Supplementary Material to "On Bayesian Sequential Clinical Trial Designs"

TIANJIAN ZHOU AND YUAN JI

S.1. FREQUENTIST SEQUENTIAL DESIGNS

We provide a brief review of frequentist sequential designs. Consider the single-arm trial example in Section 1.2. The maximum type I error rate of this sequential testing procedure is given by Equation (1.2). Frequentist group sequential designs are concerned with the specification of the stopping boundaries $\{c_1, \ldots, c_K\}$ such that Equation (1.2) holds for prespecified α, K, and $\{n_1, \ldots, n_K\}$. The solution to Equation (1.2) is not unique, thus restrictions on the stopping boundaries have been considered. We give some examples next.

S.1.1 The Pocock and O'Brien-Fleming Procedures

In the case of equal group sizes (that is, $n_j = jg$ for some g), [10] proposed to use equal stopping boundaries by setting $c_1 = \cdots = c_K = c_P(K, \alpha)$, while [9] suggested decreasing boundaries with $c_j = c_{\text{OBF}}(K, \alpha)\sqrt{K/j}$. In either case, the stopping boundaries can be solved through a numerical search. Note that $\boldsymbol{z} = (z_1, \ldots, z_K)^\top$ follows a multivariate normal distribution with $\text{E}(z_j) = \theta\sqrt{n_j}/\sigma$, $\text{Var}(z_j) = 1$, and $\text{Cov}(z_j, z_{j'}) = \sqrt{n_j/n_{j'}}$ for $j < j'$. Therefore,

$$\alpha = 1 - \Phi_K(\boldsymbol{c}; \boldsymbol{0}, \boldsymbol{\Sigma}),$$

where $\Phi_K(\cdot; \cdot, \cdot)$ is the cumulative distribution function of a multivariate Gaussian random variable, $\boldsymbol{c} = (c_1, c_2, \ldots, c_K)^\top$, and $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{z}$.

S.1.2 The Error Spending Approach

[16] first considered the idea of specifying the error rate spent at each analysis, defined as $\kappa_j = \text{Pr}(z_1 \leq c_1, \ldots, z_{j-1} \leq c_{j-1}, z_j > c_j \mid \theta = 0)$. This represents the probability of rejecting H_0 at stage j but not at any previous stages, given that $\theta = 0$. We have $\alpha = \sum_{j=1}^{K} \kappa_j$. Once the κ_j's are specified, one can successively calculate the stopping boundaries. [6] further extended this idea and suggested to use a function to characterize the rate at which the error rate is spent. This function, denoted by $h(u)$ ($0 \leq u \leq 1$), satisfies $h(0) = 0$ and $h(1) = \alpha$. The κ_j's can be chosen such that $\kappa_j = h(n_j/n_K) - h(n_{j-1}/n_K)$ (with the understanding that $n_0 = 0$). Common choices of $h(u)$ include

$$h_1(u) = \alpha \log\left(1 + (e-1)u\right),$$

$$h_2(u) = 2 - 2\Phi\left(q_{\alpha/2}/\sqrt{u}\right),$$

$$h_3(u) = \alpha u^b \quad \text{for} \quad b > 0.$$

Here, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and $q_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ is the upper $(\alpha/2)$ quantile of the standard normal distribution, $\Phi(q_{\alpha/2}) = 1 - \alpha/2$. It has been shown that in the case of equal group sizes, $h_1(u)$ and $h_2(u)$ produce stopping boundaries similar to those given by Pocock's and O'Brien-Fleming's procedures, respectively. Function h_3 is known as the power spending function and has been studied by [5]. The error spending approach introduces greater flexibility to sequential designs, as the frequency and timing of the interim analyses do not need to be specified in advance.

