c + m_c} \tag{15}$$

and

$$\sigma^2_{\hat{\Delta}^*|\boldsymbol{y}} = \frac{m_t \widehat{p}^t_{n_t}(1-\widehat{p}^t_{n_t})}{n_t(n_t+m_t)} + \frac{m_c \widehat{p}^c_{n_c}(1-\widehat{p}^c_{n_c})}{n_c(n_c+m_c)}. \tag{16}$$

Note that (15) implicitly uses the consistency of $\widehat{p}^t_{n_t}$ and $\widehat{p}^c_{n_c}$ and Slutsky's Theorem. Exact calculation of the Bayesian Predictive Power by Bayesian clinical trial simulation (BCTS) is proposed by O'Hagan et al. (2005).

2.3 Time-to-event outcomes

Here the goal is to test hypotheses of the form

$$\begin{cases} \mathcal{H}_0 & : \theta > \gamma \quad \text{vs.} \\ \mathcal{H}_1 & : \theta \leqslant \gamma \end{cases}$$

where θ denotes the log-hazard ratio between treatment and control and the test in use is the Wald test based on the maximum partial likelihood estimator $\widehat{\theta}$. We will use the *int* (for "interim") and *max* subscripts to denote estimation and analysis performed after the initial $n_t + n_c$ patient outcomes have been recorded and at the final analysis, in which all $n_t + m_t + n_c + m_c$ outcomes have been observed (including censored outcomes), respectively. The results presented in this section largely mirror the work of Tang (2015), and are based on the asymptotic Brownian motion properties of the logrank test statistic as detailed by Tsiatis (2014).

Denoting the respective average arm size in the early and follow-up trials by

$$\bar{n} = \frac{n_t + n_c}{2}, \quad \bar{m} = \frac{m_t + m_c}{2}, \tag{17}$$

as well as the average event rate

$$\bar{p} = \frac{p_t + p_c}{2} := \frac{d_{\text{int}}}{2\bar{n}} = \frac{d_{\max}}{2\left(\bar{n} + \bar{m}\right)}, \tag{18}$$

The posterior predictive distribution of the follow-up log-hazard ratio estimator when previously collected data are ignored is given by

$$\widehat{\theta}_{-\text{int}} \big| \widehat{\theta}_{\text{int}} \sim \mathcal{N}\left(\widehat{\theta}_{\text{int}}, \frac{1}{4\bar{p}r(1-r)}\left[\frac{1}{\bar{n}} + \frac{1}{\bar{m}}\right]\right), \tag{19}$$

while incorporating all collected data into the final analysis results in

$$\widehat{\theta}_{\max} \big| \widehat{\theta}_{\text{int}} \sim \mathcal{N}\left(\widehat{\theta}_{\text{int}}, \frac{1}{4\bar{p}r(1-r)}\frac{\bar{m}}{\bar{n}\left(\bar{n} + \bar{m}\right)}\right), \tag{20}$$

where r is the proportion of patients randomized to the treatment arm. Together, (19) and (20) lay out the blueprint for the the calculation of between- and within-trial BPP, respectively.

The details behind the derivation of these expressions are provided in Appendix A.

2.4 Intermediate summary and an example

The formulas derived in the above discussion can be conveniently summarized as follows: The Bayesian Predictive Power is given by

$$\text{BPP} = \Phi\left(\frac{\widehat{\Delta} + \gamma - z_{1-\alpha}\sigma_{\widehat{\Delta}*}}{\sigma_{\widehat{\Delta}*|\boldsymbol{y}}}\right), \tag{21}$$

where $\widehat{\Delta}$, $\sigma_{\widehat{\Delta}*}$ and $\sigma_{\widehat{\Delta}*|\boldsymbol{y}}$ are given in Table 1.

Example 2.5. We will now illustrate the application of the BPP to a two-arm trial involving a dichotomous outcome, assuming $n_t = n_c = 100$, $m_t = m_c = 500$, $y^t = 5$, $y^c = 10$ and $\gamma = 0$

$\widehat{\Delta}$	BPP type	$\sigma^2_{\widehat{\Delta}*}$	$\sigma^2_{\widehat{\Delta}*\|\boldsymbol{y}}$
Normal Outcomes			
$\overline{y}^t_{n_t} - \overline{y}^c_{n_c}$	Cross-trial	$\dfrac{\sigma_t^2}{m_t} + \dfrac{\sigma_c^2}{m_c}$	$\sigma_t^2\left[\dfrac{1}{m_t} + \dfrac{1}{n_t}\right] + \sigma_c^2\left[\dfrac{1}{m_c} + \dfrac{1}{n_c}\right]$
	Within-trial	$\dfrac{\sigma_t^2}{n_t + m_t} + \dfrac{\sigma_c^2}{n_c + m_c}$	$\dfrac{m_t\sigma_t^2}{n_t(n_t + m_t)^2} + \dfrac{m_c\sigma_c^2}{n_c(n_c + m_c)^2}$
Dichotomous Outcomes			
$\widehat{p}^c_{n_c} - \widehat{p}^t_{n_t}$	Cross-trial	$\dfrac{\widehat{p}^t_{n_t}(1-\widehat{p}^t_{n_t})}{m_t} + \dfrac{\widehat{p}^c_{n_c}(1-\widehat{p}^c_{n_c})}{m_c}$	$\widehat{p}^t_{n_t}(1-\widehat{p}^t_{n_t})\left[\dfrac{1}{n_t} + \dfrac{1}{m_t}\right] + \widehat{p}^c_{n_c}(1-\widehat{p}^c_{n_c})\left[\dfrac{1}{n_c} + \dfrac{1}{m_c}\right]$
	Within-trial	$\dfrac{\widehat{p}^t_{n_t}(1-\widehat{p}^t_{n_t})}{n_t + m_t} + \dfrac{\widehat{p}^c_{n_c}(1-\widehat{p}^c_{n_c})}{n_c + m_c}$	$\dfrac{m_t\widehat{p}^t_{n_t}(1-\widehat{p}^t_{n_t})}{n_t(n_t + m_t)} + \dfrac{m_c\widehat{p}^c_{n_c}(1-\widehat{p}^c_{n_c})}{n_c(n_c + m_c)}$
Time-to-event Outcomes			
$\widehat{\theta}_{\text{int}}$	Cross-trial	$\dfrac{1}{4\bar{p}r(1-r)\bar{m}}$	$\dfrac{1}{4\bar{p}r(1-r)}\left[\dfrac{1}{\bar{n}} + \dfrac{1}{\bar{m}}\right]$
	Within-trial	$\dfrac{1}{4\bar{p}r(1-r)\left(\bar{n} + \bar{m}\right)}$	$\dfrac{1}{4\bar{p}r(1-r)}\dfrac{\bar{m}}{\bar{n}\left(\bar{n} + \bar{m}\right)}$

