An Introduction to Bayesian Sequential Monitoring of Clinical Trials

Matthew A. Psioda

Department of Biostatistics

The University of North Carolina at Chapel Hill

December 14, 2018

U.S. Food and Drug Administration

Outline

Why a Bayesian Approach and What is Sequential Monitoring?

Bayesian vs. Frequentist Approaches – Interpretability

- Bayesian probabilities provide direct information about quantities of interest.
- Frequentist measures provide indirect information.
- Consider the following statements regarding whether treatment A is better then treatment B:
  - Bayesian: The probability that treatment A is better than B is 0.05.
  - Frequentist: The probability is 0.05 of observing data as extreme as or more extreme than that observed here, assuming there is no difference in treatment efficacy and based on the study design.
- Most (All?) practitioners would prefer the Bayesian interpretation and many incorrectly use it even when applying frequentist methods.

Bayesian vs. Frequentist Approaches – Flexibility

- Bayesian inferences can be updated continually as data accumulate and the evidence in favor of a hypothesis at any point time can be evaluated against a single standard.
- Because Bayesian inference obeys the likelihood principle, it does not depend on the stopping rule for data collection or how many times the data have been analyzed before.

  "*It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience.*" – Edwards et al. (1963)
- Strictly speaking, Frequentist inference requires a complete experiment that is conducted exactly according to prespecified design.

Bayesian vs. Frequentist Approaches – Flexibility

- Cornfield (1966) provides and interesting illustration of the rigidity of the Frequentist approach.
  - An experimenter made n observations in the expectation that they would permit the rejection of a particular hypothesis, at some predesignated significance level, say 0.05, and finds that he has not quite attained this critical level.
  - He still believes that the hypothesis is false and asks how many more observations would be required to have reasonable certainty of rejecting the hypothesis if the means observed after n observations are taken as the true values.
  - He also makes it clear that had the original n observations permitted rejection he would simply have published his findings.
  - Under these circumstances there is no amount of additional observation, no matter how large, which would permit rejection at the 0.05 level.

Bayesian vs. Frequentist Approaches – Interpretability

- The Bayesian approach provides a more intuitive framework for sequentially monitoring accumulating data.
- Consider testing two hypotheses H_0 and H_1 where H_1 corresponds to the favorable outcome that treatment A is better than treatment B.
- Summarizing Spiegelhalter et al. (1993) and others: A Bayesian may monitor data continually and stop collection when any of the following criteria have been met:
  - A sufficiently *skeptical person* is convinced H_1 is true.
  - A sufficiently *enthusiastic person* is convinced H_1 is false or that the benefit of treatment is not likely to be what was expected.
  - The probability of *eventually* proving that H_1 is true is sufficiently low.
  - The resources allocated have been exhausted.
- Assuming we can give a satisfactory definition for what it means to be a *skeptical* or *enthusiastic* person, most (all?) people would agree the criteria above are quite intuitive.

Bayesian vs. Frequentist Approaches – Interpretability

- Frequentist group sequential designs allow for sequential analyses of accumulating data while adhering strictly to type I error control requirements through α-spending functions.
- The test statistic is (in general) compared to different thresholds depending on the timing of the analysis making its value difficult to interpret.
- Application group sequential methods require adherence to the prespecified plan regarding how many analyses can be performed and hence are inherently rigid.

Common Complaints Against Bayesian Methods

- Bayesian inference depends on the prior. How does one pick the prior?
  - Most Bayesians advocate performing analyses using multiple priors instead of simply picking one with the possible exception being when the prior is informed by concrete, objective information (i.e., data).

    Much work has been done in this area. See Berger and Berliner (1986), Greenhouse and Waserman (1995), Carlin and Sargent (1996), and Spiegelhalter et al. (1993) for examples.
  - We will illustrate pragmatic approaches for elicitation of skeptical and enthusiastic priors by linking elicitation to the set of hypotheses to be tested.
  - Monitoring accumulating data using skeptical and enthusiastic priors offers a natural way of viewing the data from two relatively extreme, but reasonable perspectives.

Common Complaints Levied Against Bayesian Methods

- Repeated analysis of the data with a goal of early stoppage for efficacy results in an inflated type I error rate. In extreme cases sequential monitoring can lead to *sampling to a foregone conclusion*.
  - Sampling to a foregone conclusion is more of a theoretical concern than a practical one.

    See Spiegelhalter et al. (1993) for some discussion that traces back to Cornfield (1966).
  - The monitoring strategies we discuss have good frequentist properties even when one analyzes data more frequently than is likely to happen in practice (i.e., after each two outcomes are ascertained).

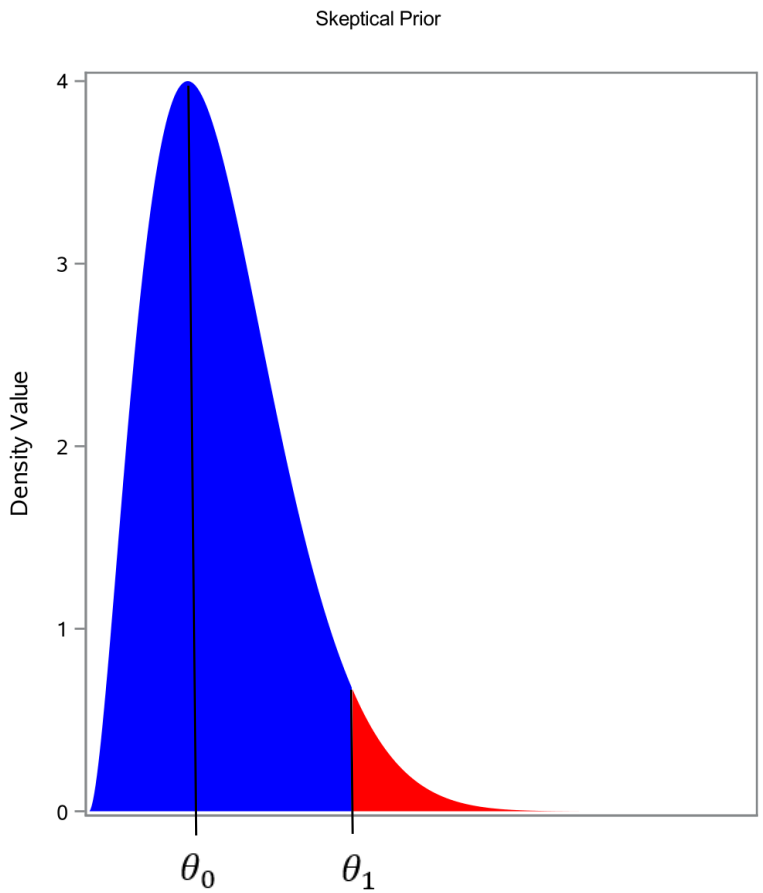
Priors for Sequential Monitoring & Estimation

Skeptical/Enthusiastic Priors

- Consider testing the hypotheses $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ where θ is a treatment effect of interest.
- Suppose an effect $\theta_1 > \theta_0$ is thought to be highly clinically relevant and plausible given what limited data is available.
- Consider a clinical trial with a single analysis and fixed sample size chosen so that there is high probability of proving H_1 when $\theta = \theta_1$.
- A standard Bayesian decision rule would reject H_0 when $P\left(\theta > \theta_0 \middle| \mathbf{D}\right) \geq 0.95$ which will result in a type one error rate of 0.05 (approximately) if $\theta = \theta_0$ when the analysis prior is non-informative (a so-called reference or flat prior).
- The standard evidence threshold for the posterior odds is $0.95/0.05 = 19$. When the prior is flat and prior probabilities on the two hypotheses are 0.5, the posterior odds are also the Bayes Factor.

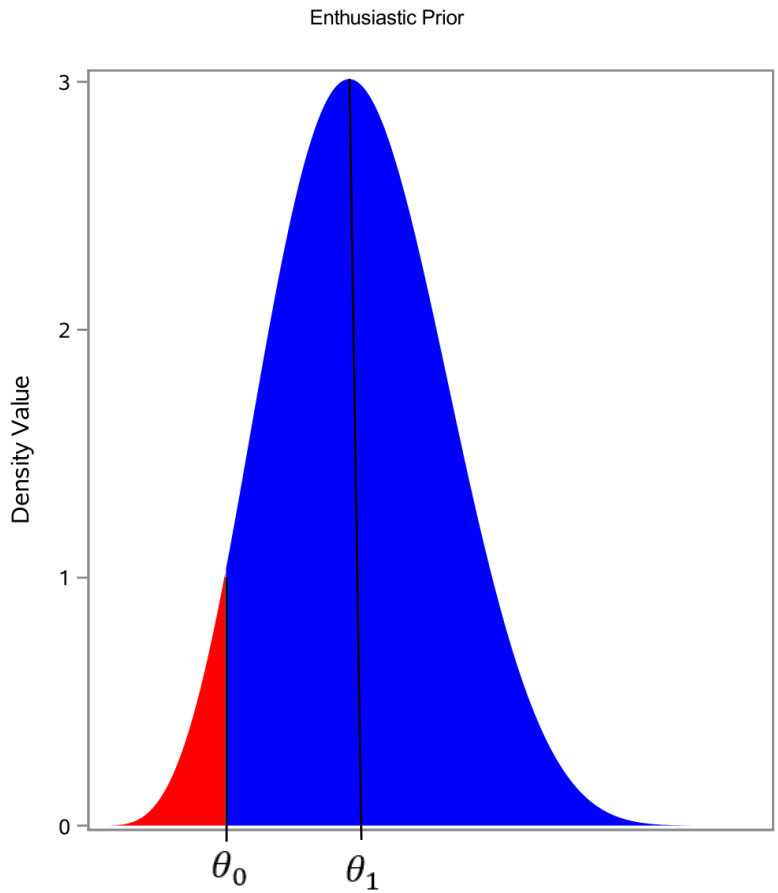
Skeptical/Enthusiastic Priors

- If one were to observe a Bayes Factor exactly equal to $^{0.95}/_{0.05} = 19$, you might say they would be *all but convinced* H_1 is true.
- We define a skeptic as someone who is *all but convinced* that $\theta < \theta_1$ and that believes θ is most likely to equal θ_0.
- Let $\pi_{\text{Skeptical}}$ denote a skeptical prior.



