Web Appendix A: Bayesian Hypothesis Testing

Consider the hypotheses $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ where $\Theta_0 \bigcup \Theta_1 = \Theta$ and $\Theta_0 \bigcap \Theta_1 = \emptyset$. Formal Bayesian hypothesis testing requires the specification of prior probabilities on the hypotheses (e.g. $p(H_i)$ for $i = 0, 1$) and prior distributions for θ specified over the parameter space defined with respect to each of the hypotheses (e.g. $\pi(\theta \big| H_i)$ for $i = 0, 1$).

The posterior probability of hypothesis H_i is given by

$$p(H_i|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|H_i) \cdot p(H_i)}{p(\boldsymbol{D}|H_0) \cdot p(H_0) + p(\boldsymbol{D}|H_1) \cdot p(H_1)}, \tag{1}$$

where $p(\boldsymbol{D}|H_i) = \int_{\Theta_i} p(\boldsymbol{D}|\theta)\pi(\theta|H_i)d\theta$ is the marginal likelihood associated with hypothesis H_i. In practice, most Bayesian hypothesis testing methods are based on the posterior probability of the *event defining* H_i. For this approach, one simply needs to specify a prior $\pi(\theta)$ representing belief about θ and compute the posterior distribution. The posterior probability that $\theta \in \Theta_i$ is given by

$$P(\theta \in \Theta_i|\boldsymbol{D}) = \frac{\int_{\Theta_i} p(\boldsymbol{D}|\theta)\pi(\theta|\theta \in \Theta_i)d\theta \cdot P(\theta \in \Theta_i)}{\sum_{j=0,1} \int_{\Theta_j} p(\boldsymbol{D}|\theta)\pi(\theta|\theta \in \Theta_j)d\theta \cdot P(\theta \in \Theta_j)} \tag{2}$$

where $P(\theta \in \Theta_i) = \int_{\Theta_i} \pi(\theta)d\theta$. We can readily see that the $P(\theta \in \Theta_i|\boldsymbol{D})$ is equal to $p(H_i|\boldsymbol{D})$ if one takes $p(H_i) = P(\theta \in \Theta_i)$ and $\pi\left(\theta \big| H_i\right) = \pi\left(\theta \big| \theta \in \Theta_i\right)$ for $i = 0, 1$. If in fact $\pi(\theta)$ does represent belief about θ, these choices are perhaps the most intuitive and thus we should have no reservation referring to $P(\theta \in \Theta_i|\boldsymbol{D})$ as the probability that hypothesis H_i is true.

Web Appendix B: Parameterizing Flattened and Concentrated Monitoring Priors

Recall the value of the normal density at the mode is $\frac{1}{\sqrt{2\pi}\sigma}$ and note that the value of a generalized normal density at the mode is $\frac{\beta}{2\alpha\Gamma(1/\beta)}$. These are equivalent when $\beta = 2$ and $\alpha = \sqrt{2}\sigma$ (i.e. the normal density is a special case of the generalized normal density at these parameter values). Let $F_{\mu,\alpha,\beta}$ denote the cumulative distribution function of the generalized normal distribution $\mathcal{GN}(\mu,\alpha,\beta)$, which can be expressed as (Griffin, 2018)

$$P(\theta \leqslant q|\mu,\alpha,\beta) = \frac{1}{2} + \frac{\text{sign}(q-\mu)}{2}\int_0^{|q-\mu|^\beta} \frac{w^{1/\beta-1}}{\alpha\Gamma(1/\beta)}\exp\left\{-\left(\frac{1}{\alpha}\right)^\beta w\right\} dw.$$

