Efficient use of futility and efficacy interim analyses in group-sequential designs

Kaspar Rufibach & Marcel Wolbers
Methods, Collaboration & Outreach Group, PD Data Sciences, Roche Basel
CEN Basel, 3rd September 2023

Generated 2023-09-03 at 07:48:56.

Further resources

- Accompanying markdown file.
- **rpact** vignettes.
- Wassmer and Brannath (2016).

Agenda

1. Example trial
2. What and how much do we gain with interim analyses?
3. Optimal use and timing of interim analyses: efficacy
   - Bias and HA view on it
   - Recommendations for efficacy interims
4. Optimal use and timing of interim analyses: futility
   - Binding vs. non-binding
   - Power loss
   - Recommendations for futility interims

BACKUP

5. Operational considerations
6. Portfolio view
7. Efficacy interims
   - MDD
8. Futility interims
   - How to set futility bound?
   - False-decision probabilities
   - β-spending
   - Other criteria
   - Futility interims: Case study: MIRROS
9. Regulatory guidance on adaptive designs
   - General concerns with confirmatory adaptive designs
   - FDA regulatory guidance on adaptive designs
   - EMA regulatory guidance on adaptive designs
   - Questions that regulators want answers to

Design specifications:

- 2-sided significance level: $\alpha = 0.05$.
- Power: $\alpha = 80\%$.
- Hazard ratio to detect: **0.75**.

Timing specifications:

- $n = 1200$.
- Medians in months: 72 and 96.
- Accrual: ramp-up first six months, then 42/month.

Single-stage design (no interim):

- **380** events needed in any case.
- Time to cutoff (months): 60 under H_0, 66 under H_1.

How much do we gain with interim analyses in group-sequential trials?

Add interim analyses:

- **Futility** interim after 30% of events: stop if hazard ratio > 1.
- **Efficacy** interim after 66.7% of events. O'Brien-Fleming α-spending.

Increases **maximal** number of events (all designs 80% power):

- Fixed design: **380**.
- (Non-binding) futility + efficacy: **408** events, + **7.4**%.
- Efficacy only: **385** events, + **1.3**%.

Probability to stop after respective stage:

Analysis	# events	No effect, i.e. under H_0	Effect size to have 80% power
futility interim	123	0.500	0.060
efficacy interim	272	0.006	0.440
final	408	(1 - 0.500 - 0.006) = 0.494	(1 - 0.060 - 0.440) = 0.500

Expected number of events:

- Under H_0: $0.500 \cdot 123 + 0.006 \cdot 272 + 0.494 \cdot 408 = \mathbf{264}$.
- Under H_1: $0.060 \cdot 123 + 0.440 \cdot 272 + 0.500 \cdot 408 = \mathbf{331}$.

Conclusions: compared to single-stage design,

- if H_1 is true, group-sequential needs on **average** 380 - 331 = 49 = 12.9% less events to show same effect.
- if H_0 is true, group-sequential needs on **average** 380 - 264 = 116 = 30.4% less events to show that drug is useless.

Time to cutoff in months:

Single-stage: 60 under H_0, 66 under H_1.

Analysis	# events	No effect, i.e. under H_0	Effect size to have 80% power
futility interim	123	29	31
efficacy interim	272	46	50
final	408	64	71

Expected duration:

- Under H_0: $0.500 \cdot 29 + 0.006 \cdot 46 + 0.494 \cdot 64 = \mathbf{46}$.
- Under H_1: $0.060 \cdot 31 + 0.440 \cdot 50 + 0.500 \cdot 71 = \mathbf{59}$.

Bias and HA view on it

Efficacy interim: bias

Finally, conventional fixed sample estimates of the treatment effect such as the sample mean tend to be biased toward greater effects than the true value when a group sequential design is used. Similarly, confidence intervals do not have the desired nominal coverage probabilities. Therefore, a variety of methods exist to compute estimates and confidence intervals that appropriately adjust for the group sequential stopping rules (Jennison and Turnbull 1999). To ensure the scientific and statistical credibility of trial results and facilitate important benefit-risk considerations, an approach for calculating estimates and confidence intervals that appropriately accounts for the group sequential design should be prospectively planned and used for reporting results.

FDA guidance on "Adaptive Designs for Clinical Trials of Drugs and Biologics"

U.S. Food and Drug Administration (2019).

How large is bias in practice?

Based on **simulation studies:**

> *For trials with a well-designed interim-monitoring plan, stopping after 50% or more events had been collected has a* **negligible impact** *on estimation.*

Freidlin and Korn (2009)

> *Group sequential designs with stopping rules seek to minimize exposure of patients to a disfavored therapy and speed dissemination of results, and* **such designs do not lead to materially biased estimates**. *. . . Superiority demonstrated in a randomized clinical trial stopping early and designed with appropriate statistical stopping rules is* **likely a valid inference**, *even if the estimate may be slightly inflated.*

Wang et al. (2016)

(At least) two types of bias. **rpact** can easily compute unbiased estimators.

Recommendations for efficacy interims

Efficacy interim - recommendations

- **Not too many** interims for efficacy.
- Not earlier than **50%** of information.
- Always discuss **MDDs** (see backup).
- Prepare for discussion of **bias**.
- Adding further efficacy interims: Easily feasible using α-spending. Neither timing nor decision to add one allowed to rely on earlier **unblinded** looks into data!

	quantity	info = 0.67	info = 0.85	final
Design 1	MDD	0.731		0.816
	local significance level	0.0121		0.0463
Design 2	MDD	0.733	0.784	0.813
	local significance level	0.0121	0.0265	0.0404

rpact can do all that.

Futility interim

Stop trial early $\Rightarrow$ conclude drug does not work.

We look into data multiple times. Still, no adjustment of overall significance level α^* needed. **Why?**

No free lunch: occassionally, trial for working drug stopped for futility $\Rightarrow$ adding futility analysis **reduces study power**.

Choice of futility boundary

Various criteria:

- Primary endpoint estimate in "wrong direction".
- No signal in "early" secondary endpoints (response, PFS, etc.).
- Low conditional power.
- Trade-off in false-decision probabilities.
- Change in Bayesian predictive power ("PTS").
- β-spending.
- Etc.

Binding vs. non-binding

Binding futility interim

Adding futility interim reduces power, i.e.

$$\text{P(reject } H_0 \mid H_1 \text{ is true)}$$

but also

$$\text{P(reject } H_0 \mid H_0 \text{ is true)}$$

$\Rightarrow$ **overprotects** type I error.

Increase critical value(s) to "fully exploit" α again $\Rightarrow$ **reduce sample size**.

Type I error only protected if futility boundary is adhered to.

Not recommended:

- Power gain small.
- iDMC "forced" to stop trial.
- Discouraged by Health Authorities.

Non-binding futility interim

Non-binding:

- No adjustment of critical value(s).
- Type I error protected even if futility boundary is ignored.

Wrap-up maximal number of events (futility boundary HR = 1):

- Fixed design: **380**.
- Efficacy only: **385**.
- Binding futility + efficacy: **401**.
- Non-binding futility + efficacy: **408**.

Power loss

Quantify power loss when adding interim

Once interim boundary chosen:

- **Quantify** power loss.
- Account for it by increasing sample size?

	boundary	power
Design 1 (informal)	1.00	0.78
Design 2 (conditional power)	1.28	0.80
Design 3 (stopping probabilities)	0.90	0.72
Design 4 (beta-spending)		0.80

For simplicity, second interim not accounted for.

Analytical bound: Proschan et al. (2006), Result 3.1:

$$\text{Power}_{\text{new}} \quad \geq \quad 1 - \frac{\beta}{1 - CP(\theta_1)} \quad = \quad 0.75.$$

Futility interim - choice of boundary

Tradeoff between:

1. **Early phase** or **pivotal** trial?
2. Mitigate aggressive development.
3. Timing.
4. Clinically meaningful bound.
5. Kill a drug early that works.
6. Power loss.

...finding right tradeoff can be difficult.

Anderson (2014):

> *Sensible futility boundaries correspond to observed effects* **much weaker** *than those that would achieve success in a trial's final results; otherwise, they could stop a disproportionate number of studies that might eventually succeed. It is important that* **this aspect is understood by trial personnel** *so that expectations are accurate and realistic.*

Gallo et al. (2014).

Recommendations for futility interims

Recommendations

Timing:

- Early $\Rightarrow$ high variability.
- How are costs (fixed vs. variable) distributed over trial?
  - **Stopping late might not save much**.
  - Recruitment ends after 31.6 months $\Rightarrow$ 152 events $\Rightarrow$ information fraction $= 152 / 408 = 37\%$.
- Anderson (2014):

  ...at **25-50%** *[of information] seems potentially useful.*

- At readout of randomized Phase 2 $\sim$ MIRROS (backup).

Quantify and/or compensate **power loss**.

Aggressive boundary $\Rightarrow$ **early peek** at efficacy!

Strategic use of futility interim: Inform other trials + programs!