Skeptical/Enthusiastic Priors

- We define an enthusiastic person as someone who is *all but convinced* that $\theta > \theta_0$ and that believes θ is most likely to equal θ_1.
- The prior odds for $P(\theta > \theta_0)$ based on the skeptical prior are $1/19$ and so the Bayes Factor required to convince the skeptic that $\theta > \theta_0$ is $19^2 = 361$.
- Let $\pi_{\text{Enthusiastic}}$ denote an enthusiastic prior.



Point and Interval Estimation

- Let $\pi \equiv \pi(\theta)$ denote a prior distribution for θ and define $p\left(\mathbf{D}\middle|\pi\right) = \int_\theta \mathcal{L}\left(\theta\middle|\mathbf{D}\right)\pi(\theta)\mathrm{d}\theta$ to be the marginal likelihood for the data given the prior π.
- To estimate θ, one can average the posterior means from the analyses using the skeptical and enthusiastic priors.

$$\begin{aligned} \mathrm{E}\left[\theta\middle|\mathbf{D}\right] &= \omega_{\text{Skeptical}} \times \mathrm{E}\left[\theta\middle|\mathbf{D}, \pi_{\text{Skeptical}}\right] \\ &+ \left(1-\omega_{\text{Skeptical}}\right) \times \mathrm{E}\left[\theta\middle|\mathbf{D}, \pi_{\text{Enthusiastic}}\right] \end{aligned}$$

where $\omega_{\text{Skeptical}} = p\left(\mathbf{D}\middle|\pi_{\text{Skeptical}}\right)/\left(p\left(\mathbf{D}\middle|\pi_{\text{Skeptical}}\right)+p\left(\mathbf{D}\middle|\pi_{\text{Enthusiastic}}\right)\right)$

- Equivalently, $\mathrm{E}\left[\theta\middle|\mathbf{D}\right]$ is the posterior mean based on the prior $\pi = 0.5 \times \pi_{\text{Skeptical}} + 0.5 \times \pi_{\text{Enthusiastic}}$.
- Credible intervals can easily be obtained using numeric methods.

Example: Single-Arm Trial /w Binary Endpoint

Example: Single-Arm Trial /w Binary Endpoint

- The goal of the trial is to test the hypotheses:

$$H_0 : \theta \leq 0.20 \text{ versus } H_1 : \theta > 0.20$$

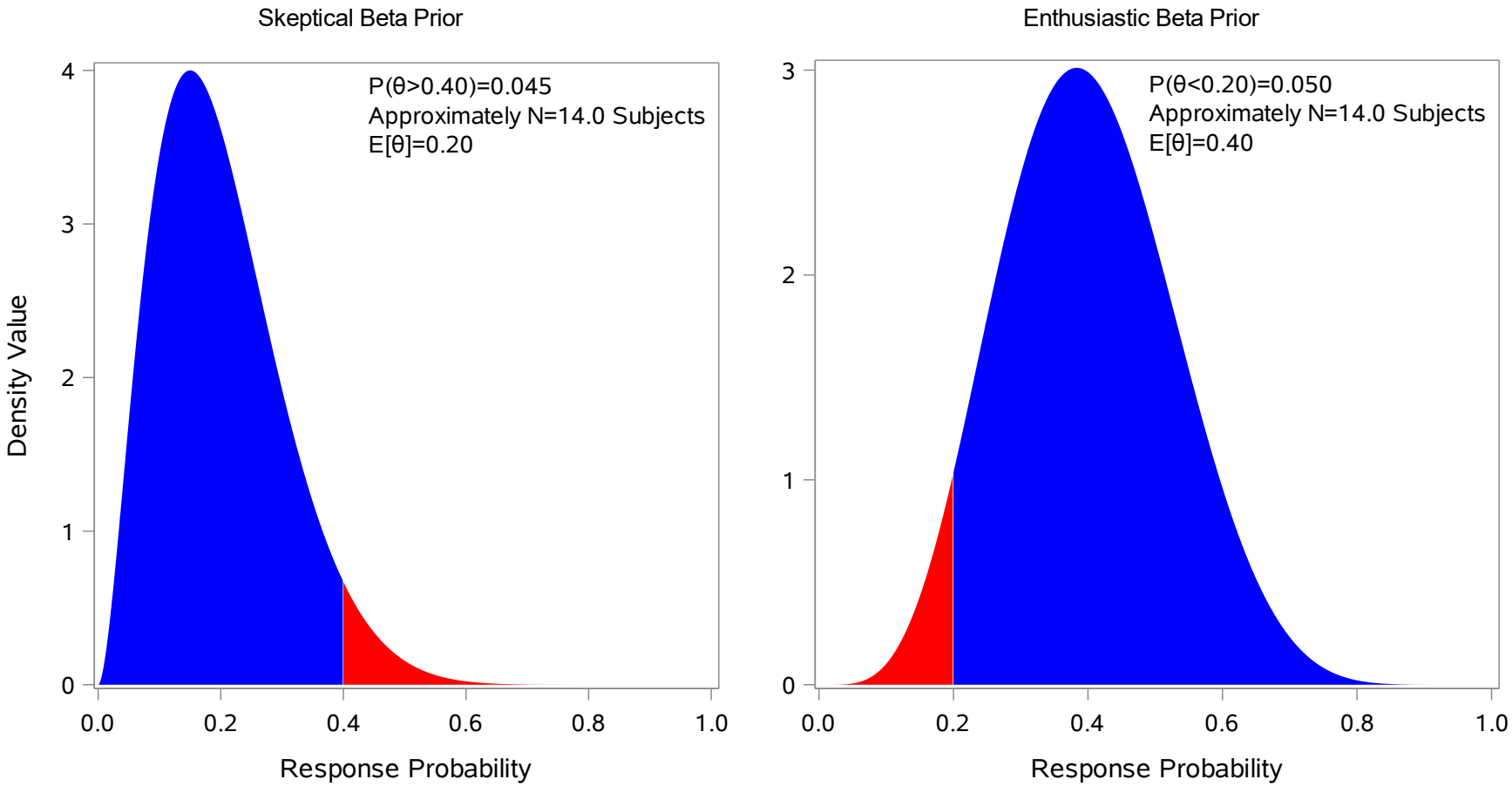
  where θ is the probability of response to treatment.

- Let $\mathbf{D}$ represent the data at an arbitrary point in time.
- The trial will enroll patients until one of three criteria is met:
  - Efficacy Criteria: $P\left(\theta > 0.20 \middle| \mathbf{D}, \pi_{\text{Skeptical}}\right) \geq 0.95$
  - Futility Criteria: $P\left(\theta \leq 0.30 \middle| \mathbf{D}, \pi_{\text{Enthusiastic}}\right) \geq 0.85$
  - Exhausted Resources: $N = 76$ patient outcomes ascertained
- Inferences are updated after every 2 outcomes are ascertained (maximum of 38 analyses).

