successR: comprehensive resource for training, methodology, and computational tools around success probabilities

successR task force
Roche PD DSS
May / June 2021

go.roche.com/successR

successR task force

Basel	Kaspar Rufibach
	Simona Rossomanno
Welwyn	Bernie Surujbally
Shanghai	Yifan Wang
Beijing	Iris Yan
Mississauga	Kasra Yousefi
SSF	Godwin Yung

go.roche.com/successR/index.html

Acknowledgments

- Everybody who provided input so far!
- Paul Jordan.
- Victor Huang, Nina Qi, Jiawen Zhu.

Ideal state

Ideal state

Every PD DSS colleague ...

- ...has good **understanding of methodology**, purpose, computation, and tools for DDCP and PTS.
- ...uses **standardized** approach to DDCP and predictive probability computations.
- ...has easy access to **software** to perform computations quickly.
- ...has easy access to library of **business-relevant** examples.
- ...has easy access to **templates** to be shared with broader project teams.

Scope of successR

DDCP computations for randomized trials.

Prior choice.

Update after not stopping at interim.

Predictive probability computations for single-arm trials (Unicycle).

Agenda

1. What is PTS?
2. How can we quantify probability of technical success?
3. DDCP typically lower than power
4. Prior sample size and DDCP
5. What triggers a DDCP update?
   - Update DDCP with external information
   - Update DDCP after an interim analysis
6. Choice of prior
7. What about single-arm studies?
8. Resources
9. Take home messages
10. Backup

Agenda

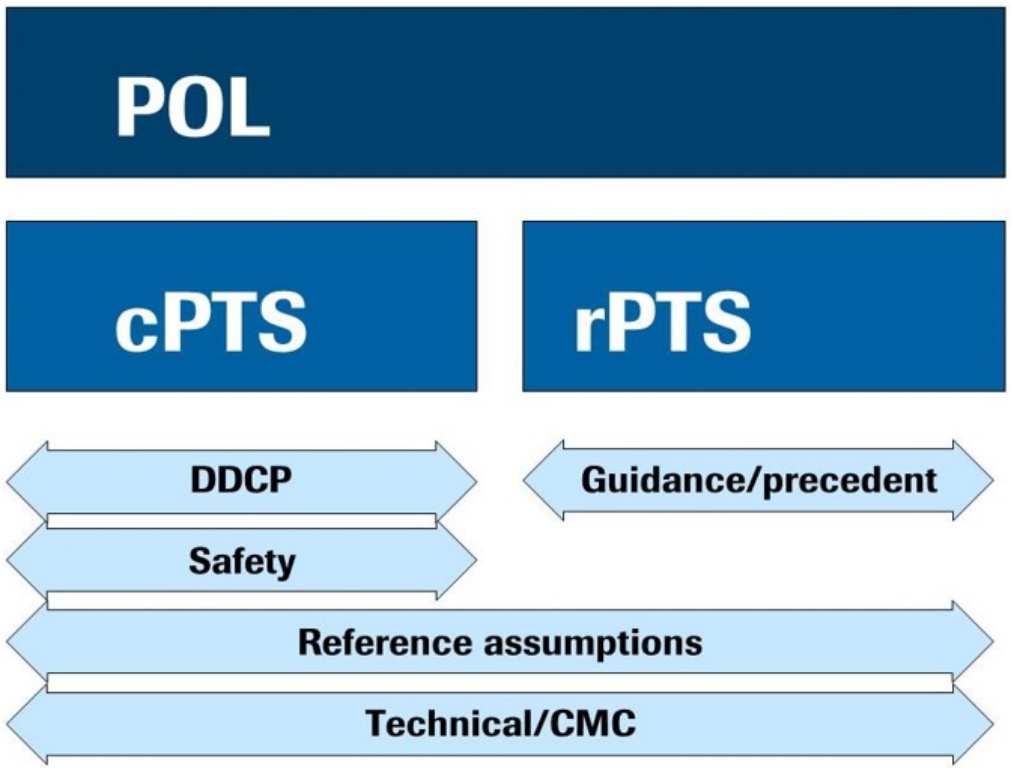
1. What is PTS?
2. How can we quantify probability of technical success?
3. DDCP typically lower than power
4. Prior sample size and DDCP
5. What triggers a DDCP update?
   - Update DDCP with external information
   - Update DDCP after an interim analysis
6. Choice of prior
7. What about single-arm studies?
8. Resources
9. Take home messages
10. Backup

What is PTS?