Table 1: A summary of BPP calculations for the various scenarios.

(i.e. a *superiority test*) and that the eventual test will be performed at the 5% significance level.

Figure 1 shows the quality of the large sample approximation (10), by plotting a histogram of the difference of two proportions calculated from two random samples drawn from the Beta-Binomial distribution (9) alongside the $\mathcal{N}\left(\hat{p}_n^c - \hat{p}_n^t, \sigma^2_{\hat{\Delta}^*|\boldsymbol{y}}\right)$, with $\sigma_{\hat{\Delta}^*|\boldsymbol{y}}$ calculated as in (14). The quality of the approximation comes as no surprise, considering the known Binomial approximation for the Beta-Binomial distribution (Teerapabolarn, 2008, 2014).

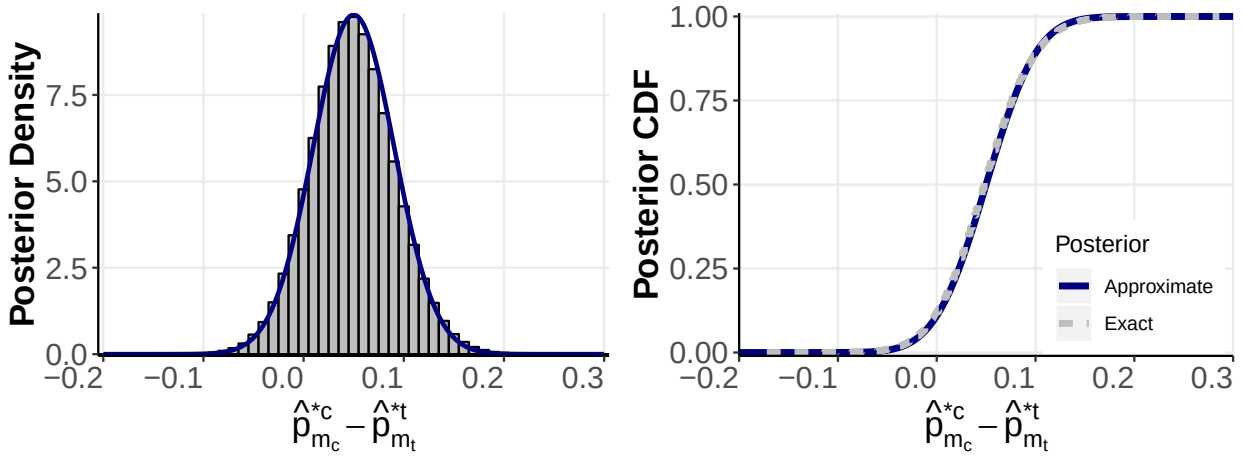

Figure 1: Assessing the quality of the normal approximation (10) to the posterior predictive distribution of proportions based on (9) by simulation. Left panel: the solid line represents the approximate density, while the histogram is based on a size 10^6 Monte Carlo sample. Right panel: comparison of the CDFs based on the same approximation and random sample.

When plugging the numbers into (21), using the binomial arguments from Table (1), we learn that the cross-trial BPP for these specifications is 71.2%, while incorporating the existing data in a seamless trial yield a within-trial BPP of 77.1%. Furthermore, a simple line search reveals that enrolling 1212 and 775 future patients per arm in an independent and seamless trial, respectively, will guarantee meeting a target BPP of 80%, as shown in Figure 2. This is akin to Spiegelhalter and Freedman (1986) with fewer potential conclusions. Note that meeting a prescribed BPP target by simply increasing the sample size is not always achievable: if the observed control event rate is smaller than that observed in the treatment

arm, no number of future patients will suffice to ensure that, seeing as the treatment effect posterior distribution is now centered at a negative value.

For readers alarmed by the large difference between the between- and within-trial BPP sample size requirements, we will now elaborate on this delicate topic. At the point of calculating of the BPP, our belief with regard to the treatment efficacy is being updated in the form of the treatment effect posterior distribution. If our next step, however, is to conduct an independent follow up trial using only future data for the final analysis, there is still a considerable posterior probability that the newly collected data will not yield results as favourable as those previously obtained – especially if the prior distribution used was very diffuse. If we are to continue seamlessly, though, we do so with both the reinforced belief and the early data in support of treatment efficacy to boot. This reduces the uncertainty regarding the way that the eventual analysis will turn out, consequently reducing sample size demands.