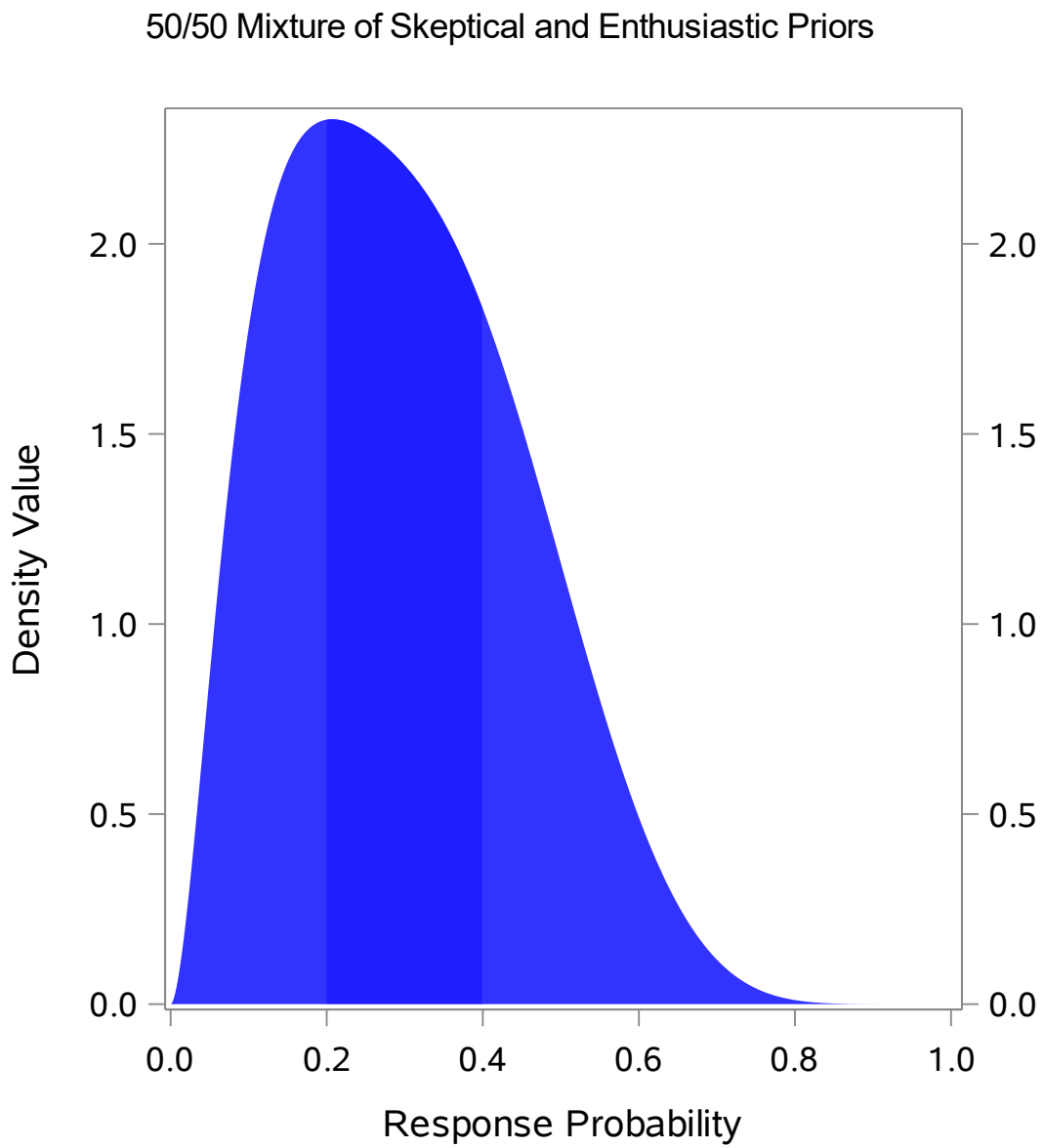
Patient Accrual and Outcome Ascertainment

- For this example we assume that outcomes are ascertained after approximately 4 months of follow-up.
- We consider varying rates of patient accrual.
  - Slow accrual: 2 patients per month on average
  - Fast accrual: 8 patients per month on average
- The study may stop enrollment at a given point in time but all enrolled patients will be followed for outcome ascertainment.