Futility interim - literature

General discussion of interims: Anderson (2014).

FDA guidance on adaptive designs: U.S. Food and Drug Administration (2019).

Background and criteria for futility interims: Gallo et al. (2014).

Statistical monitoring of clinical trials (book): Proschan et al. (2006).

All computations done in **rpact** or simple manual coding.

What does *stopping a trial for efficacy* **mean?**

Stopping for efficacy - not an automatic decision!

Decision to prematurely stop trial $\Rightarrow$ **not based on statistical criteria alone**:

- **Robust** and clinically convincing. Sensitivity analyses.
- Data should be sufficiently **mature**, i.e. have enough follow up: new drug might be more effective early, but not in the long run (or vice versa).
- All patients should have received treatment: if not $\Rightarrow$ ethical imperative to allow for cross-over of control patients $\Rightarrow$ makes estimation of long-term effect estimates, e.g. overall survival, difficult.

Studies stopped too early for success might not have accumulated sufficient safety information, **regulators are more concerned with safety than efficacy**.

Van Norman (2019).

What does *stopping a trial for efficacy* mean?

Statistically:

- **Reject null hypothesis** of "no effect of drug" in hypothesis test.
- (Typically) Unblind trial and **file**.

Operationally:

- Trial continues as before: patients finish treatment, remain on assessment schedule.
- Data collection might be reduced: IRC-PFS only necessary for approval - that's done!
- Other **efficacy and safety** data remains important: survival follow-up, long-term follow-up of primary endpoint and safety. We will keep taking **follow-up snapshots**!

What does *stopping a trial for futility* **mean?**

What does *stopping a trial for futility* mean?

Low probability you reject null hypothesis at final analysis $\Rightarrow$ stop trial now.

- **Save resources**. Maybe not for this trial (often lots of $$$ already spent), but may reallocate resources.
- **Prevent further exposure** of patients to new therapy.
- Inform other programs.

If we do not stop at futility interim? **Trial can still be a failure**! Probability of success goes up!

Group-sequential designs in drug development

Group-sequential designs with **efficacy** interims generally well-accepted by Health Authorities:

- Plain vanilla Phase 3 design, especially in oncology.
- Strong control of type I error generally **non-negotiable** for confirmatory studies $\Rightarrow$ group-sequential designs have this property.
- **Pre-specification** is key.
- **Timing** of efficacy interim needs to be carefully considered and **pre-defined**.
  - Decision to stop trial pre-maturely not to be driven by **early** effect only.
  - Ideally, all patients should have finished treatment.
  - Time-to-event endpoint: ratio of #events / #patients should not be too small.
- Inference after stopping trial early **in principle** not straightforward.

Futility interims less controversial $\Rightarrow$ risk is with the company.

Thank you for your attention.

kaspar.rufibach@roche.com
http://go.roche.com/dss-mco

http://www.kasparrufibach.ch

References I

- Anderson, K. M. (2014). Timing and frequency of interim analyses in confirmatory trials. In *Practical Considerations for Adaptive Trial Design and Implementation*. Springer, 115–123.
- Bauer, P. and Koenig, F. (2006). The reassessment of trial perspectives from interim data–a critical view. *Stat. Med.* **25** 23–36.
- Committee for proprietary medicinal products (2007). Reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan. Tech. rep.
- Freidlin, B. and Korn, E. L. (2009). Stopping clinical trials early for benefit: impact on estimation. *Clin Trials* **6** 119–125.
- Gallo, P., Mao, L. and Shih, V. H. (2014). Alternative views on setting clinical trial futility criteria. *Journal of Biopharmaceutical Statistics* **24** 976–993. PMID: 24933121. `https://doi.org/10.1080/10543406.2014.932285`
- Lachin, J. M. (2005). A review of methods for futility stopping based on conditional power. *Statistics in medicine* **24** 2747–2764.
- Meller, M., Beyersmann, J. and Rufibach, K. (2019). Joint modeling of progression-free and overall survival and computation of correlation measures. *Statistics in medicine* **38** 4270–4289.
- Proschan, M., Lan, K. and Wittes, J. (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*. Springer, New York.

References II

- Rufibach, K., Heinzmann, D. and Monnet, A. (2020). Integrating phase 2 into phase 3 based on an intermediate endpoint while accounting for a cure proportion – with an application to the design of a clinical trial in acute myeloid leukemia. *Pharmaceutical Statistics* **19** 44–58. Code available on github: `https://github.com/numbersman77/integratePhase2.git`.
- Rufibach, K., Jordan, P. and Abt, M. (2016). Sequentially updating the likelihood of success of a Phase 3 pivotal time-to-event trial based on interim analyses or external information. *J Biopharm Stat* **26** 191–201.
- Rufibach, K., Jordan, P. and Abt, M. (2021). *bpp: Computations Around Bayesian Predictive Power*. R package version 1.0.2. `https://CRAN.R-project.org/package=bpp`
- U.S. Food and Drug Administration (2019). *Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics*. `https://www.fda.gov/media/78495/download`
- Van Norman, G. A. (2019). Phase ii trials in drug development and adaptive trial design. *JACC: Basic to Translational Science* **4** 428–437. `https://www.sciencedirect.com/science/article/pii/S2452302X19300658`
- Wang, H., Rosner, G. L. and Goodman, S. N. (2016). Quantifying over-estimation in early stopped clinical trials and the "freezing effect" on subsequent research. *Clin Trials* **13** 621–631.

References III

- Wassmer, G. and Brannath, W. (2016). Adaptive group sequential tests. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials* . http://dx.doi.org/10.1007/978-3-319-32562-0%5F6

Portfolio:

- 50 trials.
- $P(H_0 = \text{true}) = 0.35$.
- $P(H_1 = \text{true}) = 0.65$.

Single-stage designs: $50 \cdot 380 = \mathbf{19000}$ events.
Group-sequential designs: $0.35 \cdot 50 \cdot 264 + 0.65 \cdot 50 \cdot 331 = \mathbf{15385}$ events.

Efficacy interim

Anderson (2014).

FDA guidance on adaptive designs: U.S. Food and Drug Administration (2019).

> *Early stopping for a positive efficacy finding can be a* **controversial** *topic.*
>
> *My recent experience with FDA oncology regulators suggested no interim efficacy analyses until after* **50 % of efficacy data** *have been collected.*
>
> *In addition to FDA suggestions to limit early efficacy analyses, the European Medicines Agency (EMA) has also strongly suggested* **limiting the number of interim efficacy analyses**.
>
> *Anderson (2014)*

MDD

Trial powered for hazard ratio 0.75.

What hazard ratio do we need to see at efficacy interim analysis to be significant?

Trial powered for hazard ratio 0.75.

What hazard ratio do we need to see at final analysis to be significant?

Power assumption vs. MDD at efficacy interim

Minimal detectable difference (MDD):

- Largest observed hazard ratio for which trial will **just be significant**, i.e. give a p-value of α.
- MDD is **analysis-dependent**:
  - Significance level α different at interim and final.
  - MDD depends on standard error $\Rightarrow$ number of events analysis is performed at.
- Efficacy interim: $\alpha = 0.012, d = 272 \Rightarrow$ MDD $=$ **0.738**. "Target TPP".
- Final analysis: $\alpha = 0.046, d = 408 \Rightarrow$ MDD $=$ **0.821**. "Minimal TPP".
- Compare MDDs to **0.75** used for powering:
  - MDDs say something about **null hypothesis**.
  - Effect for powering is specification of **alternative** hypothesis.

Choice of scale

Scale:

- z-statistic.
- Effect scale $\Rightarrow$ hazard ratio.
- β-spending $\Rightarrow$ local type II error.
- Conditional power: tricky in **rpact**, better interpretability.
- Bayesian predictive power: own implementation, better interpretability.

Translation:

$$z \quad = \quad \log(\theta)\sqrt{\kappa(1-\kappa)d},$$

$\kappa = P(\text{randomized to arm } A)$.

`go.roche.com/adaptr`, **Q&A 3.2**.

How to set futility bound?

How to set futility bound?

Power: Given assumed effect what is P(success)?

$$\pi(\theta) \quad = \quad P_\theta(\text{reject } H_0 \text{ at final}).$$

Conditional power: Given interim data and assumed effect after interim what is P(success) if we continue?

$$CP(\theta) \quad = \quad P_\theta(\text{reject } H_0 \text{ at final} \mid \widehat{\theta}_{\text{int}}).$$

Random variable! Bauer and Koenig (2006). See also Lachin (2005).