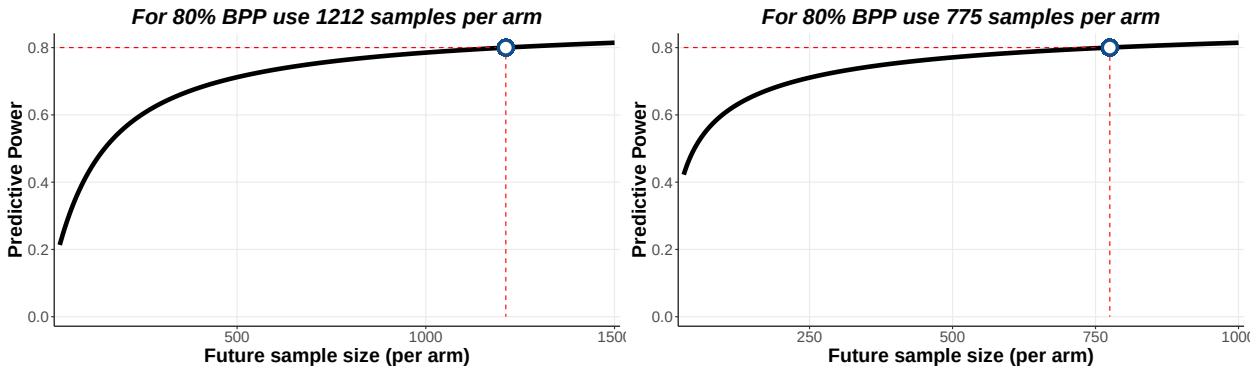

Figure 2: Finding the future sample size (per arm) required to ensure meeting the target 80% BPP in independent (left panel) and seamless (right panel) trials, based on the data used in Example 2.5.

3 Bayesian Predictive Power as a screening tool in phase II trials

We may now consider using Bayesian Predictive Power to screen out proposed treatments in a multi-arm phase II clinical trial. A similar approach was employed in Stallard et al. (2005) and Tang (2015) for time-to-event outcomes, with the latter's focus being futility testing.

3.1 Predictive power distribution and selection probabilities

The application is straight forward:

1. Choose a cutoff p_{cut},
2. At the end of the phase II trial calculate Bayesian Predictive Power for all treatment arms, and
3. Proceed to phase III with those arms whose predictive power exceeded p_{cut}.

Note that other statistical criteria may be used in stages (2) and (3) of the above schema. However, the Bayesian Predictive Power frames nicely our updated belief with regard to future success in a confirmatory trial in light of the data observed, as well as the amount of data yet to be observed. A decision rule along the lines of "proceed only with treatments that have at least 60% of being approved" is easy to communicate. This kind of statement cannot otherwise be made in the case of independent analyses (i.e. when cross-trial BPP is used) if the Bayesian approach is not employed. More importantly, unlike other potential criteria like the p-value, the BPP brings into consideration both past and future sample sizes, hence the decision whether to continue or discontinue an arm may be different between trials with different resources, despite showing the same level of evidence. This resonates with the advocating of Saville et al. (2014) for the use of Bayesian predictive probabilities in monitoring futility.

Combining the above criterion with (21) and denoting the BPP at the end of the phase II (as a function of the model parameters $\boldsymbol{\theta}$) by

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{y}) \coloneqq \mathbb{P}_{\boldsymbol{\theta}}\left(Z \geqslant z_{1-\alpha} \,\middle|\, \boldsymbol{y}\right),$$

we gather that for a phase III test at level α of the form

$$\begin{cases} \mathcal{H}_0 : \Delta(\boldsymbol{\theta}) < -\gamma \quad \text{vs.} \\ \\ \mathcal{H}_1 : \Delta(\boldsymbol{\theta}) \geqslant -\gamma \end{cases}$$

(here $\Delta\left(p^c, p^t\right) = p^c - p^t$ for dichotomous outcomes for example), the BPP CDF is then

$$\begin{aligned} F_{\pi_{\boldsymbol{\theta}}(\boldsymbol{y})}(t) &= \mathbb{P}\left(\Phi\left(\frac{\widehat{\Delta} + \gamma - z_{1-\alpha}\sigma_{\widehat{\Delta}*}}{\sigma_{\widehat{\Delta}*|\boldsymbol{y}}}\right) \leqslant t\right) \\ &= \mathbb{P}\left(\frac{\widehat{\Delta} - \Delta}{\sigma_{\widehat{\Delta}}} \leqslant \frac{z_{1-\alpha}\sigma_{\widehat{\Delta}*} - \gamma - \Delta + z_t\sigma_{\widehat{\Delta}*|\boldsymbol{y}}}{\sigma_{\widehat{\Delta}}}\right) \\ &= \Phi\left(\frac{z_{p_{\text{cut}}}\sigma_{\widehat{\Delta}*|\boldsymbol{y}} + z_{1-\alpha}\sigma_{\widehat{\Delta}*} - \gamma - \Delta}{\sigma_{\widehat{\Delta}}}\right), \end{aligned} \tag{22}$$

where

$$\sigma^2_{\widehat{\Delta}} = \begin{cases} \dfrac{\sigma_t^2}{n_t} + \dfrac{\sigma_c^2}{n_c} & \text{Normal outcomes} \\ \\ \dfrac{p^t_{n_t}(1 - p^t_{n_t})}{n_t} + \dfrac{p^c_{n_c}(1 - p^c_{n_c})}{n_c} & \text{Dichotomous outcomes} \\ \\ \dfrac{1}{4\bar{p}r(1-r)\bar{n}} & \text{Time-to-event outcomes} \end{cases}$$

is the standard error of $\widehat{\Delta}$ (based on phase II data only). The expected predictive power at

interim analysis can then be calculated by

$$\overline{\pi}_{\boldsymbol{\theta}} := \mathbb{E}_{\boldsymbol{y}}\left[\pi_{\boldsymbol{\theta}}(\boldsymbol{y})\right] = \int_0^1 \left[1 - F_{\pi_{\boldsymbol{\theta}}(\boldsymbol{y})}(t)\right] \mathrm{d}t. \tag{23}$$