S.1.3 Stochastic Curtailment Based on Conditional Power

[7] proposed the idea of *stochastic curtailment* that at any point in a sequential clinical trial, if the result at the end of the trial is inevitable, the study can be terminated early. Consider the single-arm trial example. Suppose that at the final analysis, H_0 will be rejected if the final z-statistic $z_K > q_\eta$, where q_η is the upper η quantile of the standard normal distribution. Then, at analysis $j \in \{1, \ldots, K-1\}$, the probability that H_0 will be rejected upon completion of the study, given θ, is given by

$$\text{CP}_j(\theta) = \text{Pr}(z_K > q_\eta \mid \theta, \boldsymbol{y}_j),$$

where $\boldsymbol{y}_j = (y_1, \ldots, y_{n_j})$ is the vector of accumulating data up to analysis j. This is known as the *conditional power*. A simple calculation shows that

$$\text{CP}_j(\theta) = 1 - \Phi\left[\frac{\frac{q_\eta \sigma \sqrt{n_K} - n_j \bar{y}_j}{n_K - n_j} - \theta}{\sigma\sqrt{(n_K - n_j)^{-1}}}\right].$$

If based on current data, H_0 will likely be rejected at the final analysis even if the investigational drug has no treatment effect ($\theta = 0$), then the trial may be stopped early. Mathematically, one may stop the trial early if $\text{CP}_j(0) > \gamma$ for some threshold γ. This is equivalent to

$$z_j > q_\eta\sqrt{n_K/n_j} + q_{1-\gamma}\sqrt{(n_K - n_j)/n_j}.$$

If desirable, one may use different thresholds γ_j's at different interim analyses. An important consideration is the type I error rate of this procedure, but [7] showed that the error rate is upper bounded by η/γ, regardless of the number of interim analyses. Therefore, if η and γ are chosen such that $\eta/\gamma \leq \alpha$, the type I error rate is maintained at or below α, even if interim analyses are conducted at arbitrary times. The stopping boundaries based on this argument are typically conservative. However, if the timing of the interim analyses is specified in advance, tighter stopping boundaries can be constructed by calculating the exact type I error rate numerically.

S.1.4 Analysis at the Conclusion of a Sequential Trial

Once a sequential trial has been completed, it is often of interest to construct a point estimate and a confidence interval for the treatment effect θ. Consider again the single-arm trial example. The results of the trial can be represented by a bivariate random vector (t, z_t), where t denotes the time of stopping,

$$t = \begin{cases} \min\{j : z_j > c_j\}, & \text{if } \exists j \in \{1, \ldots, K\} \text{ s.t. } z_j > c_j; \\ K, & \text{if } z_j \leq c_j \text{ for all } j, \end{cases}$$

and z_t is the corresponding test statistic. Following [1] or [3] (Chapter 8), the density of (t, z_t) is

$$f(t, z_t \mid \theta) = \begin{cases} \tilde{f}(t, z_t \mid \theta), & \text{if } z_t > c_t \text{ or } t = K; \\ 0, & \text{if } z_t \leq c_t \text{ and } t \in \{1, \ldots, K-1\}, \end{cases}$$

where

$$\tilde{f}(1, z_1 \mid \theta) = \phi(z_1 - \theta\sqrt{n_1}/\sigma),$$

and for $t = 2, \ldots, K$,

$$\tilde{f}(t, z_t \mid \theta) = \int_{-\infty}^{c_{t-1}} \tilde{f}(t-1, u \mid \theta) \cdot \frac{\sqrt{n_t}}{\sqrt{n_t - n_{t-1}}} \cdot \\ \phi\left(\frac{z_t\sqrt{n_t} - u\sqrt{n_{t-1}} - (n_t - n_{t-1})\theta/\sigma}{\sqrt{n_t - n_{t-1}}}\right) \mathrm{d}u,$$

with $\phi(\cdot)$ denoting the standard normal density.

The sample mean estimator, $\hat{\theta} = \bar{y}_t$, is a straightforward point estimator for θ. It can be shown that $\hat{\theta}$ is also the maximum likelihood estimator (MLE). However, it is known that the MLE following a sequential trial is biased, and one may correct it by subtracting an estimate of its bias. See, e.g., [18] for more details.