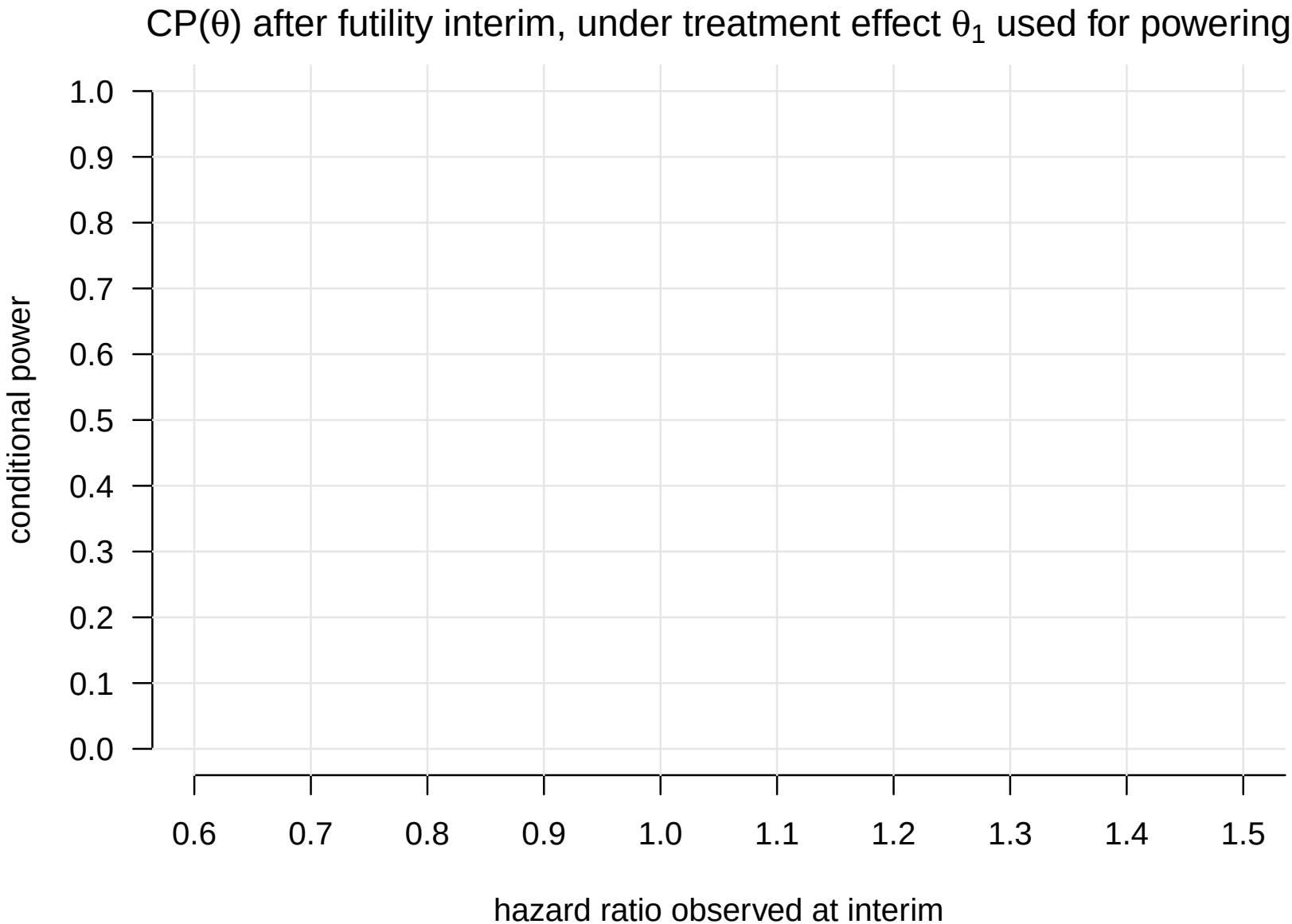
Depends on:

- $\widehat{\theta}_{\text{int}}$: effect estimate **up to interim**.
- θ: effect **beyond interim**.