Moreover, differentiation of (22) yields the (limiting) BPP probability density function

$$f_{\pi_{\boldsymbol{\theta}}(\boldsymbol{y})}(t) = \frac{\sigma_{\hat{\Delta}*|\boldsymbol{y}}}{\sigma_{\hat{\Delta}*}\phi(z_t)}\phi\left(\frac{z_t\sigma_{\hat{\Delta}*|\boldsymbol{y}} + z_{1-\alpha}\sigma_{\hat{\Delta}*} - \gamma - \Delta}{\sigma_{\hat{\Delta}}}\right), \tag{24}$$

where $\phi(\cdot)$ is the standard normal PDF. For a set cutoff p_{cut}, the probability of a treatment arm progressing to the confirmatory trial is then

$$\mathbb{P}\left(\pi_{\boldsymbol{\theta}}(\boldsymbol{y}) > p_{\text{cut}}\right) = 1 - F_{\pi_{\boldsymbol{\theta}}(\boldsymbol{y})}\left(p_{\text{cut}}\right). \tag{25}$$

3.2 Optimal cutoff by Selectivity

The question of what of constitutes a "good" cutoff p_{cut} has no clear-cut answer and varies between studies and investigators. Typically, phase II trials tend to be liberal, with more emphasis put on limiting the risk for false negatives, that is: passing on efficacious treatments. In light of this, some investigators would choose a threshold that would guarantee a high probability of qualification to the confirmatory trial for any treatment whose effect size (relative to control) meets a certain minimum.

However, oftentimes the phase III trial is limited to a fixed number of arms, in which case the role of the phase II trial is to identify the most efficacious subset of treatments with maximum precision. In attempt to achieve this goal, the authors of this paper propose the *Selectivity* criterion. Denote the parameters of the M different phase II candidate therapies, listed in a decreasing order of effect size, $\boldsymbol{\theta}_i$, $i = 1, 2, \ldots, M$, and suppose that we are tasked with selecting a maximum number of K therapies for a confirmatory trial, based on their

BPP evaluations at the end of the phase II trial. The Selectivity function is

$$\mathcal{S}(p_{\text{cut}}, \boldsymbol{\theta}, K) = \prod_{i=1}^{K} \left[1 - F_{\pi_{\boldsymbol{\theta}_i}(\boldsymbol{y})}\left(p_{\text{cut}}\right)\right] \prod_{i=K+1}^{M} F_{\pi_{\boldsymbol{\theta}_i}(\boldsymbol{y})}\left(p_{\text{cut}}\right), \tag{26}$$

where $F_{\pi_{\boldsymbol{\theta}_i}(\boldsymbol{y})}\left(p_{\text{cut}}\right)$ is as in (22). A sensible choice for p_{cut} would then be the maximizer of (26). Note that this is the cutoff that maximizes the probability of selecting the correct subset of treatments.

3.3 BPP-based trial design: a worked example

Example 3.4. As an illustrative example, suppose that the outcome of interest is a dichotomous one, and we conduct a phase II trial whose goal is to select a subset (of an undefined size as of yet) of four treatment arms to continue into a phase III confirmatory trial. Assuming a *control event rate* (CER) of 12.5%, the investigators plan to base their phase III qualification criterion on the Bayesian Predictive Power. Our goal is thus to choose a cutoff that will be useful at distinguishing between treatments whose conjectured *relative risk reductions* (RRRs) are 0%, 15%, 30% and 45%.

Assuming per-arm phase II and phase III sample sizes of 250 and 750, respectively, for all arms, and the customary 5% phase III significance level, we calculate the optimal Selectivity cutoff for both cross-trial and within-trial BPPs, assuming desired treatment subsets of size 1, 2 and 3. The Selectivity curves along with the emerging cutoffs are on display in Figure 3, and the corresponding selection probabilities for all arms can be found in Table 2. Graphical illustration of the impact of the cutoff chosen on the selection probabilities appears in Figure 4, where the corresponding BPP densities (24) are also presented. To the puzzled reader, we wish to clarify that the randomness in the BPP stems from that of the data yet to be observed. Once data for the phase II is gathered, the BPP turns into a number again.

We would like to stress again that these choices for cutoffs are but a mere recommen-