To construct a confidence interval for θ, one needs to define an ordering of the sample space [17, 4, 12]. For example, based on the stage-wise ordering, $(t', z'_{t'})$ is above (t, z_t) if either (i) $t' = t$ and $z'_{t'} > z_t$, or (ii) $t' < t$. In this case, $(t', z'_{t'})$

is indicative of a larger value of θ compared to (t, z_t). It can be shown that

$$\Pr[\text{Observing an outcome above } (t, z_t) \mid \theta]$$

is a continuous and monotonically increasing function of θ for every possible trial outcome (t, z_t) [4]. Thus, one can find unique values θ^L and θ^U which satisfy

$$\Pr[\text{Observing an outcome above } (t, z_t) \mid \theta^{\mathrm{L}}] = \alpha/2,$$
$$\Pr[\text{Observing an outcome above } (t, z_t) \mid \theta^{\mathrm{U}}] = 1 - \alpha/2.$$

The two equations can be solved numerically. Then, $(\theta^{\mathrm{L}}, \theta^{\mathrm{U}})$ is a $100(1-\alpha)\%$ confidence interval for θ.

S.2. THE CALIBRATED BAYESIAN PERSPECTIVE

We present more details about the calibrated Bayesian perspective described in Section 2.3. We consider the setup of an infinite series of single-arm trials (described in Section 1.2) with true but unknown treatment effects $\theta^{(1)}, \theta^{(2)}, \ldots \sim \pi_0(\theta)$. For each trial, patient outcomes $\boldsymbol{y}_K \sim f_0(\boldsymbol{y}_K \mid \theta)$ and are observed sequentially. The Bayesian design with stopping rules given by Equation (2.1) is applied to every trial with a prior model $\pi(\theta)$, a sampling model $f(\boldsymbol{y}_K \mid \theta)$, and threshold values $\{\gamma_1, \ldots, \gamma_K\}$. We are interested in the operating characteristics of the Bayesian design over this infinite series of trials, in particular its FDR and FPR.

S.2.1 Background

We first provide more background on the calibrated Bayesian perspective. [14] called a statistical procedure (conservatively) *calibrated* if the resulting probability statements (at least) have their asserted coverage in repeated practices. Clearly, calibrated procedures are desirable, and Rubin recommended examining operating characteristics to select calibrated Bayesian procedures. Rubin's points were echoed by [8].

The following discussion is adopted from [14]. A Bayesian procedure is calibrated if the model specification is correct, that is, if $f(\boldsymbol{y}_K \mid \theta)\pi(\theta) = f_0(\boldsymbol{y}_K \mid \theta)\pi_0(\theta)$. For example, suppose that $I(\boldsymbol{y}_K)$ is a 95% credible interval for θ under model $f(\boldsymbol{y}_K \mid \theta)\pi(\theta)$, then

$$\frac{\int_{\theta \in I(\boldsymbol{y}_K)} f(\boldsymbol{y}_K \mid \theta)\pi(\theta)\mathrm{d}\theta}{\int_\theta f(\boldsymbol{y}_K \mid \theta)\pi(\theta)\mathrm{d}\theta} =$$
$$\frac{\int_{\theta \in I(\boldsymbol{y}_K)} f_0(\boldsymbol{y}_K \mid \theta)\pi_0(\theta)\mathrm{d}\theta}{\int_\theta f_0(\boldsymbol{y}_K \mid \theta)\pi_0(\theta)\mathrm{d}\theta} = 0.95.$$

The interpretation is that, among the possible θ values from $\pi_0(\theta)$ that might have generated the observed $\boldsymbol{y}_K$ from $f_0(\boldsymbol{y}_K \mid \theta)$, 95% of them belong to $I(\boldsymbol{y}_K)$. Therefore, when the procedure of calculating $I(\boldsymbol{y}_K)$ from $f(\boldsymbol{y}_K \mid \theta)\pi(\theta)$ is repeatedly applied to data drawn from $f_0(\boldsymbol{y}_K \mid \theta)\pi_0(\theta)$, 95%

of the calculated credible intervals will cover the true parameter values. We see that posterior probabilities correspond to frequencies of actual events. Similarly, when we claim $\Pr(\theta > 0 \mid \boldsymbol{y}_K) > 0.95$, it means that among the possible θ values that might have generated $\boldsymbol{y}_K$, more than 95% are positive.