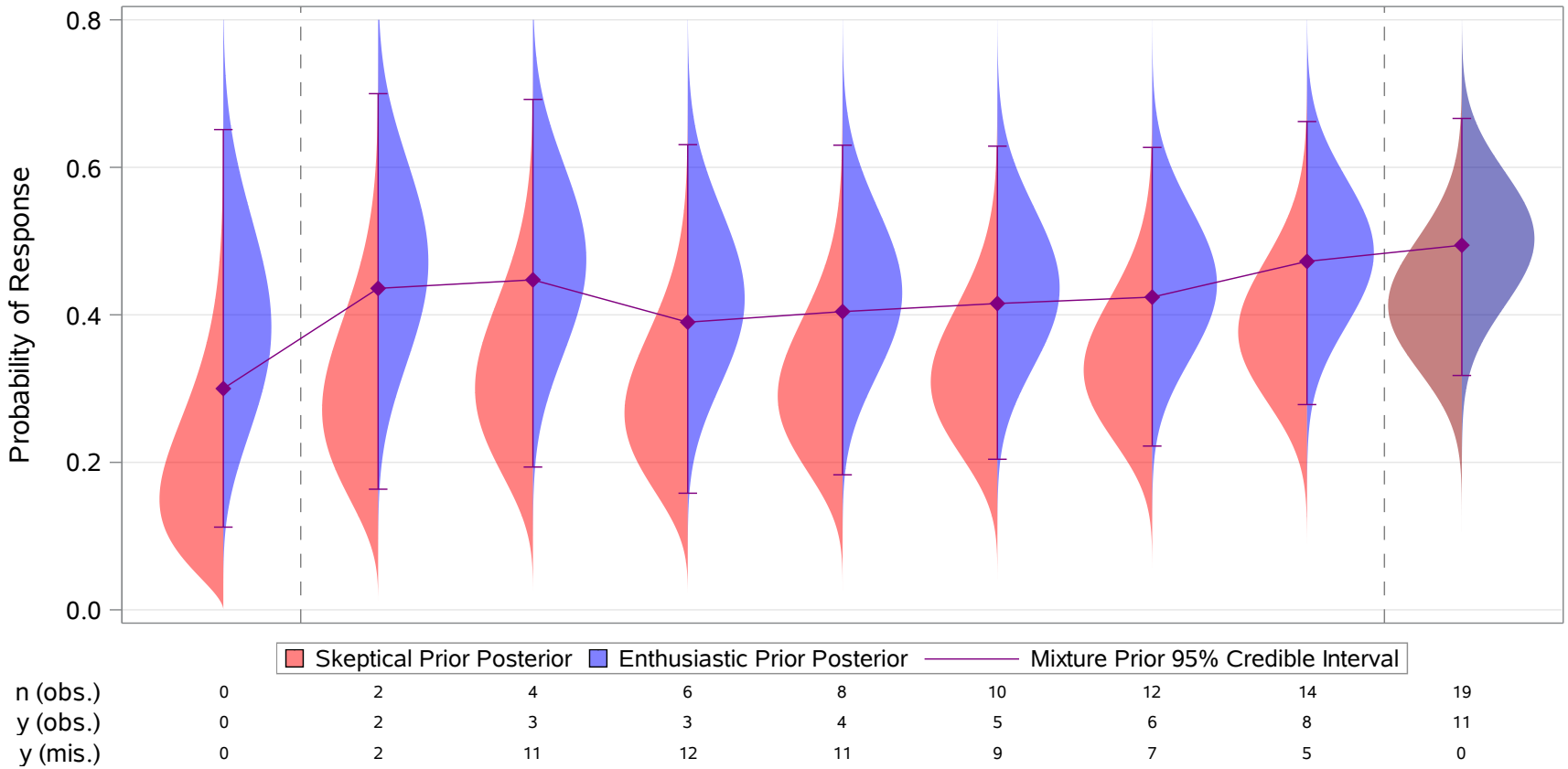
Monitoring Priors



Estimation Prior

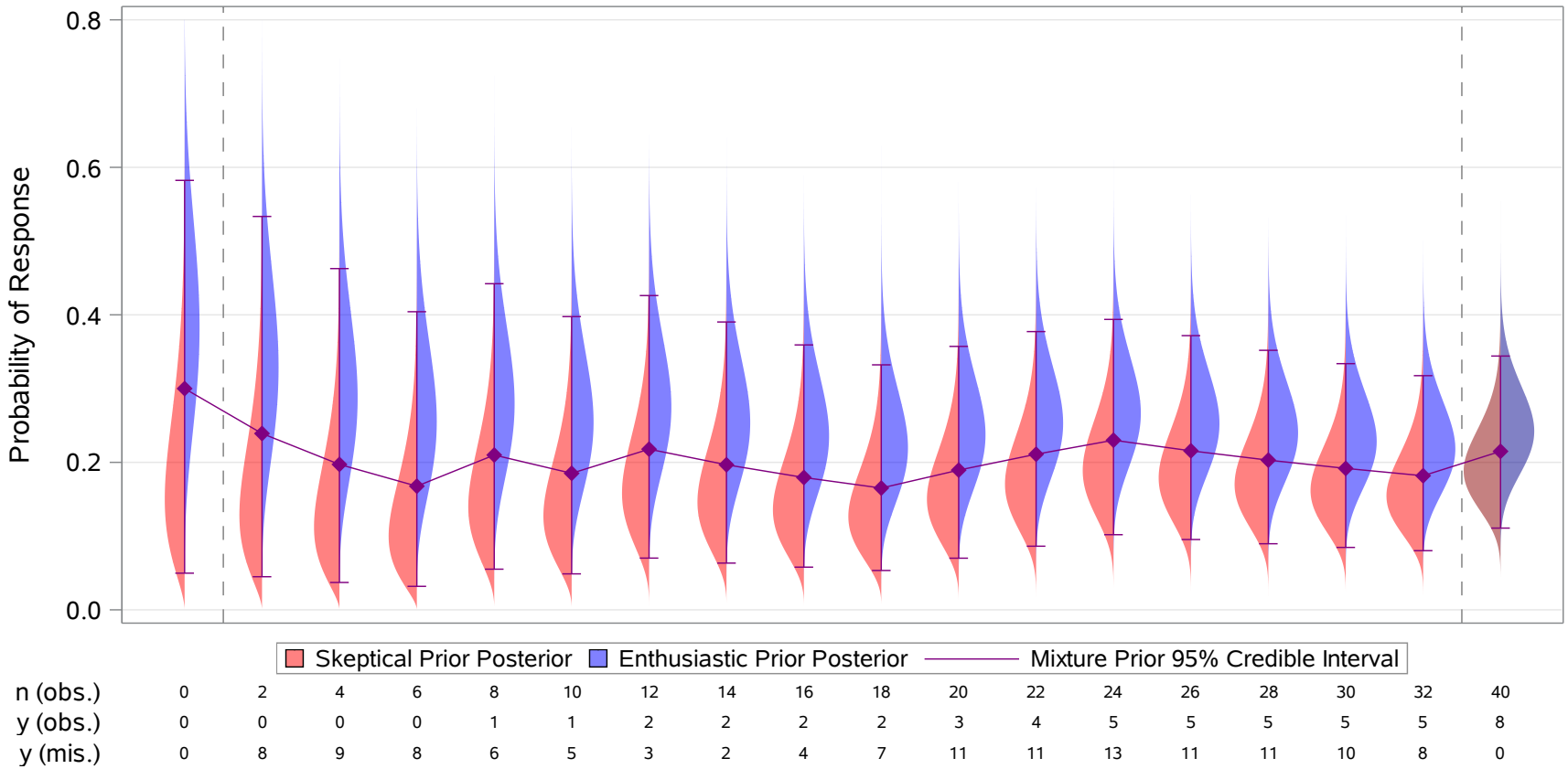


An Example Path – Early Stoppage for Efficacy



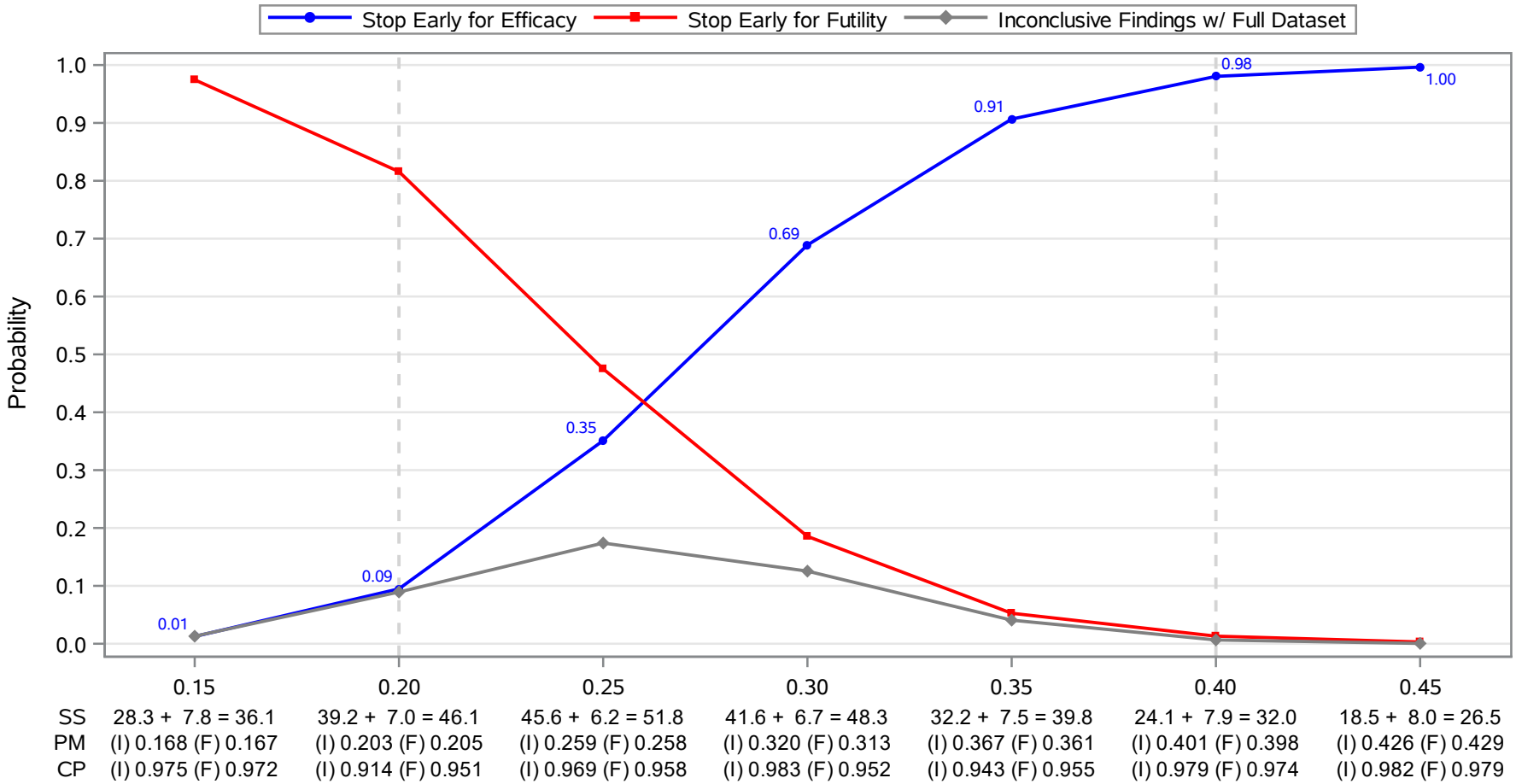
n (obs.)	0	2	4	6	8	10	12	14	19
y (obs.)	0	2	3	3	4	5	6	8	11
y (mis.)	0	2	11	12	11	9	7	5	0

An Example Path – Early Stoppage for Futility



n (obs.)	0	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	40
y (obs.)	0	0	0	0	1	1	2	2	2	2	3	4	5	5	5	5	5	8
y (mis.)	0	8	9	8	6	5	3	2	4	7	11	11	13	11	11	10	8	0

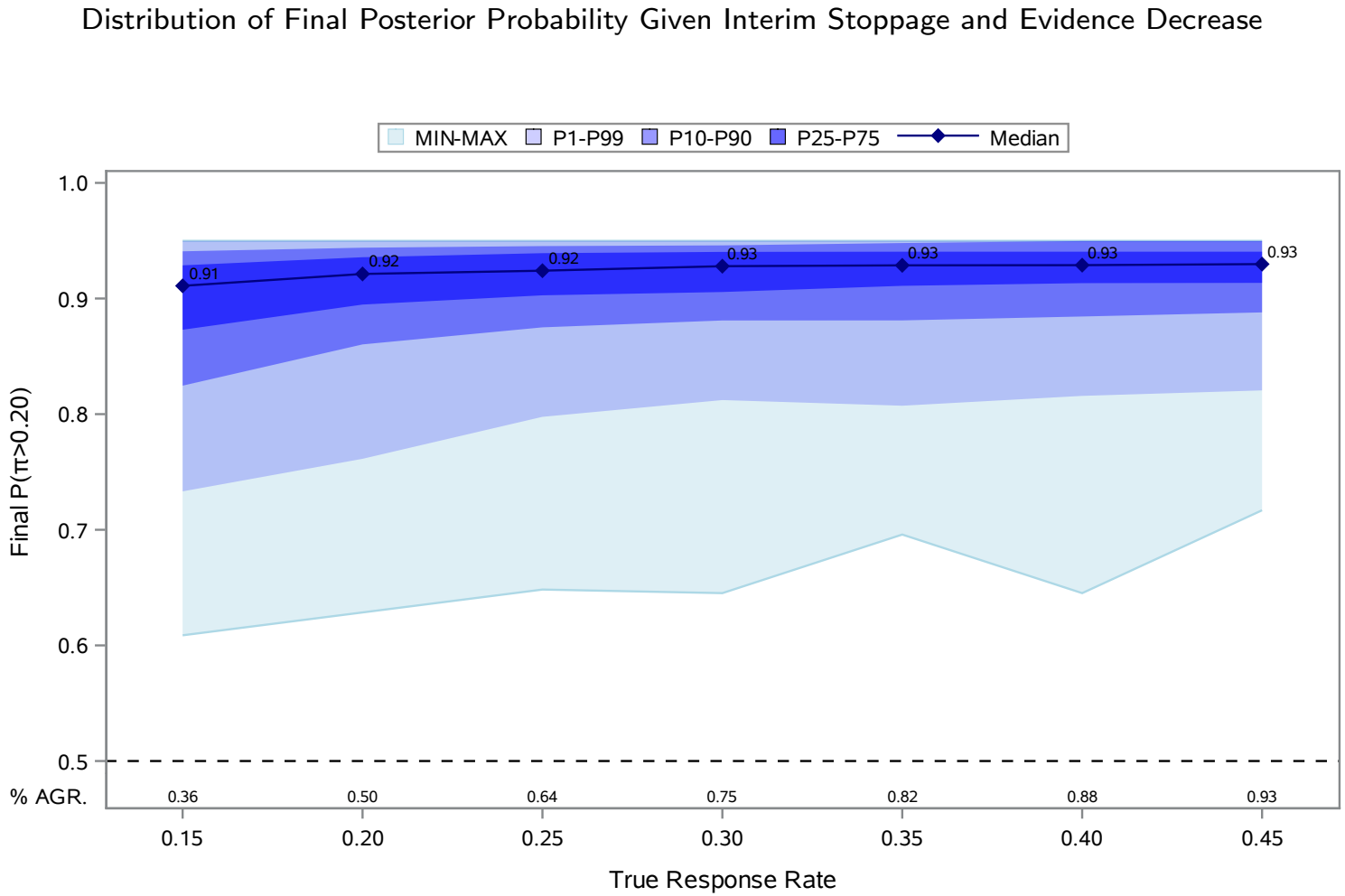
Sequential Design Properties – Slow Accrual +2/month



Notes:

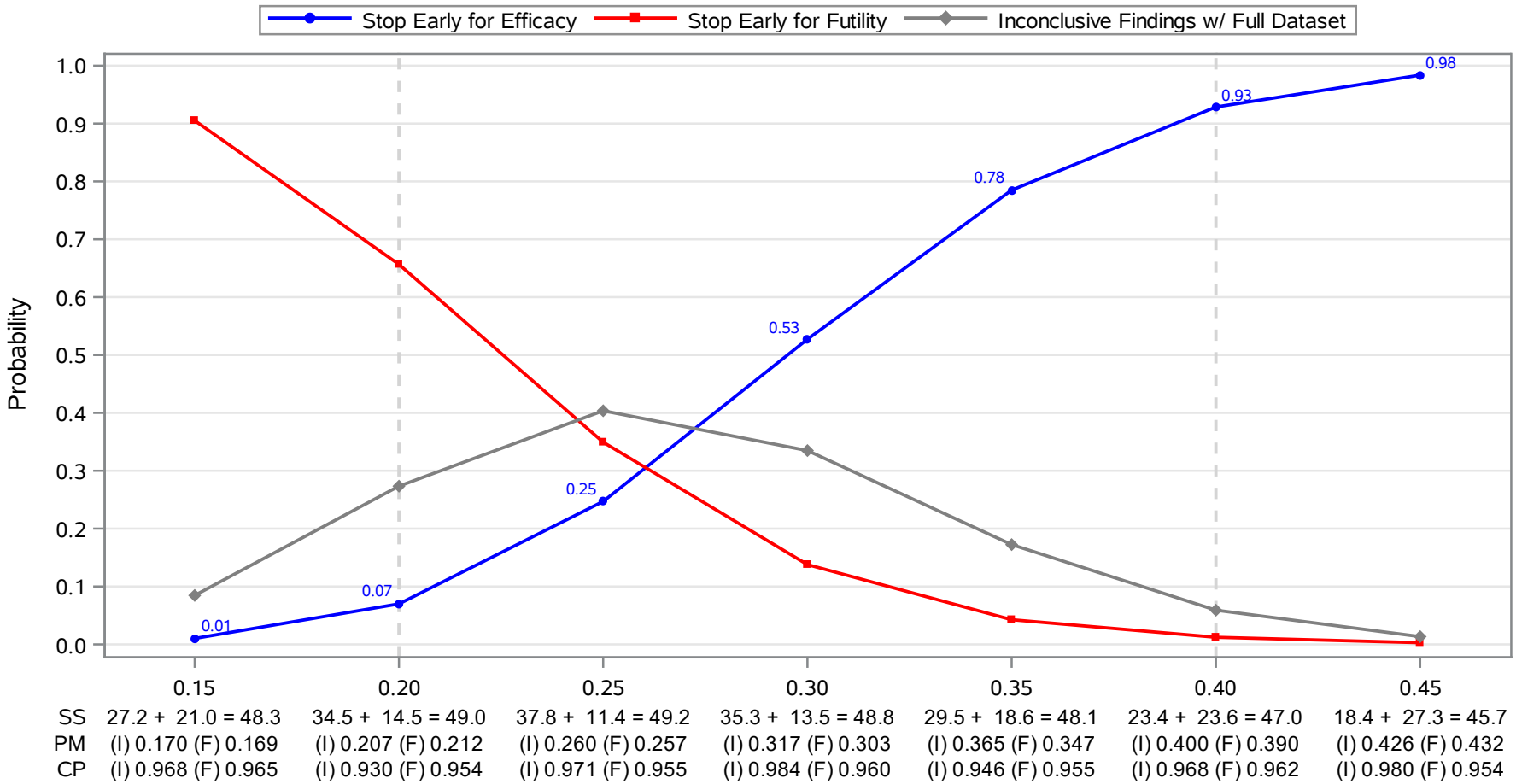
- SS = Avg. # Outcomes Ascertained at Interim + Avg. # Subjects Ongoing at Interim = Avg. Final Enrollment
- PM = Posterior Mean; CP = Coverage Probability; (I) = Interim Analysis; (F) Final Analysis;

Sequential Design Properties – Slow Accrual +2/month



Note: % AGR. = Percent of agreement between final and interim posterior probabilities relative to 0.95 threshold.

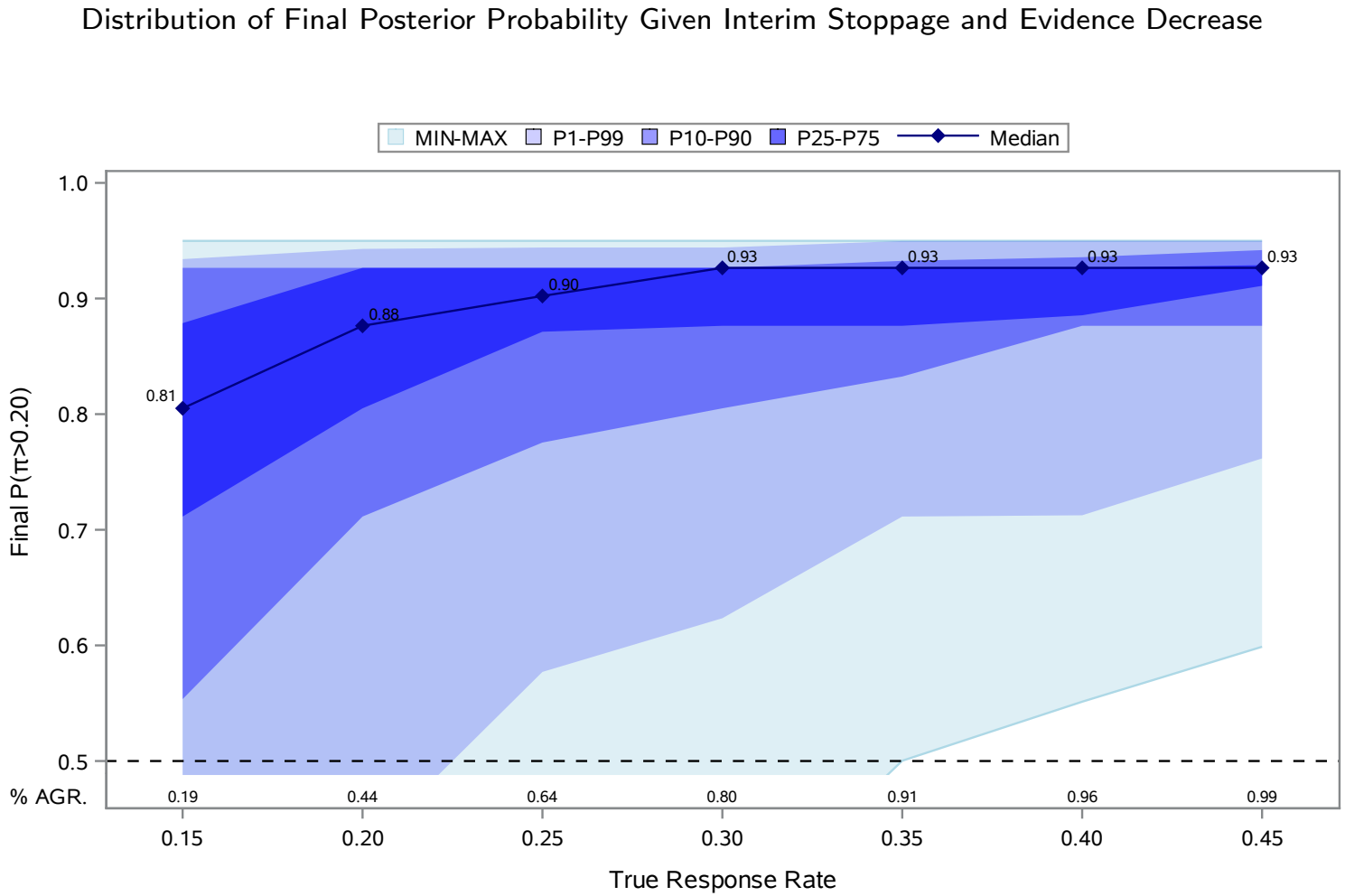
Sequential Design Properties – Fast Accrual +8/month



Notes:

- SS = Avg. # Outcomes Ascertained at Interim + Avg. # Subjects Ongoing at Interim = Avg. Final Enrollment
- PM = Posterior Mean; CP = Coverage Probability; (I) = Interim Analysis; (F) Final Analysis;
- Maximum sample size restricted to $N = 50$ patients.

Sequential Design Properties – Fast Accrual +8/month



Note: % AGR. = Percent of agreement between final and interim posterior probabilities relative to 0.95 threshold.