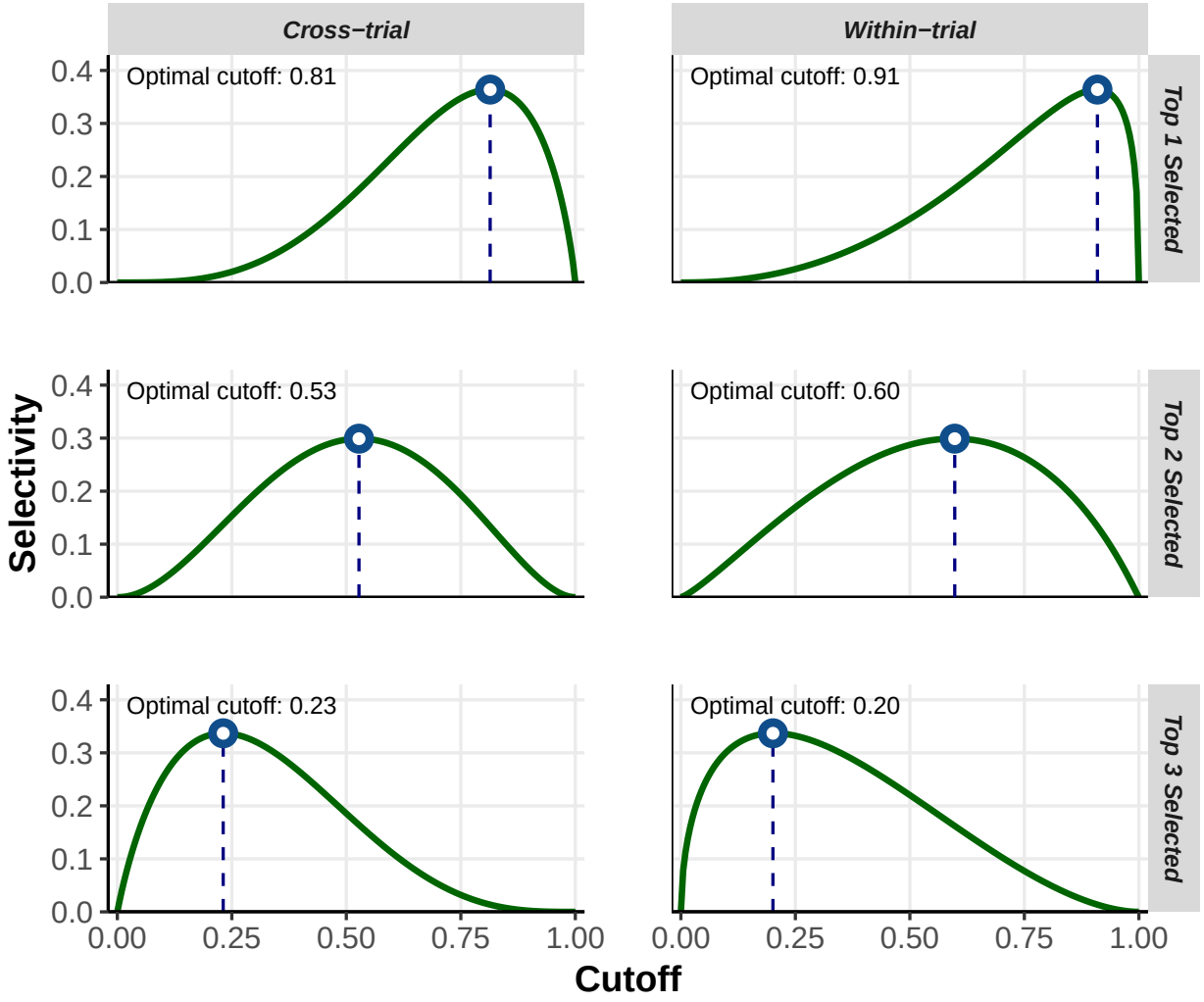

Figure 3: The Selectivity criterion evaluated at $0 \leqslant p_{\text{cut}} \leqslant 1$ for all target subset sizes, assuming both cross-trial and within-trial BPP.

		Probability of Selection		
RRR	Expected Power	Top 1 Maximum Selectivity	Top 2 Maximum Selectivity	Top 3 Maximum Selectivity
Cross-trial BPP, $p_{\text{cut}} = \{0.81, 0.53, 0.23\}$				
0%	26.8%	2.5%	15.0%	46.3%
15%	42.4%	9.6%	35.2%	71.3%
30%	60.6%	27.4%	62.8%	89.8%
45%	78.0%	56.8%	86.4%	97.9%
Within-trial BPP, $p_{\text{cut}} = \{0.91, 0.60, 0.20\}$				
0%	26.8%	2.4%	14.9%	46.2%
15%	45%	9.2%	35.0%	71.2%
30%	65.8%	26.7%	62.6%	89.7%
45%	83.8%	56.1%	86.3%	97.9%

Table 2: Dichotomous outcome treatment effects and their corresponding selection probabilities for different phase II to phase III transitions, based on the respective BPP cutoffs chosen to maximize the selectivity criterion when either one or two arms are planned to graduate to a confirmatory trial. Here we assume CER = 12.5% as well as $n_t = n_c = 250$ and $m_t = m_c = 750$ for all treatment arms. The phase III test significance level is planned to be 5%.

dation. The reader may correctly point to relatively low selection probability of arm 3, the second best treatment, when the stated goal is selecting the top 2 treatments, and is fully entitled to opt for a smaller cutoff, at the increased risk of the third best arm sneaking in. It should also be made clear at this point that the example provided here is an illustrative one and the reported numbers need not represent a realistic phase II trial design. To begin with, because of the one-sided test used, the investigator may opt for a 2.5% type I error rate rather than the customary 5%. In addition, to minimize the risk of advancing less efficacious treatments, arm sizes may be increased for both phase II and III cohorts. Also note that in our example, the probability of at least one non-efficacious arm graduating to phase III if one is aiming to advance three arms is 84% in the worst case "null scenario" that no treatment is efficacious. Worse yet: there is a $4 \times (46\%)^3 \times 54\% \approx 21\%$ probability of advancing three non-efficacious treatments to the next phase. This is clearly an exceedingly high probability of an utterly undesirable outcome and the investigators would be strongly

encouraged to abandon their initial strategy. The reader is encouraged to explore different design configuration using our BIAS tool, discussed in the Software section.