[14] and [11] also demonstrated that when the model specification is correct, the coverage and interpretation of Bayesian statements are still valid under data-dependent stopping rules. For example, if we conclude $\Pr(\theta > 0 \mid \boldsymbol{y}_j) > 0.95$ at any interim analysis j, it means that more than 95% of the possible θ values that might have generated $\boldsymbol{y}_j$ are positive, even if the trial is optionally stopped at analysis j based on the observed data.

Of course, in the presence of model misspecification, the coverage of Bayesian statements is not warranted. In particular, [14] and [11] noted that data-dependent stopping rules increase the sensitivity of Bayesian inference to model specification. Therefore, especially for sequential trial designs, one might want to examine their operating characteristics for a range of plausible $f_0(\boldsymbol{y} \mid \theta)\pi_0(\theta)$ (which may deviate from $f(\boldsymbol{y} \mid \theta)\pi(\theta)$) to select appropriate design parameters.

S.2.2 The False Discovery Rate

We show that the FDR is upper bounded if $f(\boldsymbol{y}_K \mid \theta)\pi(\theta) = f_0(\boldsymbol{y}_K \mid \theta)\pi_0(\theta)$. Note that if $\boldsymbol{y}_K \in \Gamma$, then $\Pr(\theta > 0 \mid \boldsymbol{y}_K) > \gamma_{\min}$. This is because for every $j \in \{1, \ldots, K\}$,

$$\Pr(\theta > 0 \mid \boldsymbol{y}_j) = \int_{\boldsymbol{y}_{j,K}} \Pr(\theta > 0 \mid \boldsymbol{y}_j, \boldsymbol{y}_{j,K}) f(\boldsymbol{y}_{j,K} \mid \boldsymbol{y}_j) \mathrm{d}\boldsymbol{y}_{j,K},$$

where $\boldsymbol{y}_{j,K} = (y_{n_j+1}, \ldots, y_{n_K})$. If $\Pr(\theta > 0 \mid \boldsymbol{y}_K) = \Pr(\theta > 0 \mid \boldsymbol{y}_j, \boldsymbol{y}_{j,K}) \leq \gamma_{\min}$, then $\Pr(\theta > 0 \mid \boldsymbol{y}_j) \leq \gamma_{\min}$ for every j, which contradicts with $\boldsymbol{y}_K \in \Gamma$. Therefore,

$$\begin{aligned}
\mathrm{FDR} &= \frac{\int_{\boldsymbol{y}_K \in \Gamma} \int_{\theta \leq 0} f_0(\boldsymbol{y}_K \mid \theta)\pi_0(\theta)\mathrm{d}\theta\mathrm{d}\boldsymbol{y}_K}{\int_{\boldsymbol{y}_K \in \Gamma} f_0(\boldsymbol{y}_K)\mathrm{d}\boldsymbol{y}_K} \\
&= \frac{\int_{\boldsymbol{y}_K \in \Gamma} \int_{\theta \leq 0} f(\boldsymbol{y}_K \mid \theta)\pi(\theta)\mathrm{d}\theta\mathrm{d}\boldsymbol{y}_K}{\int_{\boldsymbol{y}_K \in \Gamma} f(\boldsymbol{y}_K)\mathrm{d}\boldsymbol{y}_K} \\
&= \frac{\int_{\boldsymbol{y}_K \in \Gamma} \Pr(\theta \leq 0 \mid \boldsymbol{y}_K) \cdot f(\boldsymbol{y}_K)\mathrm{d}\boldsymbol{y}_K}{\int_{\boldsymbol{y}_K \in \Gamma} f(\boldsymbol{y}_K)\mathrm{d}\boldsymbol{y}_K} \\
&\leq 1 - \gamma_{\min}.
\end{aligned}$$