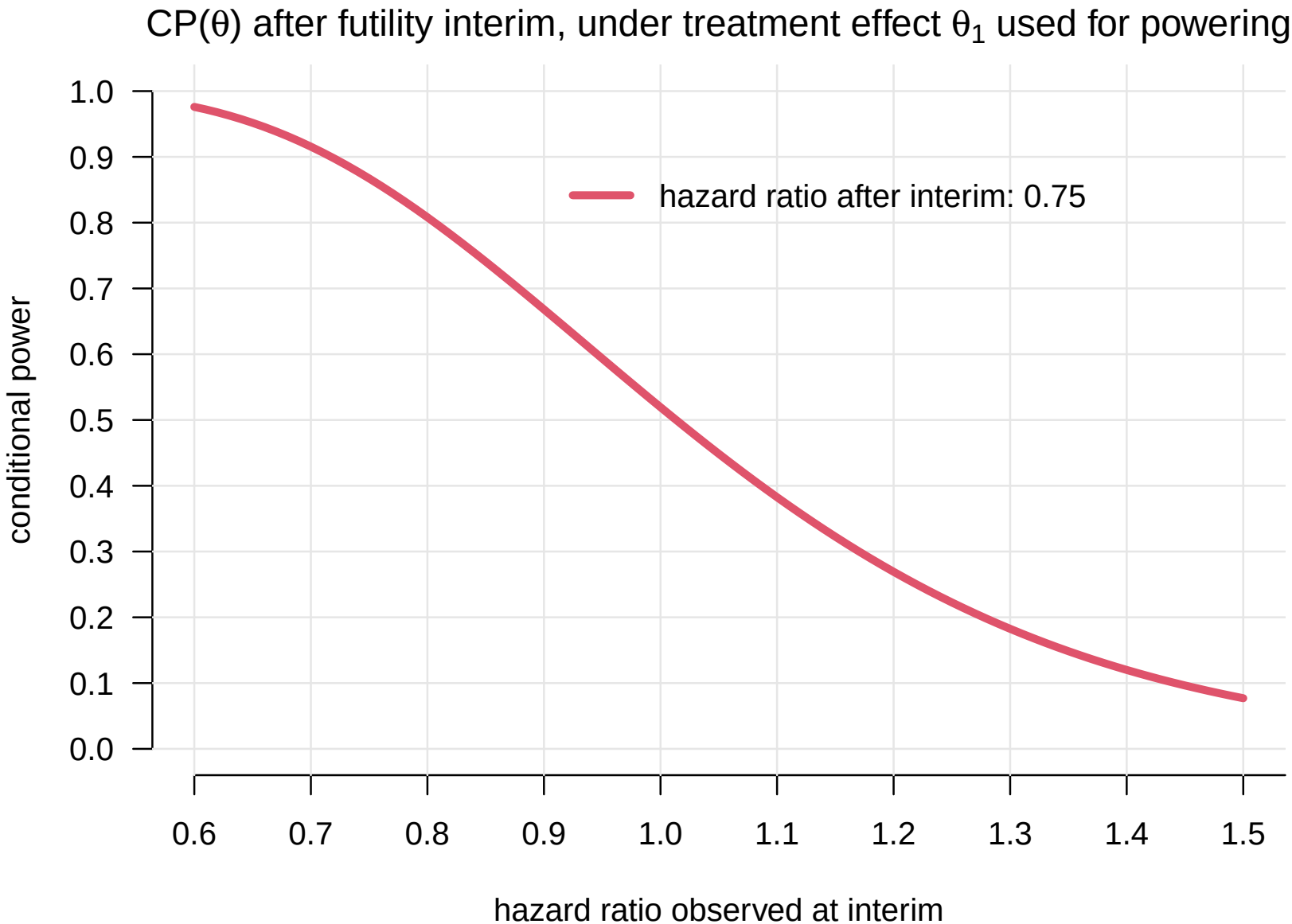
Recamp example trial

Analysis	# events
futility interim	123
efficacy interim	272
final	408

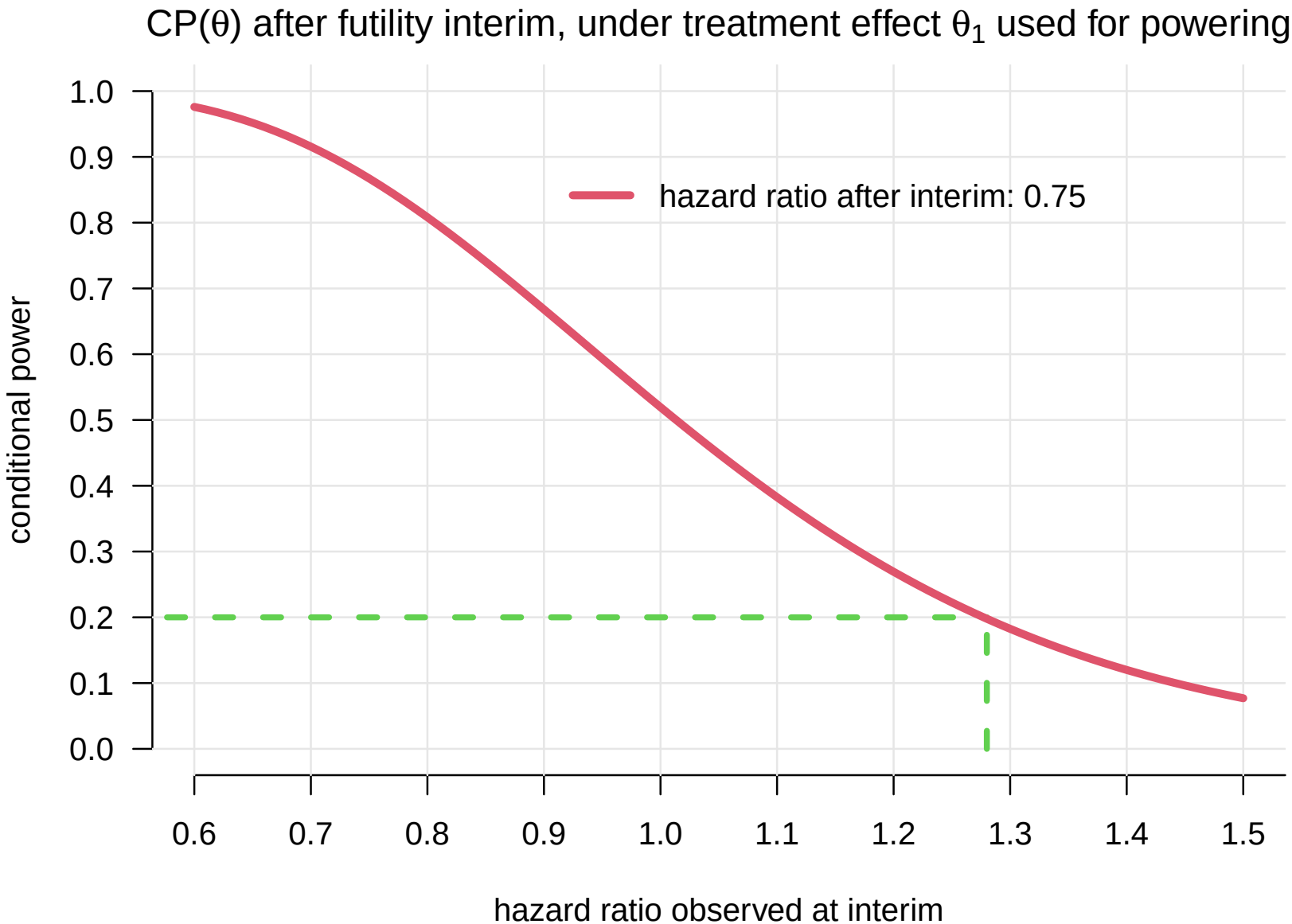
Conditional power



Conditional power

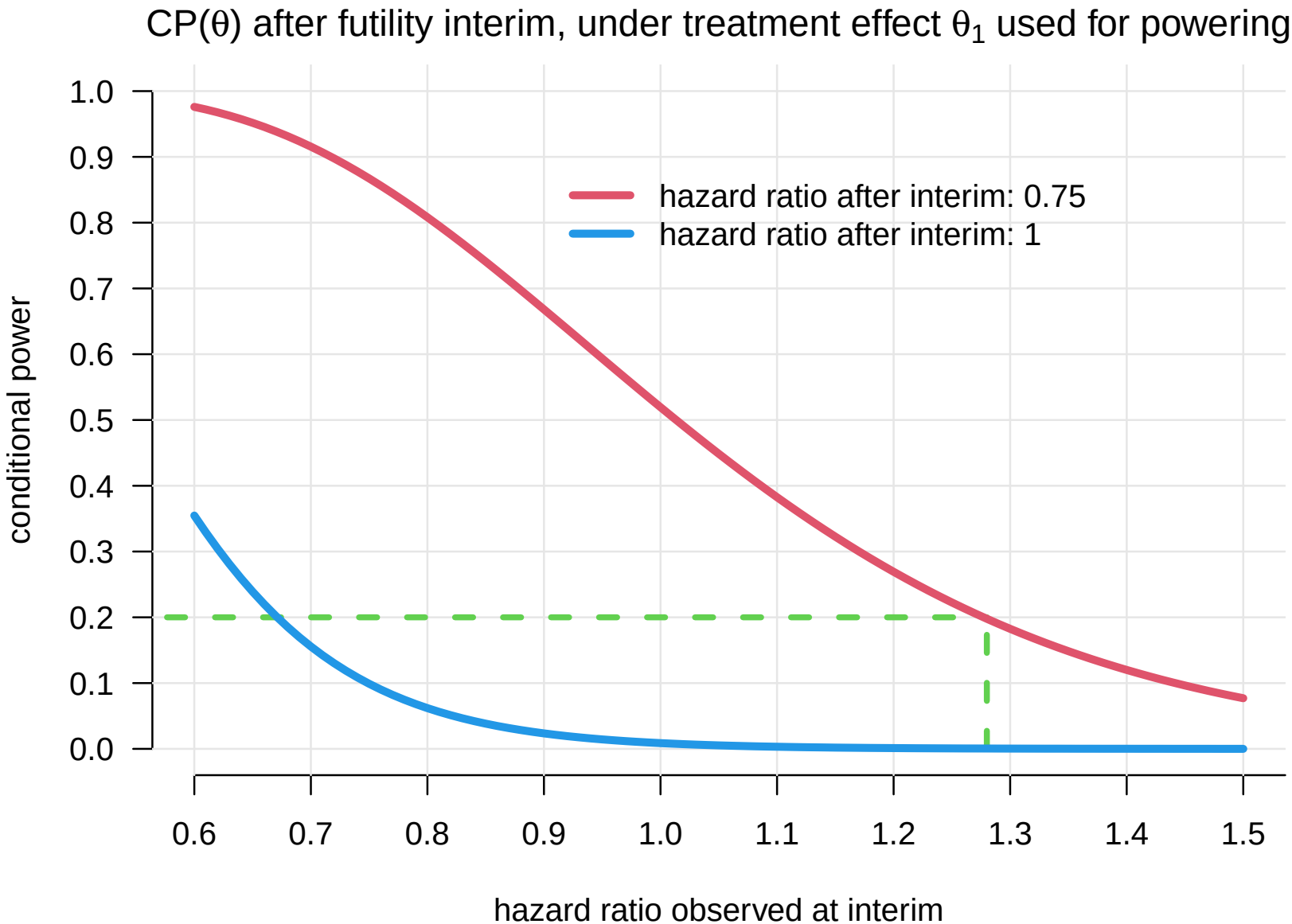


Conditional power



If futility boundary $= 1.28 \Rightarrow CP(\theta_1) = P_{\theta_1}(\text{reject } H_0 \text{ at final} \mid \widehat{\theta}_{\text{int}} = 1.28) = 0.2.$

Conditional power



If futility boundary $= 1.28 \Rightarrow CP(\theta_1) = P_{\theta_1}(\text{reject } H_0 \text{ at final} \mid \widehat{\theta}_{\text{int}} = 1.28) = 0.2.$

Conditional power

$P_\theta(\text{reject } H_0 \text{ at final} \mid \widehat{\theta}_{\text{int}} = 1.28) = \mathbf{0.2}$.

Equivalent to p-**value** ≥ 0.91. Monotonocity of $CP(\theta)$.

Conclusions for conditional power:

- Interim early $\Rightarrow$ low interim hurdle based on CP.
- What to use for θ? Matter of debate!
- Bauer and Koenig (2006):

  Using the estimated effect size for sample size reassessment seems not be a recommendable option." Too much variability!

  Trying to use the **original effect size from the planning phase** *should always be considered as a useful option.*

- Recommendation: $\theta = \theta_1$ **used for powering**.

False-decision probabilities

False-decision probabilities

Conditional power:

$$P_\theta(\text{reject } H_0 \text{ at final} \mid \widehat{\theta}_{\text{int}}).$$

LIP based on randomized Phase 2: interested in

$$\begin{aligned} \text{False-positive probability:} \quad & P_\theta(\widehat{\theta}_{\text{P2}} \leq \theta_{\text{P2}} \mid \mathbf{H_0}), \\ \text{False-negative probability:} \quad & P_\theta(\widehat{\theta}_{\text{P2}} \leq \theta_{\text{P2}} \mid \mathbf{H_1}). \end{aligned}$$