A flattened or concentrated enthusiastic monitoring prior in the generalized normal family of distributions has density at the mode equal to $k \times \frac{1}{\sqrt{2\pi}\sigma}$. The parameters for the generalized normal distribution $\mathcal{GN}(\mu,\alpha,\beta)$ are derived as follows: μ remains equal to the mode value of θ_1 and α and β are determined to minimize the function

$$\left(F_{\mu,\alpha,\beta}(\theta_0) - \epsilon\right)^2 + \left(\frac{\beta}{2\alpha\Gamma(1/\beta)} - k\frac{1}{\sqrt{2\pi}\sigma}\right)^2$$

with box-constrained optimization (Byrd et al., 1995), where $\sigma = \frac{\theta_1-\theta_0}{\Phi^{-1}(1-\epsilon)}$ is the standard deviation of the default normally distributed enthusiastic monitoring prior. The first term reflects the residual uncertainty that $\theta < \theta_0$, and the second term reflects the density at the mode value. Similarly, the parameters for a flattened or concentrated skeptical monitoring prior are as follows: μ remains equal to the mode value of θ_0 and α and β are determined to minimize the function

$$\left((1 - F_{\mu,\alpha,\beta}(\theta_1)) - \epsilon\right)^2 + \left(\frac{\beta}{2\alpha\Gamma(1/\beta)} - k\frac{1}{\sqrt{2\pi}\sigma}\right)^2,$$

where $\sigma = \frac{\theta_0-\theta_1}{\Phi^{-1}(\epsilon)}$. The parameters for a locally non-informative generalized normal distribution are derived as follows: μ is equal to $\frac{\theta_0+\theta_1}{2}$ (i.e. the midpoint between θ_0 and θ_1) and α and β are determined to minimize the function

$$\left(F_{\mu,\alpha,\beta}\left(\frac{3\theta_0-\theta_1}{2}\right) - \epsilon\right)^2 + \left(\frac{\beta}{2\alpha\Gamma(1/\beta)} - k\frac{1}{\sqrt{2\pi}\sigma}\right)^2,$$

where $\sigma = \frac{2(\theta_1-\theta_0)}{\Phi^{-1}(1-\epsilon)}$ (i.e. reflecting residual uncertainty that $\theta < \frac{\theta_0+\theta_1}{2} - 2(\theta_1 - \theta_0)$) and $k = 1.5$. Note that the standard deviation σ was chosen to be twice that of the normally distributed skeptical or enthusiastic monitoring priors, so that the locally non-informative prior would have greater dispersion around the mode value. Finally, this parameterizing procedure is applicable to a generalized normal distribution truncated to an interval domain (e.g. when θ is a response probability with domain $[0,1]$). In this case, the generalized normal distribution truncated to an interval domain $\Theta = (\theta_{min}, \theta_{max})$ has density equal to $f(\theta) = c \cdot \exp\left\{-\frac{|\theta-\mu|}{\alpha}^{\beta}\right\} I(\theta \in \Theta)$ where $c = \frac{\beta}{2\alpha\Gamma(1/\beta)}(F_{\mu,\alpha,\beta}(\theta_{max}) - F_{\mu,\alpha,\beta}(\theta_{min}))^{-1}$.

Web Appendix C: Type 1 Error Rate Depending on Enrollment Schemes

Recall main article Figure 5 showed Type 1 error properties for the single-arm design. Web Figure 1 shows results from a design that has a longer follow-up period. The interim sample sizes are the same for each monitoring frequency, however, the final sample sizes under the longer follow-up designs are much larger (over 20 patients in follow-up for monitoring frequencies of 8 or fewer, compared to approximately 6 patients in the shorter follow-up designs). The final probability of efficacy criteria being satisfied is generally slightly lower in the longer follow-up design, which is what we would expect since the larger final sample size contains more data and is therefore more likely to be consistent with the underlying true null treatment effect.

[Figure 1 about here.]

Web Appendix D: Robustness of Parameterizations of Monitoring Priors

Recall the analyses done in main article Section 3.1 used a concentrated skeptical prior and default enthusiastic prior. In this section we show the four possible designs using the combinations of skeptical and enthusiastic prior given in main article Figure 1. Web Figures 2 and 3 illustrate how the design operating characteristics change when the enthusiastic prior shifts from default to flattened, with the skeptical prior remaining fixed. Note that in the region between θ_0 and $\frac{\theta_0+\theta_1}{2}$ as the enthusiastic prior shifts from default to flattened, (a) the probability of stopping early for futility increases (b) the probability of inconclusive findings decreases and (c) the intermediate and final sample sizes decrease. This is because the enthusiastic prior gives more mass in this region of θ.