Probability of technical success

PTS late stage handbook (Slide 41):

- **PTS**: Probability of successfully transitioning to the next phase.
- **cPTS**: Probability of meeting efficacy and safety requirements.
- **Phase 3 cPTS**: probability of achieving target TPP.



cPTS

Comprehensive assessment of probability to have **successful** Phase 3 trial. Based on:

1. **Quantitative** assessment: assurance (data-driven conditional probability, "DDCP", cPTS).
2. **Qualitative** adjustment: non-quantifiable additional information e.g. on competitors, uncertainty around assumptions, change in endpoint from Phase 2 to Phase 3, safety, ...

Success: be significant or beat (minimal) **target product profile** (TPP).

Out of scope: Project Providentia: `https://pd.roche.com/pd/post/pds-participating-in-project-providentia-the-ability-to-foresee/`.

Probability of technical success

Used for:

- Calculating project **valuations**,
- supporting **funding**, trade-off, and gating decisions by senior management,
- informs due diligencies for in-licensing,
- developing **budgets** and **hiring** plans,
- planning manufacturing **capacity**,
- informs P(launch) for affiliates,
- supporting and transparency and consistency at portfolio level.

Probability of technical success - tool for decision making

Let **data** drive decisions, not the **HI**ghest **P**aid **P**erson's **O**pinion.

FDA might also be interested in DDCP!

ODAC 9th February 2021:
Pembro in high-risk TNBC.

Pembro in high-risk TNBC

Predictive Probability (PP) of EFS Effect

- What is the probability of achieving a statistically significant EFS in a future analysis?
  - **FDA's Model:** Probability at IA4 is **62%-78%** (varies based upon assumptions)
  - **Applicant's Model:**
    - Probability at IA4 is **32%-92%** (varies based upon assumption)
    - Probability for the entire trial is **63%-99%** (varies based upon assumptions)

- Limitations:
  - Models are highly sensitive to the distribution/parameter assumptions, with uncertainty and sensitivity increasing with longer looks into the future.
  - Best at forecasting PP at proximal timepoints

No details on prior, account for interims yes/no, etc. given.

- Meeting briefing document: `https://www.fda.gov/media/145654/download`.
- FDA slides: `https://www.fda.gov/media/145771/download`.
- Merck slides: `https://www.fda.gov/media/145770/download`.

Agenda

How can we quantify probability of technical success?

DDCP (assurance)

Any endpoint type, true effect δ, estimator assumed to follow Normal distribution.

Estimate $\widehat{\delta}_{\text{final}}$ at final analysis of pivotal trial, based on n_{final} observations:

$$\widehat{\delta}_{\text{final}} \quad \sim \quad N(\delta, \sigma^2_{\text{final}} = \sigma^2 / n_{\text{final}}).$$

Pivotal trial is called a success if $\widehat{\delta}_{\textbf{final}} \leq \delta_{\textbf{suc}}$ (think of log hazard ratio).

δ_{suc}: can be

- **Minimal detectable difference** (MDD), i.e. effect size such that trial is "just significant".
- Any **other quantity of interest**, e.g. alternative that gives 80% power $\Rightarrow$ target product profile (TPP).