Trial Monitoring using Predictive Probability of Success

Futility Analysis using Predictive Probability of Success

- As an alternative strategy to futility analysis, one can monitor the **Probability of Success** (PoS) for the trial.
- Let $\pi\left(\theta\middle|\mathbf{D}\right) \equiv \pi\left(\theta\middle|\mathbf{D}, \pi_{\text{Skeptical}}\right)$ denote the posterior distribution for θ based on the skeptical prior $\pi_{\text{Skeptical}}$ and current data $\mathbf{D}$.
- Let $\psi \equiv \psi\left(\mathbf{D}, \pi_{\text{Skeptical}}\right)$ denote the PoS which is given as follows:

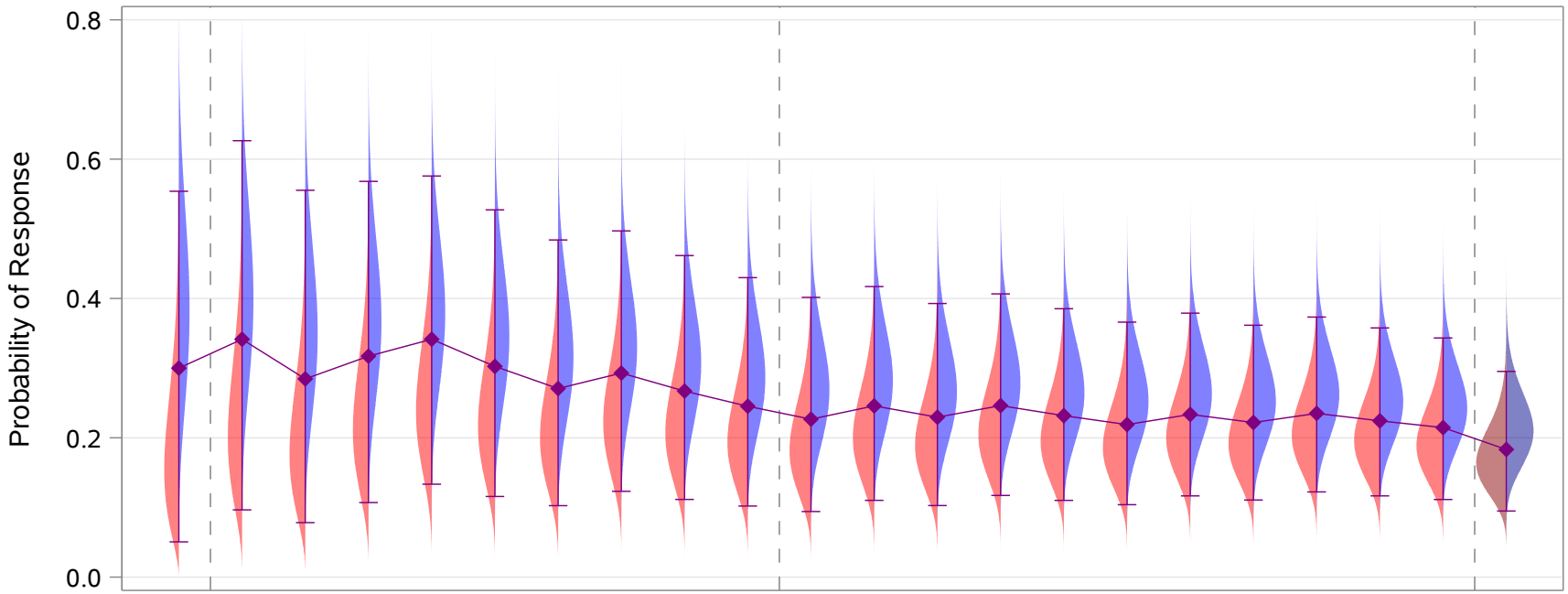
$$\psi \quad = \quad \text{E}\left[1\left\{P\left(\theta > 0.20\middle|\mathbf{D}_1, \mathbf{D}, \pi_{\text{Skeptical}}\right) \geq 0.95\right\}\right]$$

  where the expectation is taken with respect to the **posterior predictive distribution** $p\left(\mathbf{D}_1\right)$ for future data $\mathbf{D}_1$ (ongoing + subjects yet to enroll).

$$p\left(\mathbf{D}_1\right) = \int p\left(\mathbf{D}_1\middle|\theta\right) \times \pi\left(\theta\middle|\mathbf{D}\right) \text{d}\theta$$

- One may stop enrollment if ψ is sufficiently small (i.e., $\psi < 0.05$).

An Example Path – Futility – $\psi < 0.05$



PoS	0.001	0.326	0.223	0.324	0.428	0.320	0.228	0.314	0.226	0.156	0.102	0.153	0.101	0.149	0.098	0.061	0.093	0.057	0.087	0.053	0.030	0.001
n (obs.)	0	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	48
y (obs.)	0	1	1	2	3	3	3	4	4	4	4	5	5	6	6	6	7	7	8	8	8	8
y (mis.)	0	8	7	5	4	2	7	10	10	10	13	13	11	9	7	6	5	4	6	6	8	0

Example: Two-Arm Non-Inferiority Trial

Example: Two-Arm Non-Inferiority Time-to-Event Trial

- Assume subjects in arm $j = 0, 1$ have survival times following an Exponential (λ_j) distribution.
- Let $\theta = \lambda_1/\lambda_0$ be the hazard ratio for treatment versus control.
- The goal of the trial is to sequentially test the following hypotheses:

$$\text{Stage I} \rightarrow H_0 : \theta \geq 1.8 \text{ versus } H_1 : \theta < 1.8$$

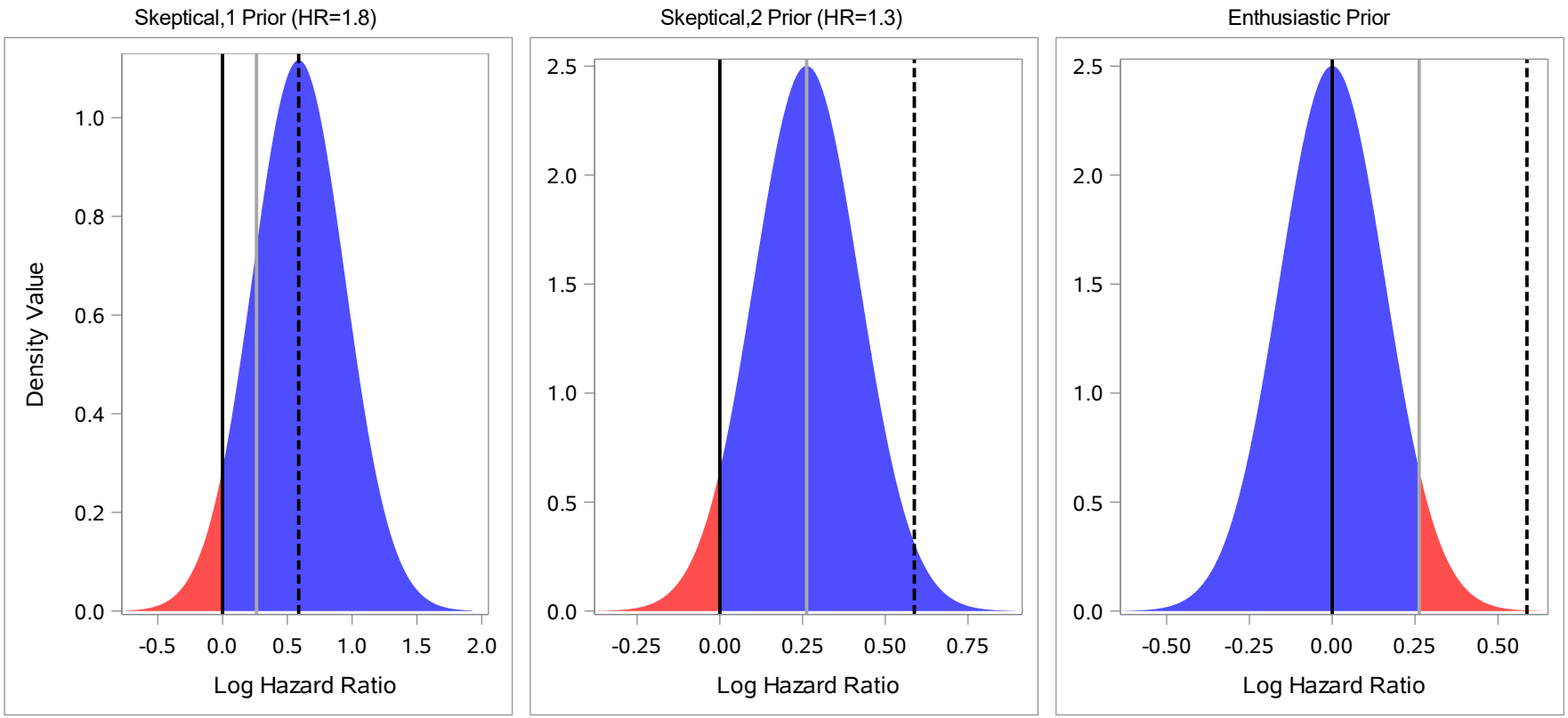
$$\text{Stage II} \rightarrow H_0 : \theta \geq 1.3 \text{ versus } H_1 : \theta < 1.3$$

- The trial continues until Criteria A & B are met or Criteria C is met:
  - Criteria A (Stage I Success): $P\left(\theta < 1.8 \middle| \mathbf{D}, \pi_{\text{Skeptical},1}\right) \geq 0.975$
  - Criteria B (Stage II Success): $P\left(\theta < 1.3 \middle| \mathbf{D}, \pi_{\text{Skeptical},2}\right) \geq 0.975$
  - Criteria C (Futility): $P\left(\theta \geq 1.3 \middle| \mathbf{D}, \pi_{\text{Enthusiastic}}\right) \geq 0.50$