Contrasting Web Figures 2 and 3, we see that the probability of stopping early for efficacy is much higher at θ_0 when the default skeptical prior is used rather than the concentrated skeptical prior. This is because the default skeptical prior has less mass around $\theta = \theta_0$, therefore it is easier to convince the skeptic that $\theta > \theta_0$ under the null result $\theta = \theta_0$.

[Figure 2 about here.]

[Figure 3 about here.]

Web Appendix E: Mixture Inference Prior Weights

Web Figure 4(A) shows the prior-data compatibility assessments $\psi^{(S)}$, $\psi^{(E)}$, $\psi^{(NI)}$ by observed risk difference which were used in main article Section 3.2. As expected, the skeptical and enthusiastic priors show highest compatibility when the observed risk difference matches the corresponding prior mode, and the non-informative prior shows high compatibility for a wide range of θ.

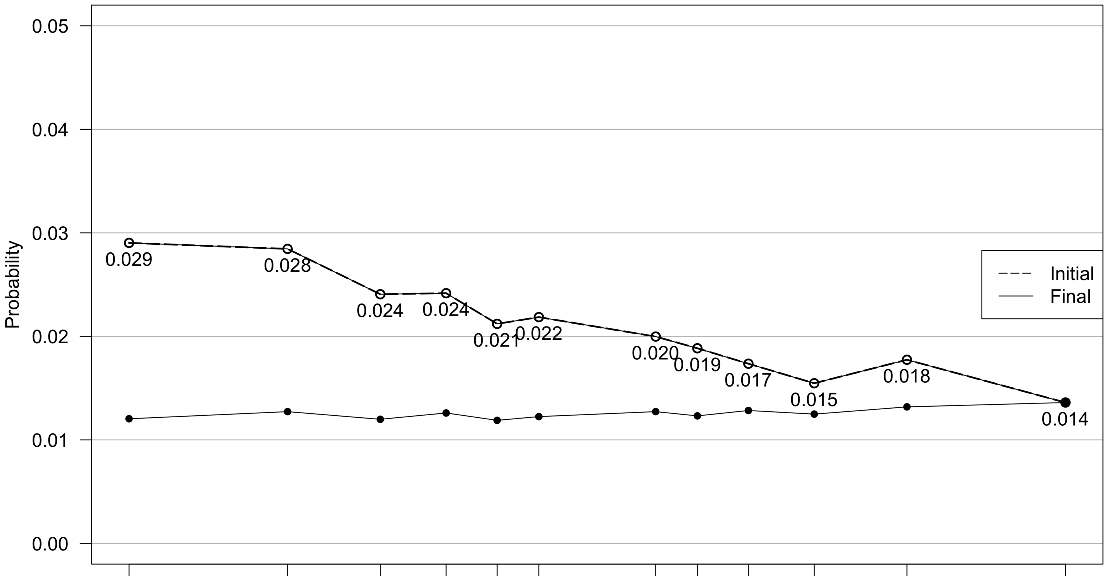
Web Figure 4(B) shows the 3-part mixture inference prior weights ω_S, ω_E, ω_{NI} by observed risk difference. Recall our goal is to create a mixture prior which favors the skeptical or enthusiastic components in areas where high compatibility is demonstrated for those components, and favors the locally non-informative prior if both the skeptical and enthusiastic components show low compatibility. To this end, the skeptical and enthusiastic components have the highest weight when the observed data is aligned with the prior mode, and the locally non-informative prior has highest weight towards extreme values of the observed response difference.

[Figure 4 about here.]