DDCP

Quantity of interest = **power function**:

$$P_\delta(\widehat{\delta}_{\text{final}} \leq \delta_{\text{suc}}) = \Phi\Big(\frac{\delta_{\text{suc}} - \delta}{\sigma_{\text{final}}}\Big).$$

Depends on true effect $\delta \Rightarrow$ assume distribution over δ with density q and average:

$$\begin{aligned} \text{DDCP}(\delta_{\text{suc}}) &= \mathbb{E}_\delta\Big(P_\delta(\widehat{\delta}_{\text{final}} \leq \delta_{\text{suc}})\Big) \\ &= \int_{-\infty}^{\infty} \Phi\Big(\frac{\delta_{\text{suc}} - \delta}{\sigma_{\text{final}}}\Big) q(\delta) d\delta. \end{aligned}$$

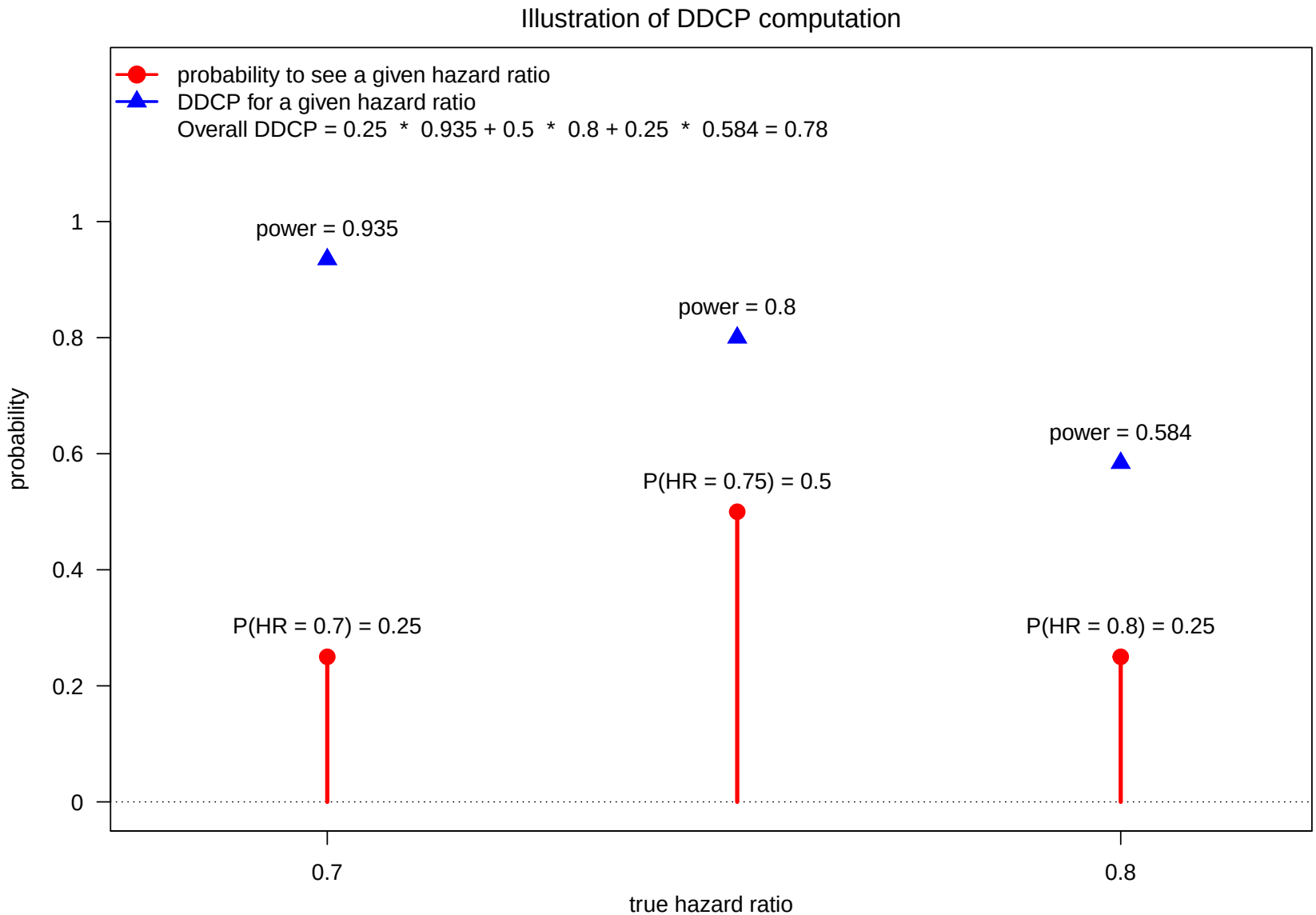
O'Hagan et al. (2001), O'Hagan et al. (2005): "assurance".