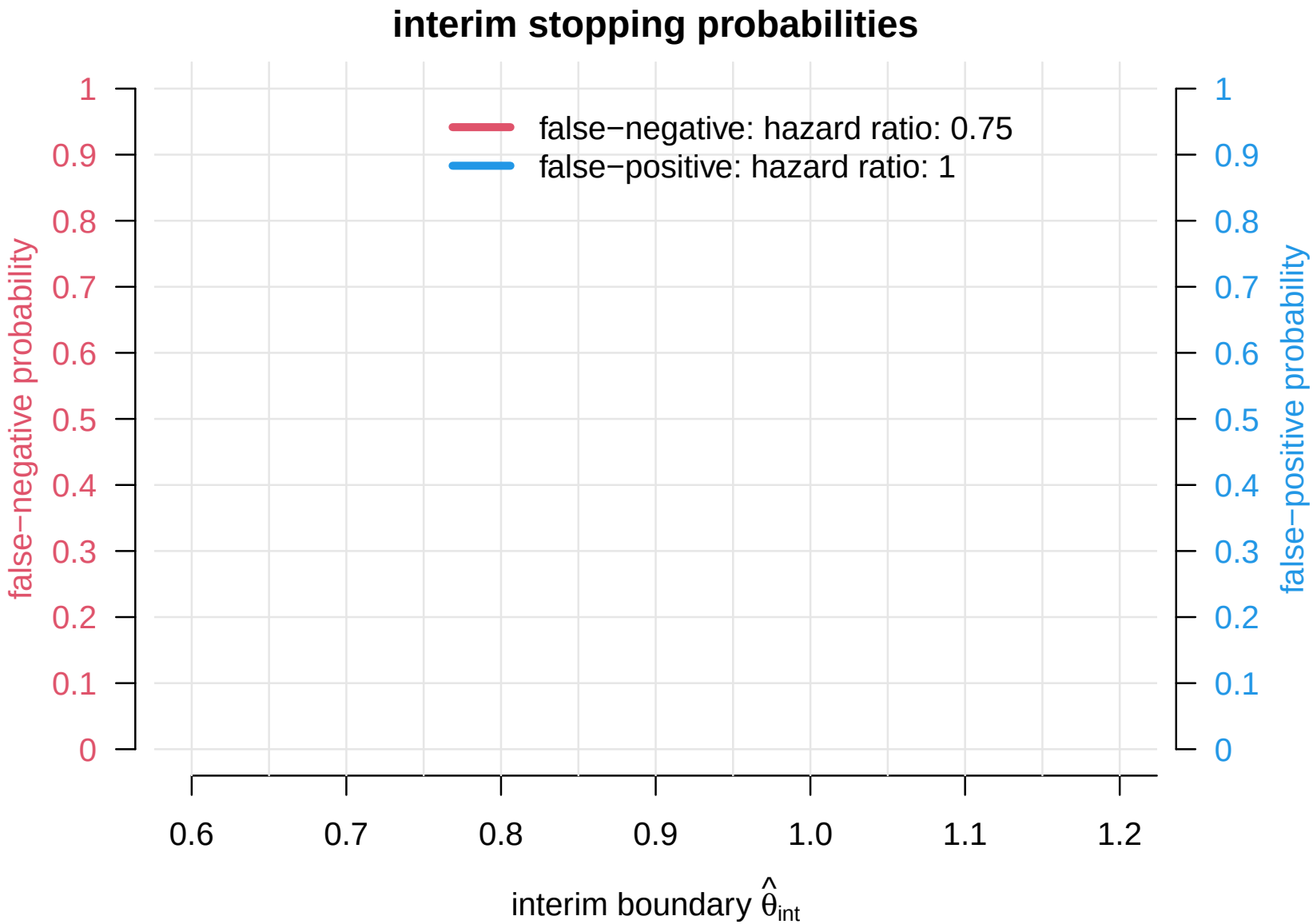
LIP built-in as futility interim in **pivotal Phase 3**: as function of interim boundary θ_{int}:

$$\begin{aligned} \text{False-positive probability:} \quad & P_\theta(\text{continue at interim} \mid \mathbf{H_0}) = P_\theta(\widehat{\theta}_{\text{int}} \leq \theta_{\text{int}} \mid \mathbf{H_0}), \\ \text{False-negative probability:} \quad & P_\theta(\text{stop at interim} \mid \mathbf{H_1}) = P_\theta(\widehat{\theta}_{\text{int}} > \theta_{\text{int}} \mid \mathbf{H_1}). \end{aligned}$$

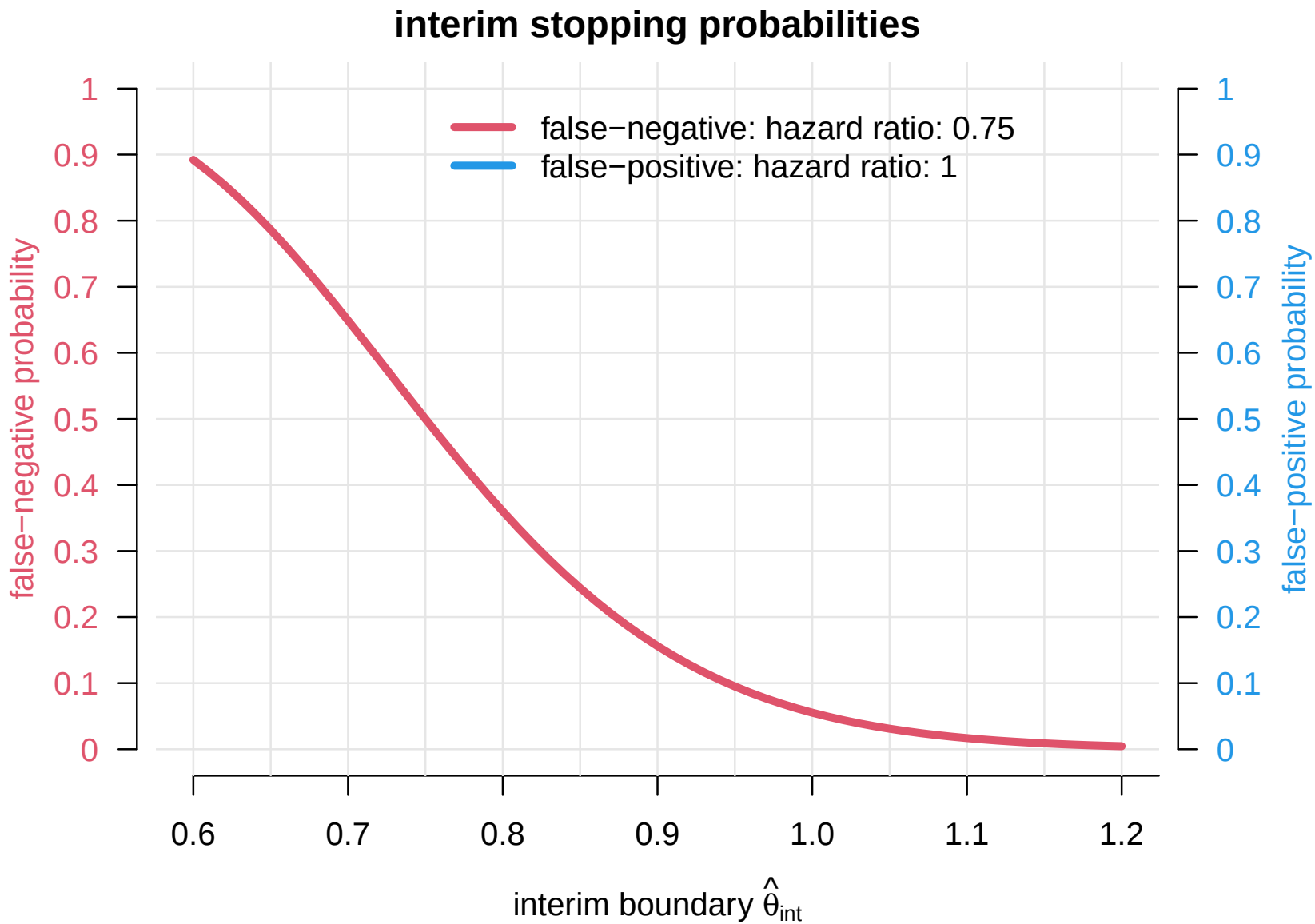
Find **sweet spot** trading these two off.

Very different from conditional power!