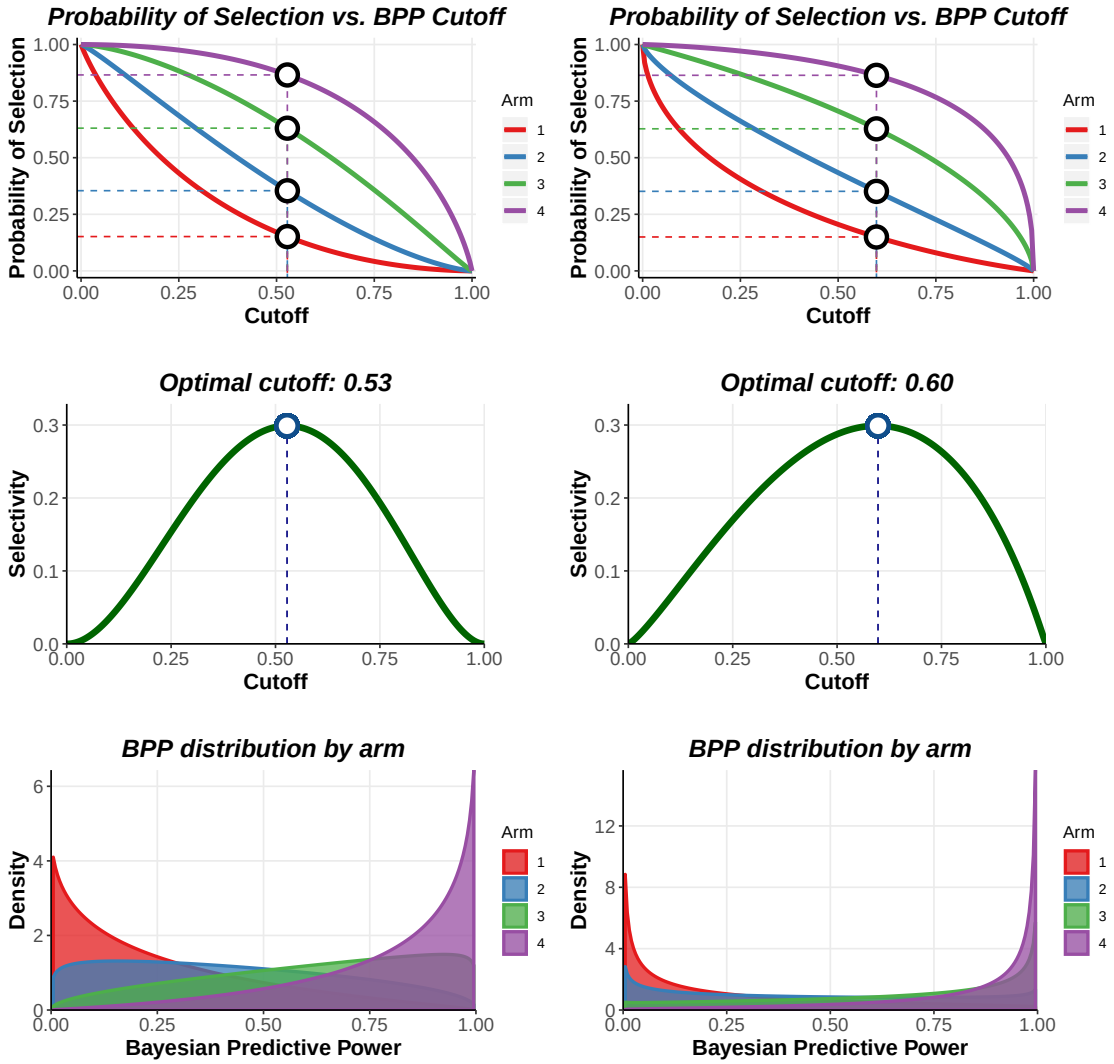

Figure 4: Selection probability curves, Selectivity plots and BPP densities for Example 3.4. Left panels: independent trials and cross-trial predictive power. Right panels: "seamless" trial and within-trial predictive power.

4 Application to Bayesian adaptive clinical trial design

Large scale simulations play an instrumental role in the design of adaptive clinical trials with interim futility monitoring (Berry et al., 2010 or Saville et al., 2014). Decision boundaries

must be determined to meet the requirements regarding operating characteristics such as type I error rate and power, different designs need be compared by means of their time saving, and sensitivity to unknowns such as the underlying treatment effect and recruitment and dropout rate should be considered. All this means that a large number of scenarios has to be simulated numerously many times to estimate said operating characteristics accurately.

Consider, for example, futility monitoring in a two-arm placebo-controlled trial with a primary dichotomous outcome. At interim analysis, the posterior predictive distribution of either arm is the Beta-Binomial distribution (9). In the absence of a closed-form expression, evaluating the probability of success with respect to any efficacy testing criterion will therefore entail –

1. Drawing a very large number of vectors of future observations from (9) for both arms,

2. add each pair of corresponding random vectors to the data accumulated prior to interim analysis and calculate the resulting test statistic or, alternatively, the posterior probability of treatment superiority in the case of Bayesian efficacy testing. Record the decision made.

3. Average the decision made over all randomly drawn posterior-predictive draws to estimate the probability of success, that is: the probability of rejecting the null hypothesis of treatment futility.

The above computationally expensive process amounts to the calculation of a single predictive probability of success for futility monitoring. Reasonable turnaround when simulating even a modest number of scenarios might require enormous resources, while MCMC sampling in cases when the posterior predictive distribution is not given in a closed form might deem the process prohibitive. The problems become even more compound by introducing designs with multiple interim analyses. The following example demonstrates how using the BPP expressions presented in this paper facilitates comprehensive trial design with considerable

time and resource saving by employing frequent futility monitoring.

Example 4.1. Suppose that we wish to design an adaptive clinical trial with a dichotomous primary outcome, aimed at testing $\mathcal{H}_0 : p_c \leqslant p_t$ vs. $\mathcal{H}_1 : p_c > p_t$ at the 2.5% level, with the investigators assessing the control event rate at $p_c = 0.3$. The follow up time is 10 days. We simulate the following trial design:

1. Patients enrol in a homogeneous Poisson process with rate λ patients per day, and are randomized (with equal probability) to treatment or control.

2. Ten days after enrolment, $100(1-\delta)\%$ of the patient outcomes become available, where δ is the loss-to-follow up rate.

3. Once at least 130 patient outcomes have become available for analysis for both the treatment and the control arm, and until a maximum of 500 patients have been enrolled, do the following at the end of every day –

   I. Stop the trial for efficacy if

   $$\mathbb{P}\left(p_c > p_t \mid \boldsymbol{y}\right) > 0.9925, \tag{27}$$

   where (27) is evaluated from independent Monte Carlo samples from the Beta$(y_c + 1, n_c - y_c + 1)$ and Beta$(y_t + 1, n_t - y_t + 1)$ posterior distributions of p_c and p_t, respectively.

   II. Stop the trial for futility if

   $$\mathrm{BPP}(y_t, y_c, n_t, n_c, m_t, m_c, \alpha) < 0.025,$$

   using the within-trial Bayesian predictive power formula for dichotomous outcomes from Table 1 with $\alpha = 0.025$.