S.2.3 The False Positive Rate

To derive the upper bound of the FPR when $f(\boldsymbol{y}_K \mid \theta)\pi(\theta) = f_0(\boldsymbol{y}_K \mid \theta)\pi_0(\theta)$, we first introduce an inequality under the Bayesian hypothesis testing framework (Section 2.5). Assume

$$\theta \mid H_0 \sim \pi^{(0)}(\theta), \qquad \theta \mid H_1 \sim \pi^{(1)}(\theta),$$

and write $f(\boldsymbol{y}_j \mid H_m) = \int_\theta f(\boldsymbol{y}_j \mid \theta)\pi^{(m)}(\theta)\mathrm{d}\theta$ for $j = 1, \ldots, K$ and $m = 0, 1$. Then, the following inequality holds

for any $0 < \epsilon < 1$ [2]:

$$\Pr\left[\exists j \in \{1, \ldots, K\} : \frac{f(\boldsymbol{y}_j \mid H_1)}{f(\boldsymbol{y}_j \mid H_0)} > \frac{1}{\epsilon} \,\bigg|\, H_0\right] \leq \epsilon,$$

where $\Pr(\cdot \mid H_0) = \int_\theta \Pr(\cdot \mid \theta)\pi^{(0)}(\theta)\mathrm{d}\theta$. This is referred to as a *universal bound* on the probability of observing misleading evidence [13, 15].

In our application, instead of specifying the priors for θ separately under H_0 and H_1, a single prior for θ is specified over the entire parameter space, $\theta \sim \pi(\theta)$. Still, the universal bound is applicable, because $\theta \sim \pi(\theta)$ is equivalent to

$$\Pr(H_0) = \int_{\theta \leq 0} \pi(\theta)\mathrm{d}\theta, \quad \Pr(H_1) = \int_{\theta > 0} \pi(\theta)\mathrm{d}\theta,$$

$$\theta \mid H_0 \sim \pi(\theta \mid \theta \leq 0) = \frac{\pi(\theta) \cdot \mathbf{1}(\theta \leq 0)}{\int_{\theta \leq 0} \pi(\theta)\mathrm{d}\theta},$$

$$\theta \mid H_1 \sim \pi(\theta \mid \theta > 0) = \frac{\pi(\theta) \cdot \mathbf{1}(\theta > 0)}{\int_{\theta > 0} \pi(\theta)\mathrm{d}\theta}.$$

Also, $\Pr(\theta > 0 \mid \boldsymbol{y}_j) = \Pr(H_1 \mid \boldsymbol{y}_j) > \gamma_j$ is equivalent to

$$\frac{f(\boldsymbol{y}_j \mid H_1)}{f(\boldsymbol{y}_j \mid H_0)} > \frac{\gamma_j \cdot \int_{\theta \leq 0} \pi(\theta)\mathrm{d}\theta}{(1 - \gamma_j) \cdot \int_{\theta > 0} \pi(\theta)\mathrm{d}\theta}.$$