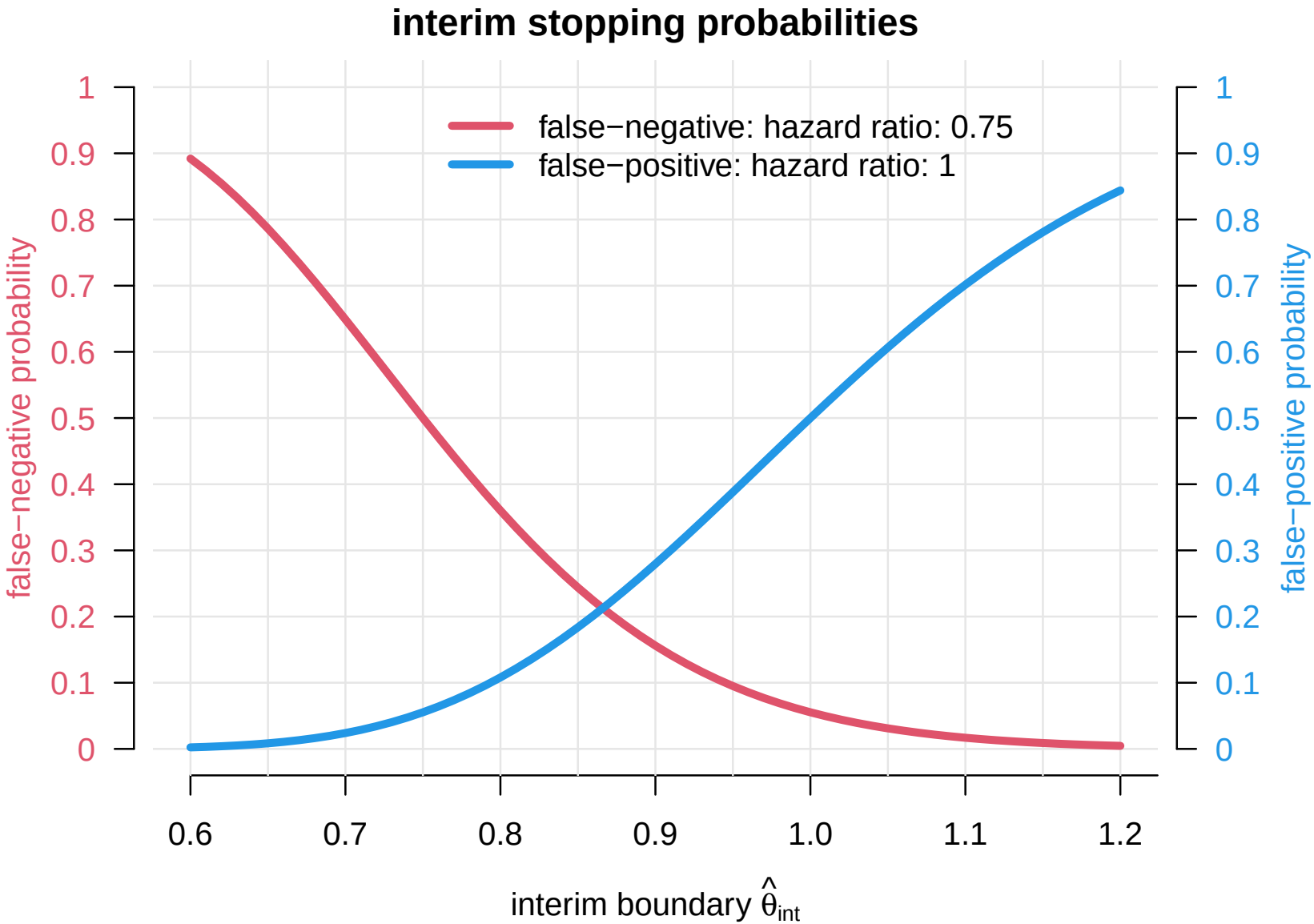
Stopping probabilities



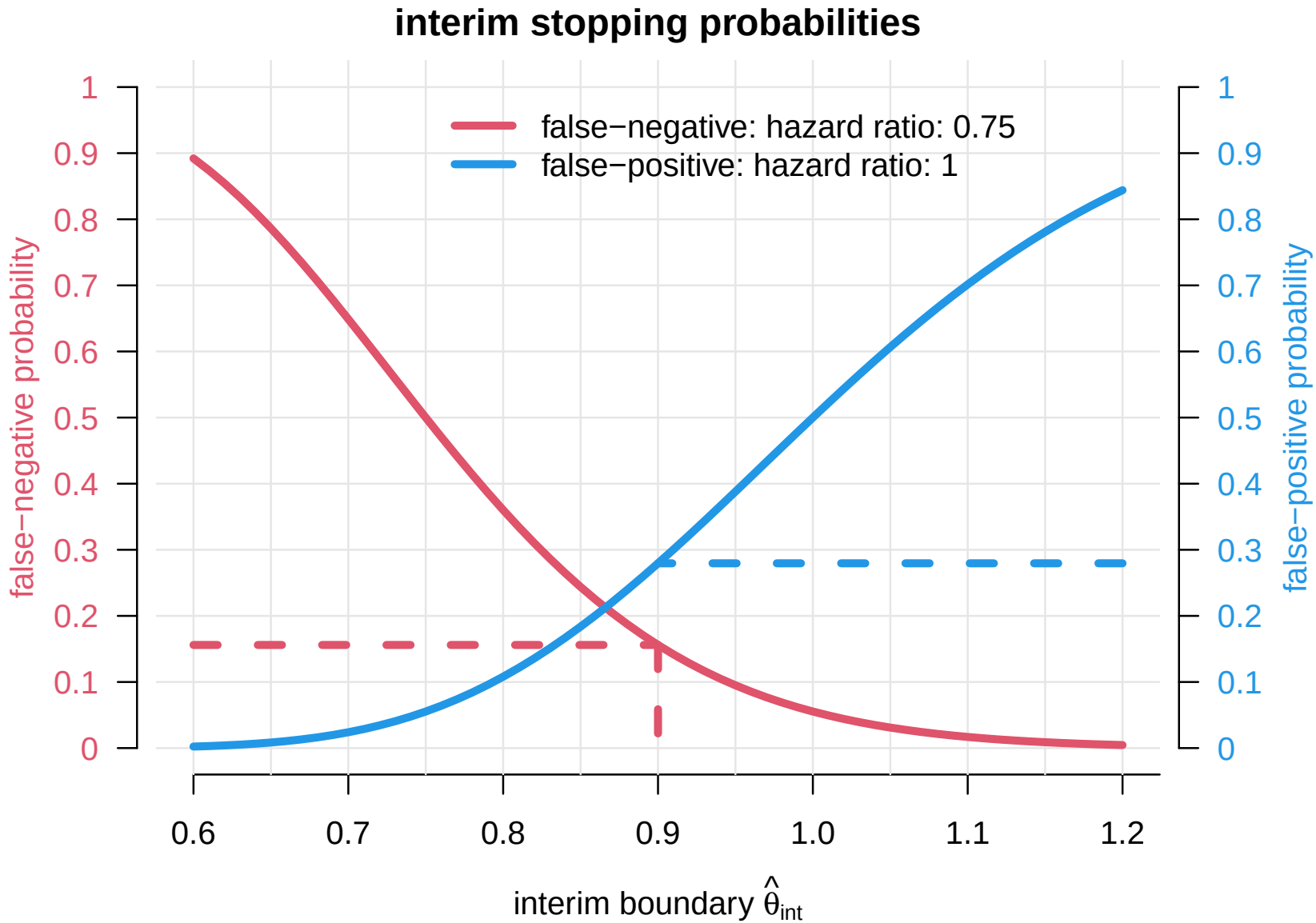
Stopping probabilities



Stopping probabilities



Stopping probabilities



$P(\text{continue at interim} \mid H_0) = 0.28$

$P(\textbf{stop at interim} \mid H_1) = \textbf{0.16.}$

β-spending

β-spending

Same design, with and without β-**spending**:

quantity	no futility interim	beta-spending
number of events	385	419
efficacy boundary 1 (effect size)	0.48	0.50
efficacy boundary 1 (p-value)	0.00004	0.00004
efficacy boundary 2 (effect size)	0.73	0.74
efficacy boundary 2 (p-value)	0.006	0.006
efficacy boundary 3 (effect size)	0.82	0.82
efficacy boundary 3 (p-value)	0.02	0.02
futility boundary 1 (effect size)		1.09
futility boundary 1 (p-value)		0.68
futility boundary 2 (effect size)		0.87
futility boundary 2 (p-value)		0.12

- Assumption: futility **adhered** to $\Rightarrow$ power loss compensated for.
- Increase number of events: from **385** to **419**.
- Power of β-spending design with 385 events: **0.77**.
- **Rarely** used.

Other criteria

Other criteria

Change in Bayesian predictive power after interim: MIRROS.

So far, this was easy.

Why?

Interim = primary endpoint.

Futility interims: Case study: MIRROS

Primary endpoint: OS.

Interim endpoint: response.

Stopping probabilities, conditional on H_0, H_1?