Roche-internal: **Data-driven conditional probability**, DDCP.

DDCP informal:

Power averaged over range of potential effect sizes, weighted with how likely we think they are.

Computation of DDCP: example



DDCP vs. Bayesian predictive power

Recall: $\widehat{\delta}_{\text{final}} \leq \delta_{\text{suc}}$ is success.

δ_{MCID}: minimally clinically interesting difference.

Rewrite DDCP:

$$\begin{aligned}
\text{DDCP}(\delta_{\text{suc}}) &= \int_{-\infty}^{\infty} \Phi\Big(\frac{\delta_{\text{suc}} - \delta}{\sigma_{\text{final}}}\Big) q(\delta) d\delta = P(\widehat{\delta}_{\text{final}} \leq \delta_{\text{suc}}) = \\
&= \underbrace{P(\widehat{\delta}_{\text{final}} \leq \delta_{\text{suc}}, \delta \leq \delta_{\text{MCID}})}_{\text{BPP}(\delta_{\text{suc}})} \\
&\quad + \underbrace{P(\widehat{\delta}_{\text{final}} \leq \delta_{\text{suc}}, \delta_{\text{MCID}} < \delta \leq \delta_{\text{mdd}})}_{\text{P(reject but effect irrelevant)}} \\
&\quad + \underbrace{P(\widehat{\delta}_{\text{final}} \leq \delta_{\text{suc}}, \delta_{\text{mdd}} < \delta)}_{\text{P(average type I error)}}.
\end{aligned}$$

- BPP: Bayesian predictive power, Spiegelhalter et al. (1986).
- DDCP: considers "only" significance $\Rightarrow$ irrelevant effects + type I errors are "success".
- Make sure $\delta_{\text{MCID}} \approx \delta_{\text{mdd}}$.

DDCP vs. Bayesian predictive power

Often, $\text{BPP}(\delta_{\text{mdd}}) \approx \text{DDCP}(\delta_{\text{mdd}})$, see Kunzmann et al. (2021).

go.roche.com/successR/methodology.html

Time-to-event endpoint

Approximate distribution of **estimated log(hazard ratio)** $\widehat{\theta} := \log(\widehat{\text{HR}})$:

$$\widehat{\theta} \quad \approx \quad N(\theta, 4/d).$$

- $\theta = \log(\text{HR})$: **true underlying effect**, true log-hazard ratio.
- d: total number of events in both arms.
- 1:1 randomized trial: $\text{Var}(\widehat{\theta}) = 4/d$.
- Non-1:1: $\tau = P(\text{arm A}) \Rightarrow \text{Var}(\widehat{\theta}) = [\tau(1-\tau)d]^{-1}$.

Example

Assumptions:

- Phase 2 result: $\widehat{\theta}_{\text{Phase 2}} = \log(\mathbf{0.700})$, based on $d_{\text{prior}} = 50$ **events**.
- Phase 3: 80% power to detect hazard ratio 0.75.
- Final analysis after $d_{\text{final}} = 380$ events based on estimate $\widehat{\theta}_{\text{final}} \sim N(\theta, \sigma^2_{\text{final}} = 4/d_{\text{final}})$.
- Minimal detectable difference at final analysis: $\theta_{\text{suc}} = \log(0.818)$.