References

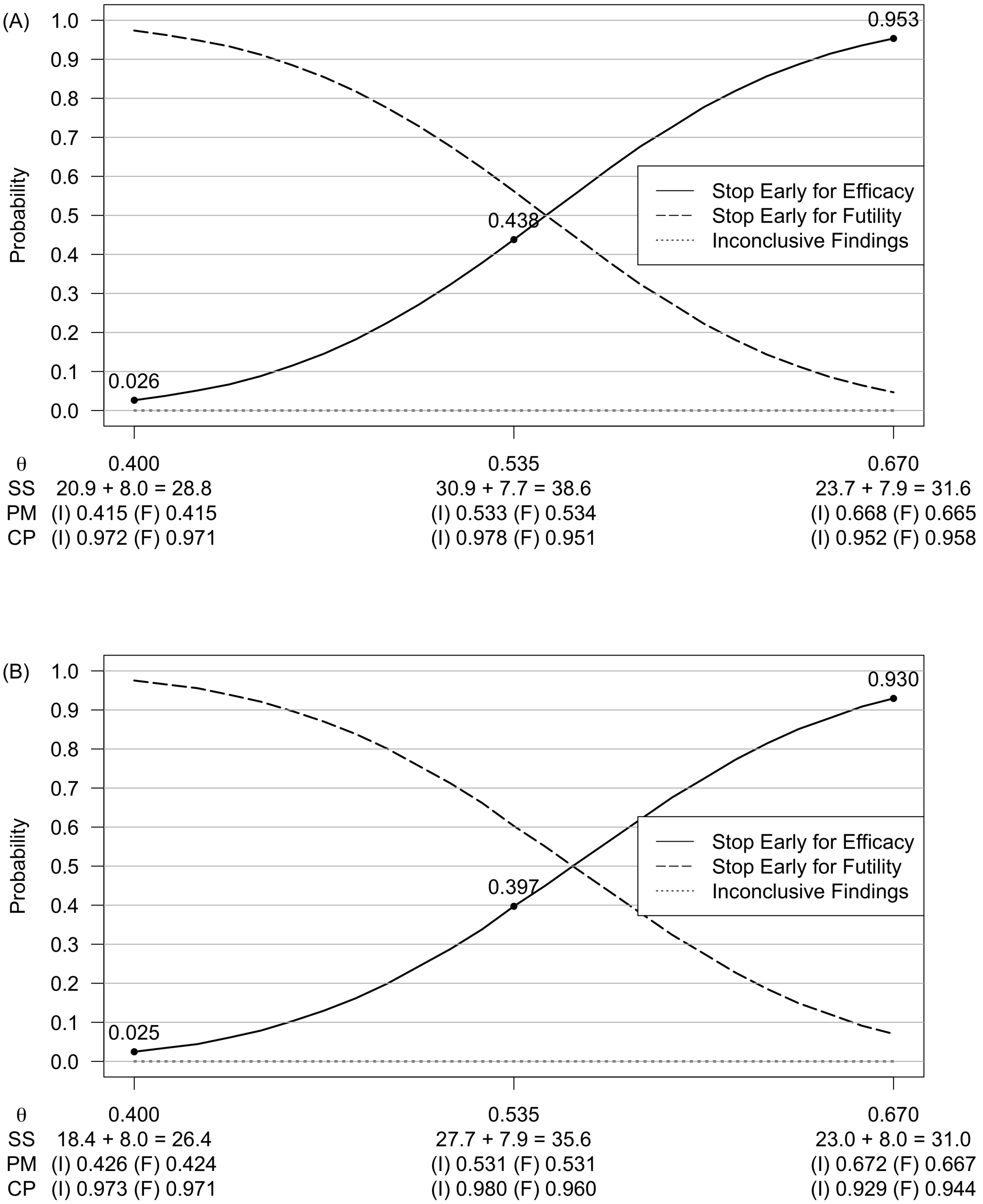
Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* **16,** 1190–1208.

Griffin, M. (2018). Working with the exponential power distribution using gnorm.