Note that futility testing in this design uses the Bayesian predictive power that was calculated assuming a standard, frequentist test is conducted for efficacy, despite using a Bayesian efficacy criterion in this particular trial design. Both superiority and futility threshold were found experimentally, and cannot be determined upfront using long-hand calculations.

We simulated the above design for all 18 combinations of recruitment rate $\lambda \in \{10, 15, 20\}$, loss-to-follow up rate $\delta \in \{5\%, 10\%\}$ and relative risk reduction RRR $\in \{40\%, 45\%, 50\%\}$. Each scenario was run 100,000 times, as per the FDA recommendationFood and Drug Administration (2019), using Monte Carlo samples of size 1,000 for the evaluation of (27). The overall simulation time, when run in parallel on Intel Xeon Scalable Processors with 96 virtual 3.6GHz CPUs, using R version 3.6.0R Core Team (2020), was 1:09:08 hours.

The simulation results, presented in Table 3, show that the futility rule in the null scenarios was as effective as the efficacy rule under the alternative hypothesis: at best – assuming slowest accrual and minimum dropouts – it should save the investigator 17.6 days on average (42.4 to 60) and 81.8 enrolled patients. Note that the numbers of patients enrolled reported in Table 3 include patients whose outcomes were not recorded, either due to loss-to-follow up or the 10 day delayed outcomes. For context, a standard placebo-controlled trial of the same power and type I error rate would require recruiting 426 patients and take on average 52.6 days to be completed.

5 Discussion

The Bayesian paradigm represents an appealing option when there is a continual accrual of data and a subsequent adjustment of expectation. The naive act of plugging in potentially small sample parameter estimates in standard conditional power and sample size calculations does not reflect any of the uncertainties carried over into future calculations. The methods presented in this paper meet this need, and can be fairly easily extended to a variety of tests

Recruitment rate (patients/day)	Dropout rate	Power	Mean duration ($\mathcal{H}_1$)	Mean enrolled ($\mathcal{H}_1$)	Type I err. rate	Mean duration ($\mathcal{H}_0$)	Mean enrolled ($\mathcal{H}_0$)
RRR = 40%							
10	5%	80.5%	44.2	429.4	2.5%	42.4	418.2
	10%	78.0%	45.9	443.3	2.4%	43.9	431.9
15	5%	79.6%	33.5	470.6	2.3%	32.2	464.8
	10%	77.0%	34.0	477.7	2.2%	32.7	472.1
20	5%	78.6%	27.6	495.6	2.1%	26.7	494.5
	10%	75.9%	28.3	498.7	2.1%	27.2	498.3
RRR = 45%							
10	5%	89.6%	42.6	417.8	2.5%	42.4	418.2
	10%	87.7%	44.4	433.8	2.4%	43.9	431.9
15	5%	89.1%	32.3	464.5	2.3%	32.2	464.8
	10%	87.2%	33.0	473.0	2.2%	32.7	472.1
20	5%	88.5%	26.8	494.5	2.1%	26.7	494.5
	10%	86.2%	27.5	498.4	2.1%	27.2	498.3
RRR = 50%							
10	5%	95.3%	41.2	407.4	2.5%	42.4	418.2
	10%	94.2%	43.1	425.1	2.4%	43.9	431.9
15	5%	95.1%	31.4	458.8	2.3%	32.2	464.8
	10%	93.8%	32.1	468.6	2.2%	32.7	472.1
20	5%	94.7%	26.0	493.4	2.1%	26.7	494.5
	10%	93.4%	26.8	498.1	2.1%	27.2	498.3

Table 3: The operating characteristics of the adaptive design from Example 4.1 for various input configurations.

and outcomes. We urge the reader to experiment with our online application – discussed in the Software section – and consider incorporating Bayesian predictive power in future clinical trial design.

Software

BIAS: Bayesian Interim Analysis Software, a web application that performs all the analyzes presented in this paper and some additional Bayesian adaptive trial features is freely available to all users at **https://mtek.shinyapps.io/Interim_Analysis_Tool/**.

Acknowledgement

The authors wish to acknowledge the Bill and Melinda Gates Foundation Knowledge Integration team for their feedback and constructive comments.

Appendix A Derivation of the expressions for time-to-event outcomes

As discussed by Tang (2015), the large sample approximation of the maximum partial likelihood estimator of the log-hazard ratio at interim and final analysis is given by

$$\widehat{\theta}_{\text{int}} \mid \theta \sim \mathcal{N}\left(\theta, \frac{1}{r(1-r)d_{\text{int}}}\right) \tag{28}$$

and

$$\widehat{\theta}_{\max} \mid \theta \sim \mathcal{N}\left(\theta, \frac{1}{r(1-r)d_{\max}}\right), \tag{29}$$

respectively, where r is the proportion of patients randomized to the treatment arm and d denotes the number of events observed at the different time points. Assigning $\pi(\theta) \propto 1$, in conjunction with (28) we have

$$\theta \mid \widehat{\theta}_{\text{int}} \sim \mathcal{N}\left(\widehat{\theta}_{\text{int}}, \frac{1}{d_{\text{int}}}\right).$$