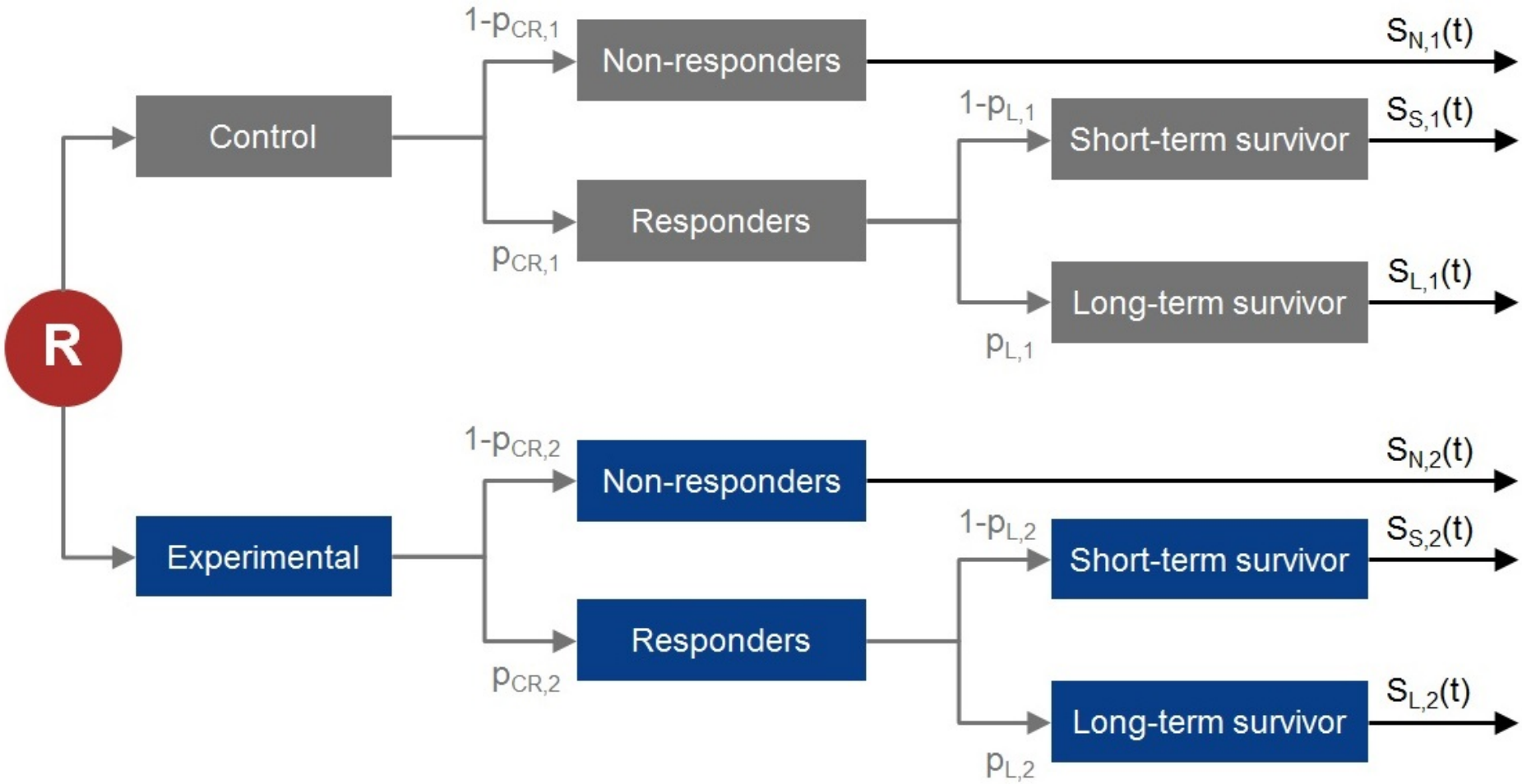
How $S_{OS}(t)$ generated involving intermediate endpoint? Allows for

- **conditioning on** H_0, H_1,
- quantification of **power loss**.

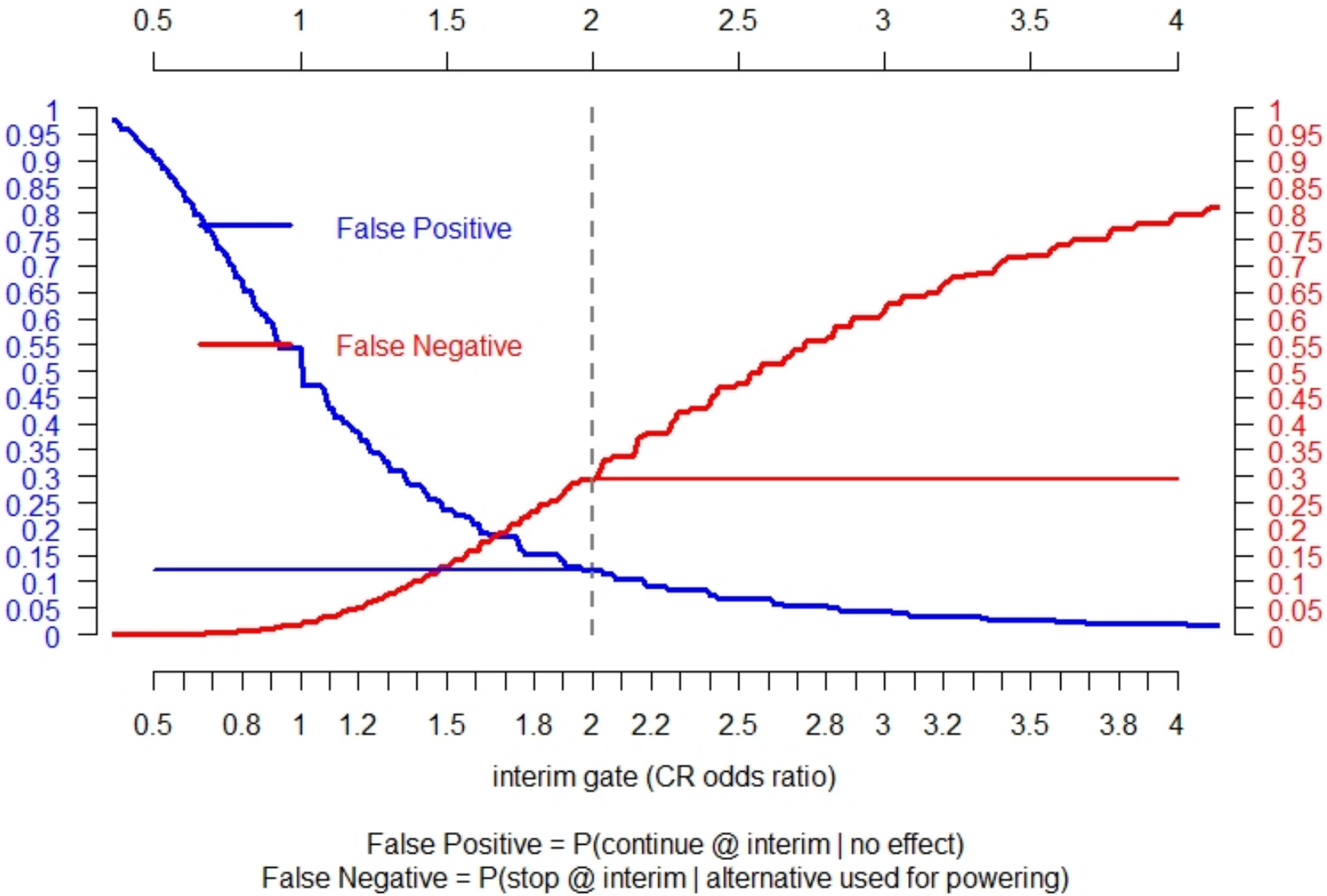
Options:

- Construct $S_{OS}(t)$ from S's in **subgrous** (responders vs. non-responders) $\Rightarrow$ MIRROS.
- $S_{OS}(t)$ prediction in **multistate model**. Opens door for response or PFS as intermediate endpoint. Model for PFS and OS: Meller et al. (2019).

Mechanistic simulation model



Operating characteristics of various interim boundaries



False Positive = P(continue @ interim | no effect)
False Negative = P(stop @ interim | alternative used for powering)