Applying the universal bound and notice that $f(\boldsymbol{y}_K \mid \theta)\pi(\theta) = f_0(\boldsymbol{y}_K \mid \theta)\pi_0(\theta)$, we have

$$\begin{aligned}
\mathrm{FPR} &= \frac{\int_{\boldsymbol{y}_K \in \Gamma} \int_{\theta \leq 0} f_0(\boldsymbol{y}_K \mid \theta)\pi_0(\theta)\mathrm{d}\theta\mathrm{d}\boldsymbol{y}_K}{\int_{\theta \leq 0} \pi_0(\theta)\mathrm{d}\theta} \\
&= \int_\theta f(\boldsymbol{y}_K \in \Gamma \mid \theta)\pi(\theta \mid \theta \leq 0)\mathrm{d}\theta \\
&= \Pr\bigg[\exists j \in \{1, \ldots, K\} : \\
&\qquad \frac{f(\boldsymbol{y}_j \mid H_1)}{f(\boldsymbol{y}_j \mid H_0)} > \frac{\gamma_j \cdot \int_{\theta \leq 0} \pi(\theta)\mathrm{d}\theta}{(1 - \gamma_j) \cdot \int_{\theta > 0} \pi(\theta)\mathrm{d}\theta} \,\bigg|\, H_0\bigg] \\
&\leq \Pr\bigg[\exists j \in \{1, \ldots, K\} : \\
&\qquad \frac{f(\boldsymbol{y}_j \mid H_1)}{f(\boldsymbol{y}_j \mid H_0)} > \frac{\gamma_{\min} \cdot \int_{\theta \leq 0} \pi(\theta)\mathrm{d}\theta}{(1 - \gamma_{\min}) \cdot \int_{\theta > 0} \pi(\theta)\mathrm{d}\theta} \,\bigg|\, H_0\bigg] \\
&\leq \frac{(1 - \gamma_{\min}) \cdot \int_{\theta > 0} \pi(\theta)\mathrm{d}\theta}{\gamma_{\min} \cdot \int_{\theta \leq 0} \pi(\theta)\mathrm{d}\theta}.
\end{aligned}$$

REFERENCES

[1] ARMITAGE, P., MCPHERSON, C. and ROWE, B. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)* **132**(2) 235–244.

[2] HENDRIKSEN, A., DE HEIDE, R. and GRÜNWALD, P. (2021). Optional stopping with Bayes factors: a categorization and extension of folklore results, with an application to invariant situations. *Bayesian Analysis* **16**(3) 961–989.

[3] Jennison, C. and Turnbull, B. W. (2000) *Group Sequential Methods with Applications to Clinical Trials.* Chapman & Hall/CRC, Boca Raton.

[4] Kim, K. and DeMets, D. L. (1987). Confidence intervals following group sequential tests in clinical trials. *Biometrics* **43**(4) 857–864.

[5] Kim, K. and DeMets, D. L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* **74**(1) 149–154.

[6] Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**(3) 659–663.

[7] Lan, K. K. G., Simon, R. and Halperin, M. (1982). Stochastically curtailed tests in long–term clinical trials. *Sequential Analysis* **1**(3) 207–219.

[8] Little, R. J. (2006). Calibrated Bayes: a Bayes/frequentist roadmap. *The American Statistician* **60**(3) 213–223.

[9] O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**(3) 549–556.

[10] Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**(2) 191–199.

[11] Rosenbaum, P. R. and Rubin, D. B. (1984). Sensitivity of Bayes inference with data-dependent stopping rules. *The American Statistician* **38**(2) 106–109.

[12] Rosner, G. L. and Tsiatis, A. A. (1988). Exact confidence intervals following a group sequential trial: a comparison of methods. *Biometrika* **75**(4) 723–729.

[13] Royall, R. (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association* **95**(451) 760–768.

[14] Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* **12**(4) 1151–1172.

[15] Sanborn, A. N. and Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review* **21**(2) 283–300.

[16] Slud, E. and Wei, L. J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association* **77**(380) 862–868.

[17] Tsiatis, A. A., Rosner, G. L. and Mehta, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40**(3) 797–803.

[18] Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika* **73**(3) 573–581.

Tianjian Zhou. Department of Statistics, Colorado State University, USA.

E-mail address: `tianjian.zhou@colostate.edu`

Yuan Ji. Department of Public Health Sciences, University of Chicago, USA.

E-mail address: `yji@health.bsd.uchicago.edu`