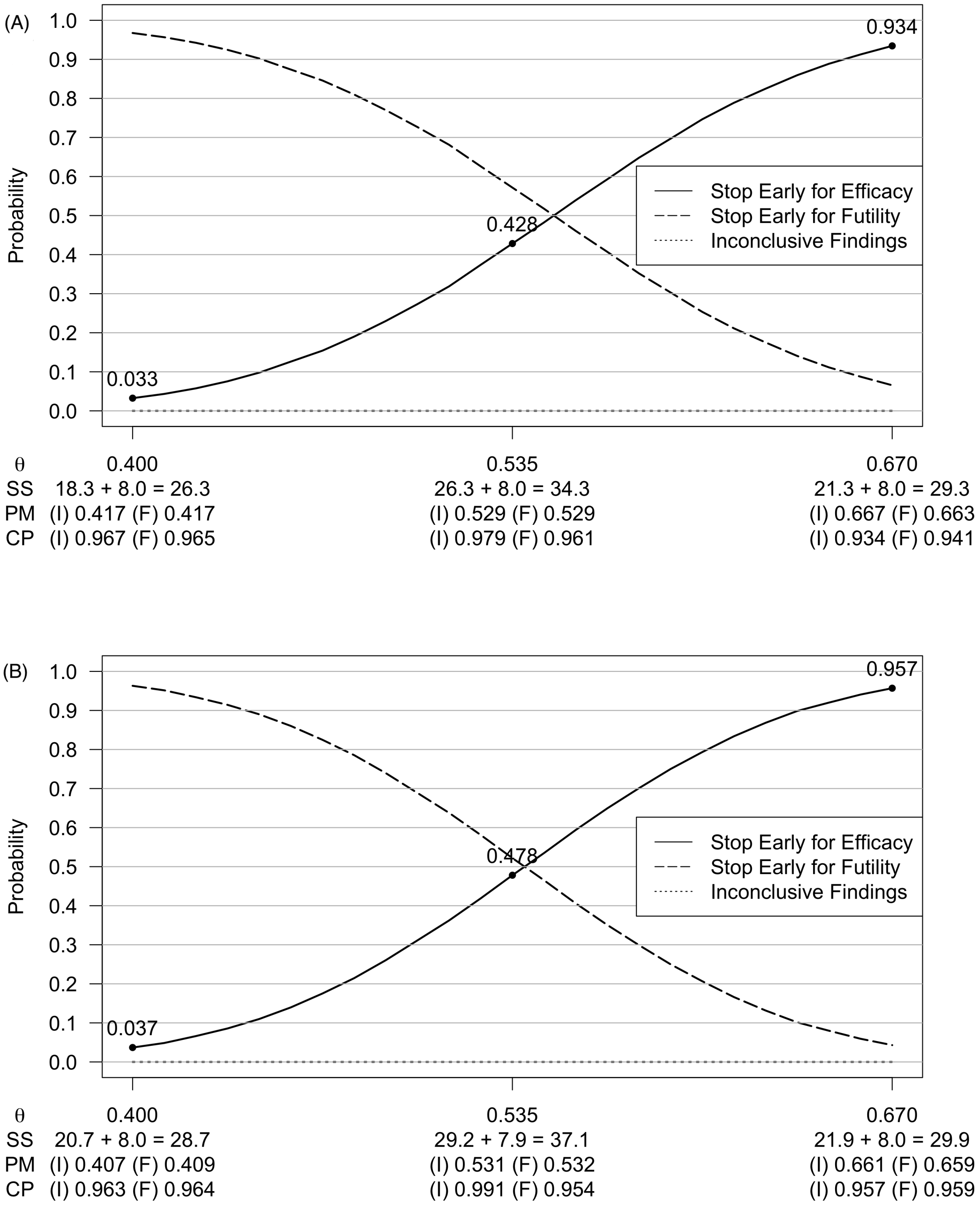


Monitor Freq	1	2	3	4	5	6	10	12	15	20	30	60
SS	20.0	20.9	21.9	22.2	21.7	24.2	24.4	27.2	26.8	29.3	35.0	60.0
Follow-up	30.1	29.8	29.4	29.2	29.6	28.5	28.3	26.6	26.8	25.4	22.6	0.0
SS Final	50.1	50.6	51.3	51.4	51.2	52.6	52.7	53.8	53.7	54.7	57.6	60.0