If we now wish to derive the posterior predictive distribution of the log-hazard ratio estimator based on the remaining of the data (excluding data collected up to interim analysis), denoted $\widehat{\theta}_{-\text{int}}$, we have

$$\mathbb{E}\left[\widehat{\theta}_{-\text{int}} \mid \widehat{\theta}_{\text{int}}\right] = \mathbb{E}\left\{\mathbb{E}\left[\widehat{\theta}_{-\text{int}} \mid \theta, \widehat{\theta}_{\text{int}}\right]\right\} = \mathbb{E}\left[\theta \mid \widehat{\theta}_{\text{int}}\right] = \widehat{\theta}_{\text{int}}$$

and

$$\begin{aligned}\text{Var}\left[\widehat{\theta}_{-\text{int}} \mid \widehat{\theta}_{\text{int}}\right] &= \mathbb{E}\left\{\text{Var}\left[\widehat{\theta}_{-\text{int}} \mid \theta, \widehat{\theta}_{\text{int}}\right]\right\} + \text{Var}\left\{\mathbb{E}\left[\widehat{\theta}_{-\text{int}} \mid \theta, \widehat{\theta}_{\text{int}}\right]\right\} \\ &= \frac{1}{r(1-r)\left(d_{\max} - d_{\text{int}}\right)} + \frac{1}{r(1-r)d_{\text{int}}},\end{aligned}$$

therefore

$$\widehat{\theta}_{-\text{int}} \mid \widehat{\theta}_{\text{int}} \sim \mathcal{N}\left(\widehat{\theta}_{\text{int}}, \frac{1}{r(1-r)}\left[\frac{1}{d_{\text{int}}} + \frac{1}{d_{\max} - d_{\text{int}}}\right]\right). \tag{30}$$

We may also define the complementary estimator

$$\widehat{\theta}_{\text{comp}} := \frac{d_{\max}\widehat{\theta}_{\max} - d_{\text{int}}\widehat{\theta}_{\text{int}}}{d_{\max} - d_{\text{int}}} \sim \mathcal{N}\left(\theta, \frac{1}{r(1-r)(d_{\max} - d_{\text{int}})}\right)$$

which is asymptotically independent of $\widehat{\theta}_{\text{int}}$ (as a result of the independent increments property of the Brownian motion) and whose posterior predictive distribution is identical to that

of $\widehat{\theta}_{-\text{int}} \mid \widehat{\theta}_{\text{int}}$ (Tang, 2015), i.e.

$$\widehat{\theta}_{\text{comp}} \mid \widehat{\theta}_{\text{int}} \sim \mathcal{N}\left(\widehat{\theta}_{\text{int}}, \frac{1}{r(1-r)}\left[\frac{1}{d_{\text{int}}} + \frac{1}{d_{\max} - d_{\text{int}}}\right]\right). \tag{31}$$

In particular, we can now decompose $\widehat{\theta}_{\max}$ into a weighted average of two independent estimators as

$$\widehat{\theta}_{\max} = \frac{d_{\text{int}}}{d_{\max}}\widehat{\theta}_{\text{int}} + \frac{d_{\max} - d_{\text{int}}}{d_{\max}}\widehat{\theta}_{\text{comp}}. \tag{32}$$

From (32) and (31) we gather that

$$\mathbb{E}\left[\widehat{\theta}_{\max} \mid \widehat{\theta}_{\text{int}}\right] = \widehat{\theta}_{\text{int}}$$

and

$$\text{Var}\left[\widehat{\theta}_{\max} \mid \widehat{\theta}_{\text{int}}\right] = \left(\frac{d_{\max} - d_{\text{int}}}{d_{\max}}\right)^2 \text{Var}\left[\widehat{\theta}_{\text{comp}} \mid \widehat{\theta}_{\text{int}}\right],$$

hence

$$\widehat{\theta}_{\max} \mid \widehat{\theta}_{\text{int}} \sim \mathcal{N}\left(\widehat{\theta}_{\text{int}}, \frac{1}{r(1-r)}\left[\frac{1}{d_{\text{int}}} - \frac{1}{d_{\max}}\right]\right). \tag{33}$$

Combining (30) and (33) with (17) and (18) yields (19) and (20).

References

Berry, S., Carlin, B., Lee, J., and Muller, P. (2010), *Bayesian Adaptive Methods for Clinical Trials*, Chapman & Hall.

Food and Drug Administration (2019), *Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry*.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013), *Bayesian data analysis*, Chapman & Hall, 3rd ed.

O'Hagan, A., Stevens, J. W., and Campbell, M. J. (2005), "Assurance in clinical trial design," *Pharmaceutical Statistics*, 4, 187–201.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rufibach, K., Burger, H. U., and Abt, M. (2016a), "Bayesian predictive power: choice of prior and some recommendations for its use as probability of success in drug development," *Pharmaceutical Statistics*, 15, 438–446.

Rufibach, K., Jordan, P., and Abt, M. (2016b), "Sequentially updating the likelihood of success of a Phase 3 pivotal time-to-event trial based on interim analyses or external information," *Journal of Biopharmaceutical Statistics*, 26, 191–201.

Saville, B. R., Connor, J. T., Ayers, G. D., and Alvarez, J. (2014), "The utility of Bayesian predictive probabilities for interim monitoring of clinical trials," *Clinical Trials*, 11, 485–493.

Spiegelhalter, D. J. and Freedman, L. S. (1986), "A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion," *Statistics in Medicine*, 5, 1–13.

Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. R. (1986), "Monitoring clinical trials: Conditional or predictive power?" *Controlled Clinical Trials*, 7, 8–17.

Stallard, N., Whitehead, J., and Cleall, S. (2005), "Decision-making in a phase II clinical trial: a new approach combining Bayesian and frequentist concepts," *Pharmaceutical Statistics*, 4, 119–128.

Tang, Z. (2015), "Optimal Futility Interim Design: A Predictive Probability of Success Approach with Time-to-Event Endpoint," *Journal of Biopharmaceutical Statistics*, 25, 1312–1319.

Teerapabolarn, K. (2008), "A Bound on the Binomial Approximation to the Beta Binomial Distribution," *International Mathematical Forum*, 3, 1355–1358.

— (2014), "An improved binomial approximation for the Beta Binomial distribution," *International Journal of Pure and Applied Mathematics*, 97, 511–514.

Tsiatis, A. A. (2014), "Repeated Significance Testing for a General Class of Statistics Used in Censored Survival Analysis," *Journal of the American Statistical Association*, 77, 855–861.