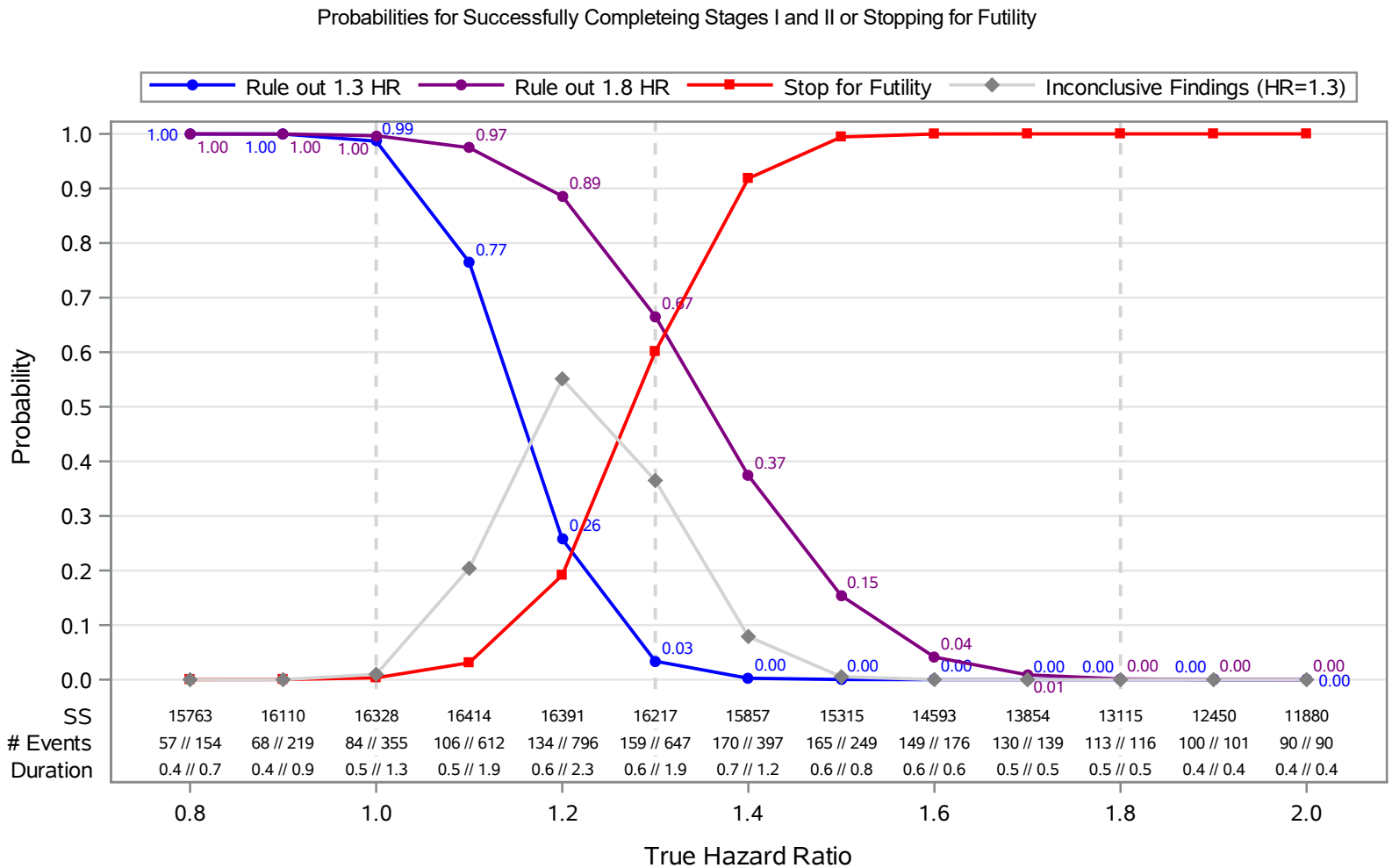
Patient Accrual and Outcome Ascertainment

- Similar to the SAVOR cardiovascular outcomes trial (Scirica et al., 2011), we assume:
  - An annualized event rate of approximately 2.1%
  - Patients accrue over a 20 month period with enrollment in that period increasing over time linearly.
  - Up to $16,500$ subjects enrolled to accrue up to $1,040$ events.
- The SAVOR trial was powered to demonstrate cardiovascular benefit (i.e., $\theta < 1$) but we do not consider that goal here.
- Due to requirements for primary endpoint adjudication, we assume that analyses are only possible after each additional 20 events are accrued.
  - Maximum number of analyses $\rightarrow 1040/20 = 52$.
- The study may stop enrollment at a given point and initiate study termination procedures but subjects are followed-up for approximately 3 months beyond study termination initiation.

Monitoring Priors



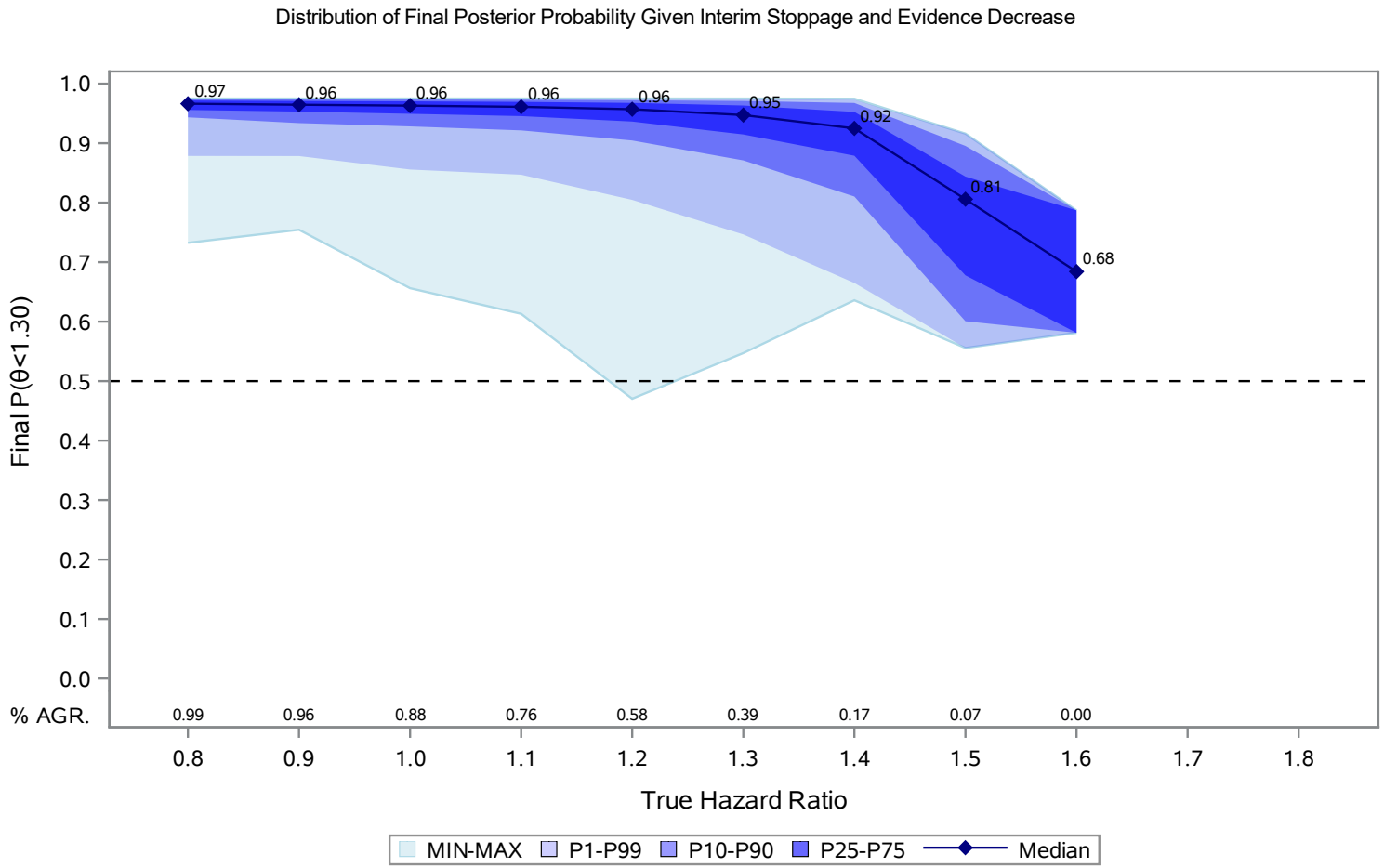
CVOT Design Properties



Notes:

- SS = Avg. # of subjects Enrolled.
- # Events and Duration correspond to analysis that rules out 1.8 hazard ratio and 1.3 hazard ratio.

CVOT Design Properties



Note: % AGR. = Percent of agreement between final and interim posterior probabilities relative to 0.95 threshold.

Closing Remarks

Closing Remarks

- The Bayesian approach provides an intuitive framework for continually updating inferences about a parameter in a sequential trial. Moreover, the quantities on which inference is based are easily interpretable.
- Frequentist properties are of secondary importance to Bayesians because Bayesian focus solely on the likelihood and prior and what they combine to tell us about the parameter in question. This is liberating as it allows for more flexibility in how trials are conducted.
- Nonetheless, good frequentist properties are desirable and so it is no surprise that reasonable Bayesian approaches exhibit them.
- For more information on "Why Bayes", see Berry (1993) and Spiegelhalter et al. (1993).
- For an introductory tutorial of Bayesian data monitoring, see Fayers et al. (1997).

References I

[1] Berger, J. and Berliner, L. M. (1986), "Robust Bayes and Empirical Bayes Analysis with Epsilon-Contaminated Priors," *Ann. Statist.*, 14, 461–486.

[2] Berry, D. A. (1993), "A case for bayesianism in clinical trials," *Statistics in Medicine*, 12, 1377–1393.

[3] Carlin, B. P. and Sargent, D. J. (1996), "Robust Bayesian Approaches for Clinical Trial Monitoring," *Statistics in Medicine*, 15, 1093–1106.

[4] Cornfield, J. (1966), "Sequential Trials, Sequential Analysis and the Likelihood Principle," *The American Statistician*, 20, 18–23.