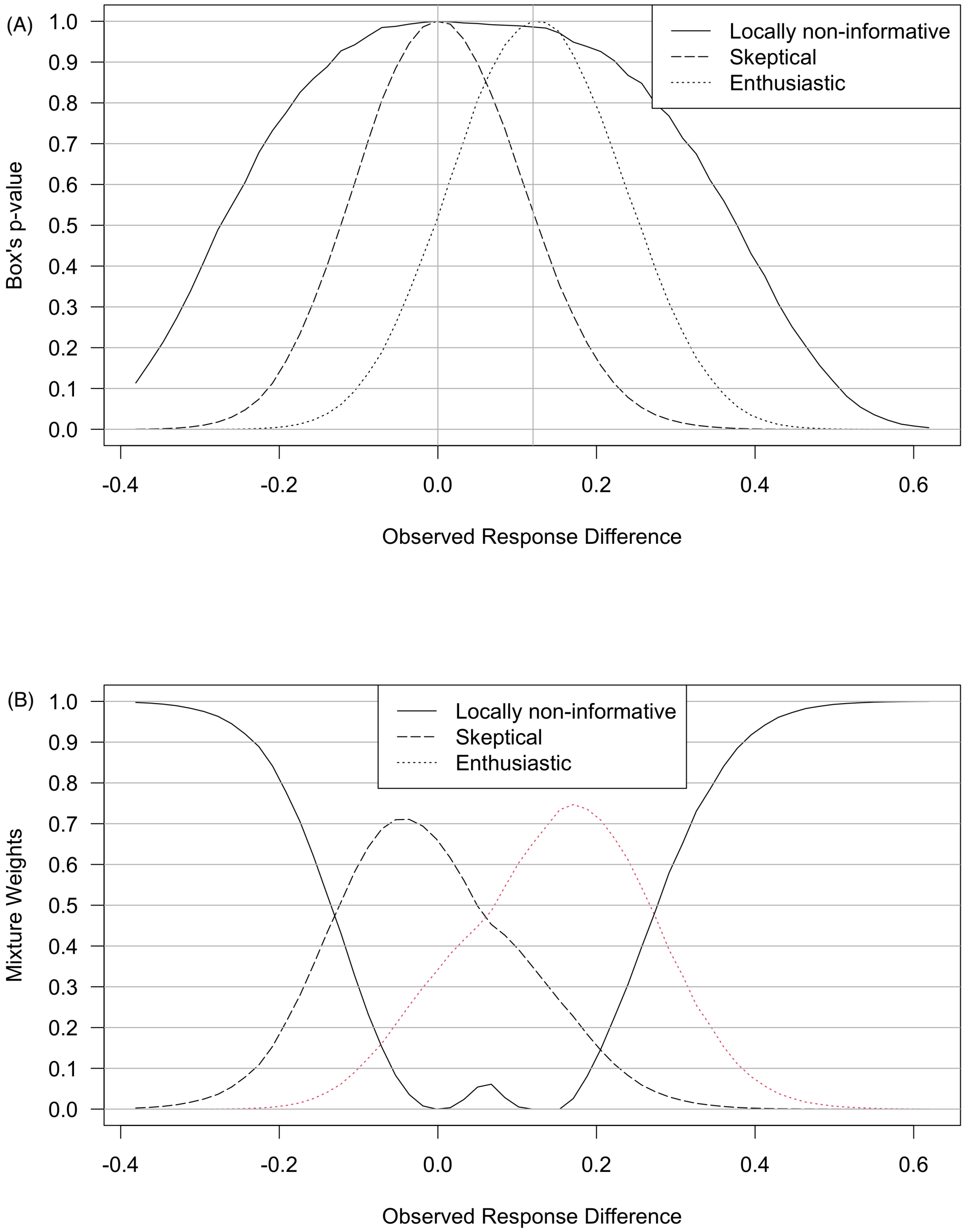
Web Figure 1: Single arm design from main article Section 3.1 with a longer follow-up period. Probability of efficacy criteria being satisfied when $\theta = \theta_0$ (SS; sample size, Monitor Freq; monitoring frequency).



Web Figure 2: Modification of enthusiastic prior parameterization from main article Section 3.1. A, Default enthusiastic prior (main article Figure 1(c)). B, Flattened enthusiastic prior (main article Figure 1(d)). Both designs use concentrated skeptical prior (main article Figure 1(b)).



Web Figure 3: Modification of enthusiastic prior parameterization in main article Section 3.1. A, Default enthusiastic prior (main article Figure 1(c)). B, Flattened enthusiastic prior (main article Figure 1(d)). Both designs use default skeptical prior (main article Figure 1(a)).



Web Figure 4: A; Prior-data compatibility assessments $\psi^{(S)}$, $\psi^{(E)}$, $\psi^{(NI)}$ by observed risk difference. B; 3-part mixture inference prior weights ω_S, ω_E, ω_{NI} by observed risk difference.