Operating characteristics of various interim boundaries

Sweet spot: **odds ratio of 2**,

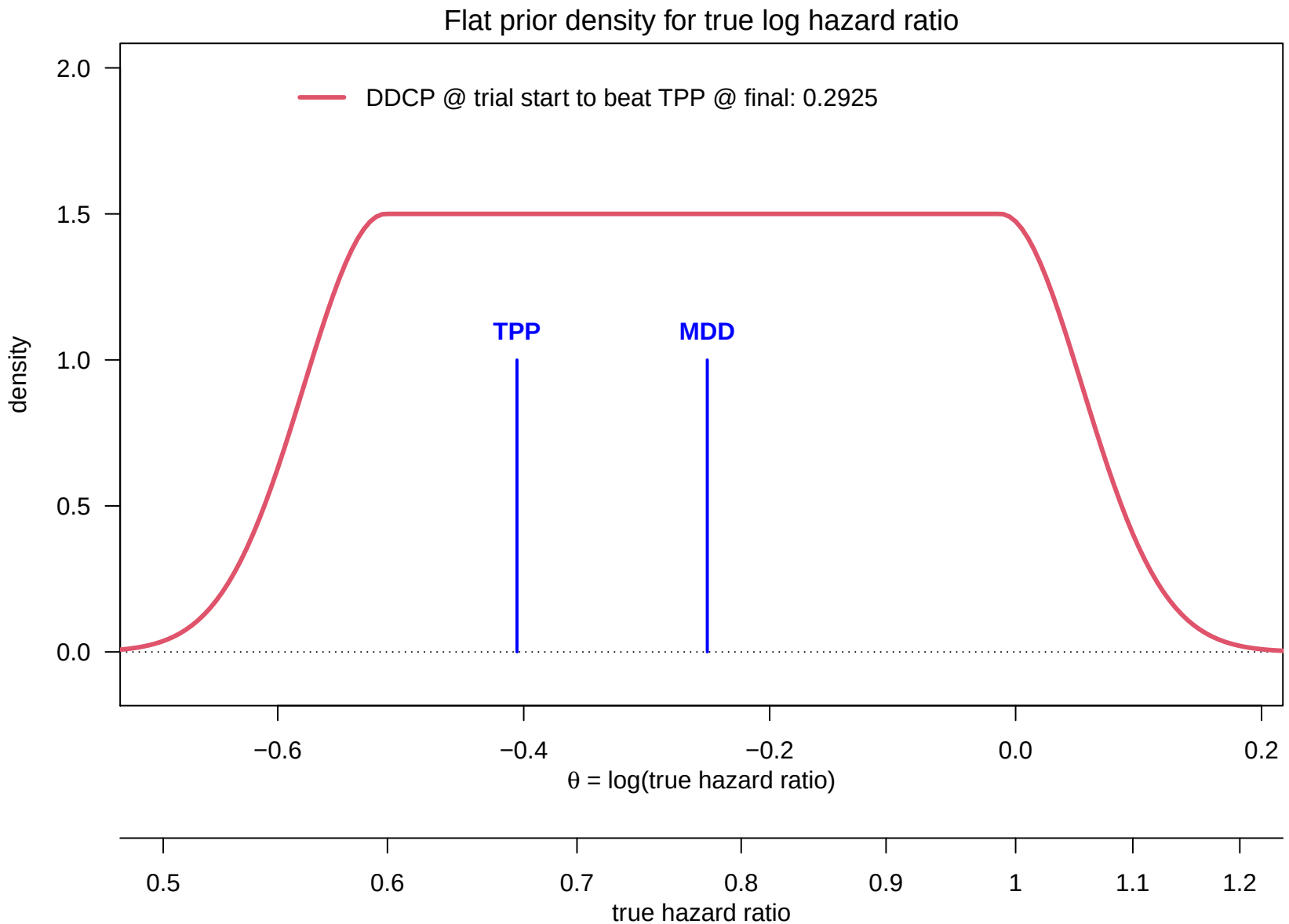
- False Positive = P(continue @ interim | H_0) $\approx$ 12%,
- False Negative = P(stop @ interim | H_1) $\approx$ 30%.

Power loss of adding futility interim

Can easily get that from simulations.

- Targeted power: 85%.
- Power taking into account futility interim: **63%!**
- Power loss not accounted for in total number of events.
- Illustrates risk-appetite $\Rightarrow$ futility interim = "informal efficacy interim".

Pessimistic priors for values of assumed initial DDCP



Challenge

Initial Bayesian predictive power ("PTS"): $0.45 \cdot 0.65 = 0.29$.

How to update assuming interim passed?

1. Simulate 10'000 trials under H_1.
2. Look at distribution of OS HRs for those simulated scenarios that jump the interim hurdle.
3. 80% are ≤ 0.865.
4. Bayesian predictive power assuming OS HR at interim was ≤ 0.865: **0.428**.

Methodology described in Rufibach et al. (2016).

R package on CRAN: **bpp**, Rufibach et al. (2021).

MIRROS

Key conclusions:

- Start Phase 3 after Phase 1 $\Rightarrow$ mitigate risk with (aggressive) futility interim.
- Use **intermediate** endpoint for futility decision. Not "established" surrogate!
- **Feasible with HAs**.

Details: Rufibach et al. (2020).

Code: https://github.com/numbersman77/integratePhase2.

General concerns with confirmatory adaptive designs

Type I error control

Bias in estimation of treatment effects

Trial planning and pre-specification

Trial conduct and integrity

Type I error control

Sources of **multiplicity**: number of

- looks,
- doses / arms,
- populations,
- endpoints,
- sample size re-assessment based on "comparative" results, ...

Or **combinations** thereof!

Statistical theory.

Simulations.

Bias in estimation of treatment effects

Raw end-of-trial treatment effect estimate: typically **biased without taking adaptation into account**. Bias depends on:

- type of adaptation and specific adaptation rule,
- true treatment effect,
- nuisance parameters.

Analytical adjustment if available.

May use **simulations** to quantify bias.

Gallium European filing

Gallium stopped at efficacy interim:

- After 245 of 370 events (248 planned, 370 for final $\Rightarrow$ 66.2% of events).
- 245 / 1202 (20.4%) of patients with event $\Rightarrow$ interim quite **early**.
- "Raw" estimate of hazard ratio: 0.66 with 95% confidence interval from 0.51 to 0.85, p-value 0.0012. **Overestimation**, since we stopped at interim for efficacy.

How large do you think is the bias?

Gallium European filing - answering strategy

Comprehensive simulation study to identify scenarios where **conditional bias** becomes non-negligible: Freidlin and Korn (2009).

Conclusions: Overestimation of hazard ratio becomes appreciable if:

- Trial is stopped **very early** ($\leq$ 40% of targeted events) $\Rightarrow$ Gallium 66.2%.
- True hazard ratio is **close to 1**. Gallium estimate was 0.66.

Gallium:

- Unbiased estimate of hazard ratio: 0.6625, with 95% CI from 0.5157 to 0.8515.
- Adjusted estimate, confidence interval, and (one-sided) p-value **virtually identical** to standard inference.

How large is bias in practice?

Based on **simulation studies:**

> *For trials with a well-designed interim-monitoring plan, stopping after 50% or more events had been collected has a* **negligible impact** *on estimation.*

Freidlin and Korn (2009)

> *Group sequential designs with stopping rules seek to minimize exposure of patients to a disfavored therapy and speed dissemination of results, and* **such designs do not lead to materially biased estimates**. *. . . Superiority demonstrated in a randomized clinical trial stopping early and designed with appropriate statistical stopping rules is* **likely a valid inference**, *even if the estimate may be slightly inflated.*

Wang et al. (2016)

For group-sequential designs. Adaptive designs might have larger bias. Unbiased estimates under assumptions e.g. from simulations.

Trial planning and pre-specification

Details of the adaptive design completely specified **prior to initiation of the trial**:

- Number and timing of interim analyses (some flexibility for group-sequential designs).
- type of adaptation,
- statistical methods: type I error, power,
- decision rules and criteria.

Sponsor-internally: decision makers may not see data for a long-time!

- Dose selection $\Rightarrow$ Gatsby.
- Phase 3 with futility interim started directly after Phase 1 $\Rightarrow$ MIRROS.

Trial conduct and integrity

Knowledge of accumulating data can affect conduct of trial: excitement among investigators after not stopping after a futility interim analysis.

Limit access to interim results on treatment effect to individuals independent of trial conduct (iDMC).

FDA regulatory guidance on adaptive designs

2019 FDA guidance on adaptive designs

U.S. Food and Drug Administration (2019)

Considerations:

- Regulatory process for obtaining formal, substantive feedback well-established.
- Guidance open towards frequentist or Bayesian designs $\Rightarrow$ as long as **operating characteristics** adequately evaluated (e.g. via simulation).
- Approach any agency **early**!
- Submit protocol and SAP plus:
  - Rationale for design.
  - Prespecified monitoring, adaptation, statistical methods.
  - Operating characteristics: type I error, power.
  - Bodies responsible for implementing adaptive design, e.g. iDMC charter.
  - Who accesses which data when? Maintain trial integrity.

EMA regulatory guidance on adaptive designs

2007 EMA guidance on adaptive designs

Committee for proprietary medicinal products (2007)

- "Adaptive designs should not be seen as a means to alleviate the burden of rigorous planning of clinical trials."
- Substantial changes of trial design:
  - Via protocol amendment, e.g. changes in duration of treatment, mandatory co-medications, or criteria for inclusion or exclusion of patients.
  - Re-size trial so that primary analysis can be based on patients randomised after change.
  - Minimal requirement: primary analysis should be stratified by randomised before or after amendment, homogeneity of results should be investigated and discussed.
  - Refers to **non-pre-specified** scenario! These are not popular with regulators at all.
- Emphasis on control of type I error.

ICH E20 guideline "Adaptive Clinical Trials" under development. Link to concept paper

Questions that regulators want answers to

Questions that regulators want answers to

1. Is there **need** for adaptive trial? Is there good **rationale**?
2. Have alternative, more standard trial designs been considered?
3. Is number of interim analysis justified? More than one interim analysis may be justified in long term clinical trials.
4. Potential advantages of adaptive design need to be weighed against **potential biases and additional complexities**.
5. Does proposal fit well in context of development program and data that will be available for the marketing authorization application?
6. Can proposal be implemented without damage to trial integrity?
7. Is **type I error** controlled?
8. Has potential bias of treatment effect estimates been evaluated? What about **endpoints other than primary**, are they interpretable?
9. Is proposal practical and feasible?

Doing now what patients need next

R version and packages used to generate these slides:

R version: R version 4.2.3 (2023-03-15 ucrt)

Base packages: stats / graphics / grDevices / utils / datasets / methods / base

Other packages: rpact / reporttools / xtable / mvtnorm

This document was generated on 2023-09-03 at 07:48:56.