[5] Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian statistical inference for psychological research." *Psychological Review*, 70, 193–242.

References II

[6] Fayers, P. M., Ashby, D., and Parmar, M. K. B. (1997), "Tutorial in Biostatistics: Bayesian Data Monitoring in Clinical Trials," *Statistics in Medicine*, 16, 1413–1430.

[7] Greenhouse, J. B. and Waserman, L. (1995), "Robust bayesian methods for monitoring clinical trials," *Statistics in Medicine*, 14, 1379–1391.

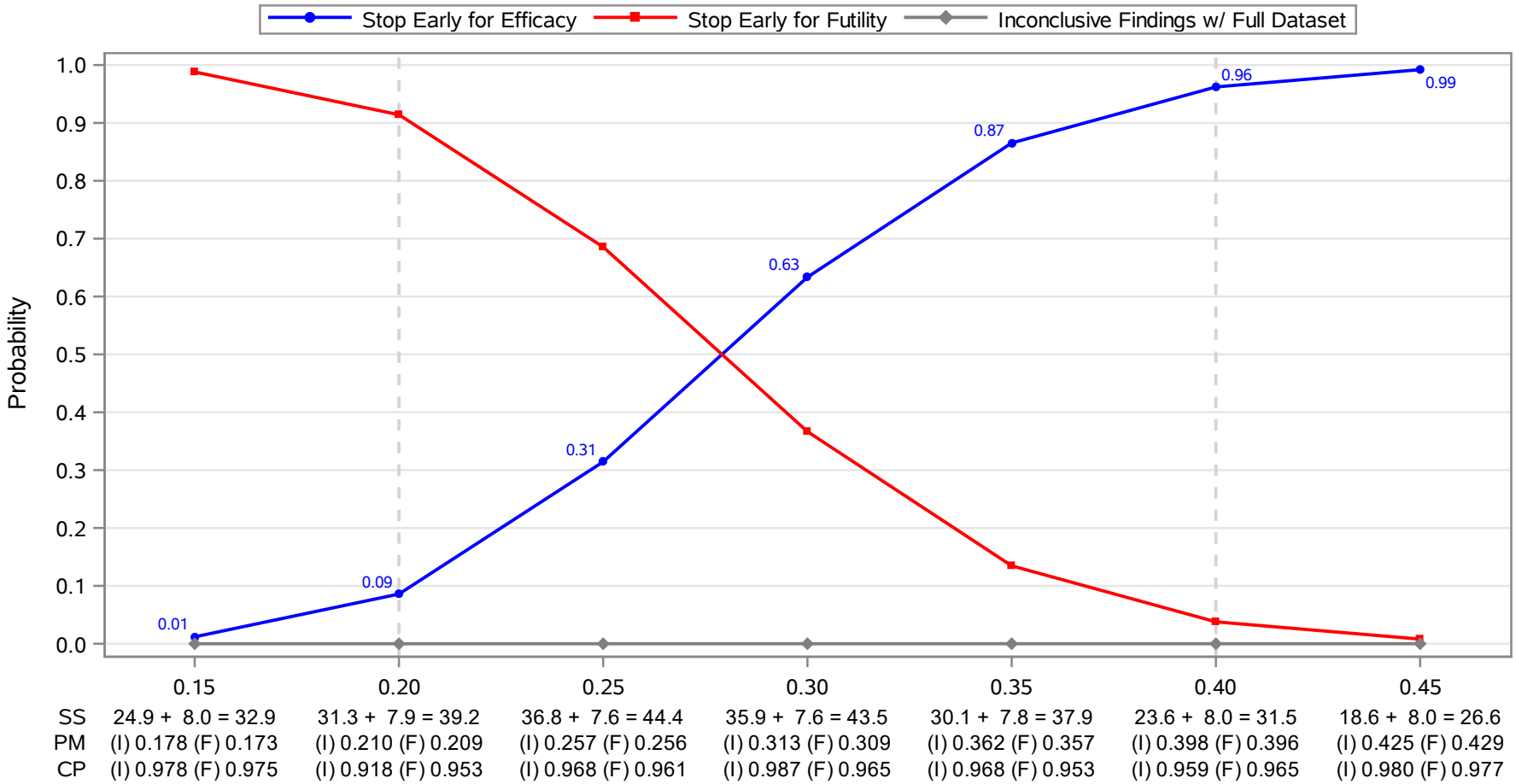
[8] Scirica, B. M., Bhatt, D. L., Braunwald, E., Steg, P. G., Davidson, J., Hirshberg, B., Ohman, P., Price, D. L., Chen, R., Udell, J., and Raz, I. (2011), "The design and rationale of the Saxagliptin Assessment of Vascular Outcomes Recorded in patients with diabetes mellitusThrombolysis in Myocardial Infarction (SAVOR-TIMI) 53 Study," *American Heart Journal*, 162, 818–825.e6.

[9] Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. B. (1993), "Applying Bayesian ideas in drug development and clinical trials," *Statistics in Medicine*, 12, 1501–1511.

THANK YOU!

Questions?

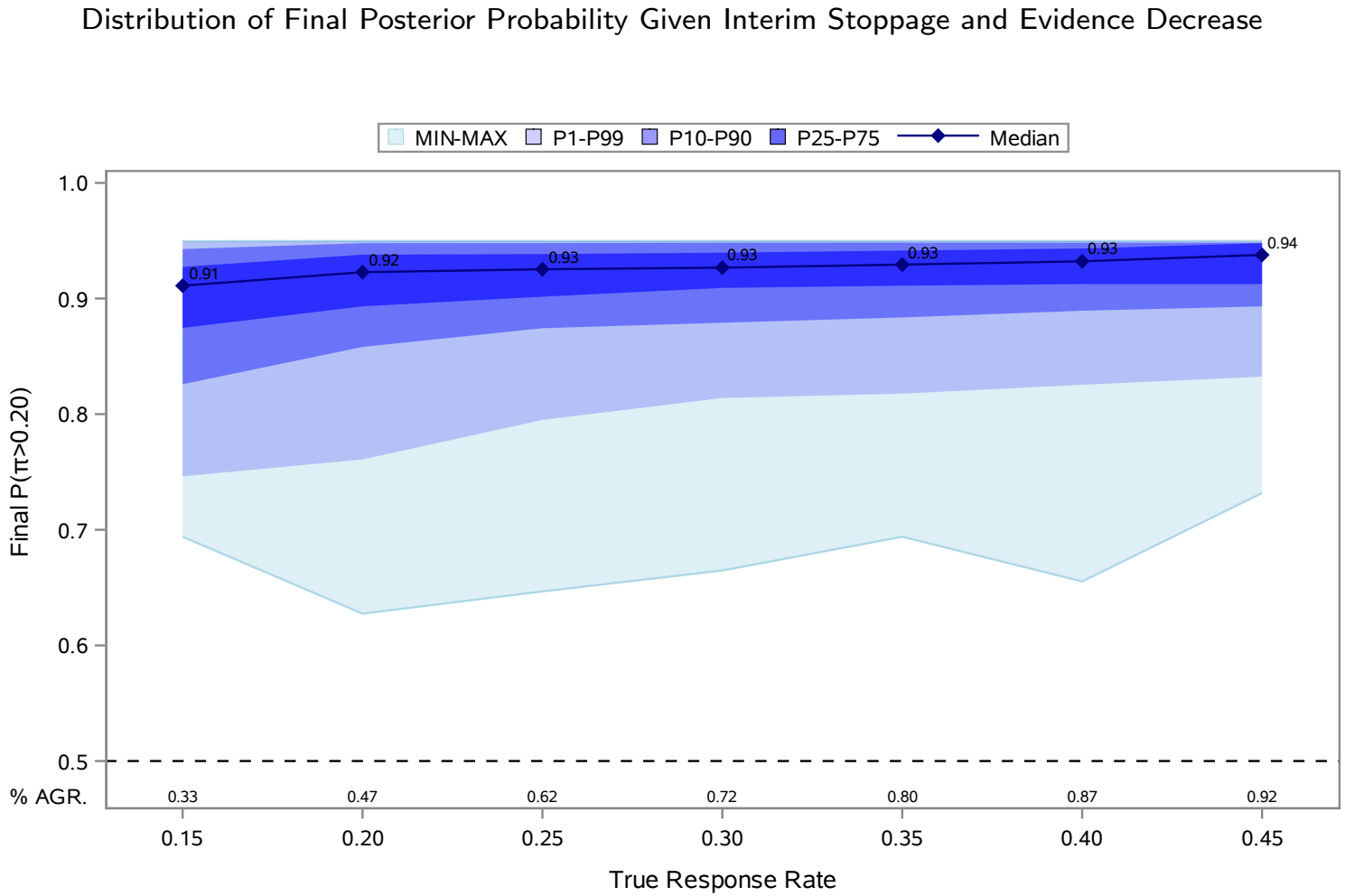
Sequential PoS Design Properties – Slow Accrual



Notes:

- SS = Avg. # Outcomes Ascertained at Interim + Avg. # Subjects Ongoing at Interim = Avg. Final Enrollment
- PM = Posterior Mean; CP = Coverage Probability; (I) = Interim Analysis; (F) Final Analysis;

Sequential PoS Design Properties – Slow Accrual



Note: % AGR. = Percent of agreement between final and interim posterior probabilities relative to 0.95 threshold.

Design Properties – Analysis Every 2 Versus 8 Outcomes