DDCP (assurance) at start of Phase 3, assuming we know Phase 2 result:

$$\text{DDCP} = \int_{-\infty}^{\infty} P_\theta(\widehat{\theta}_{\text{final}} \leq \theta_{\text{suc}}) \phi_{\mu=\log(0.700),\sigma^2=4/50}(\theta) d\theta = \mathbf{0.697}.$$

go.roche.com/successR/cont_bpp_design.html
go.roche.com/successR/t2e_bpp_design.html

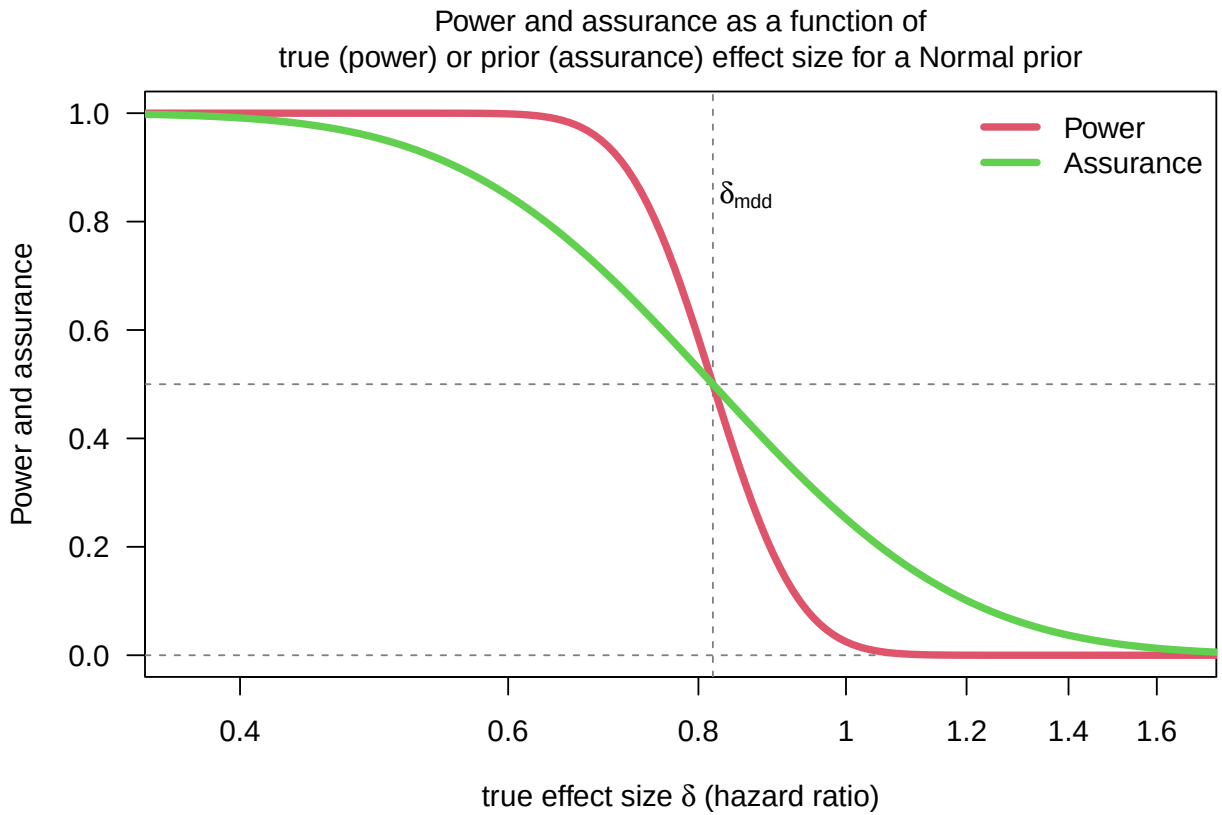
Agenda

1. What is PTS?
2. How can we quantify probability of technical success?
3. DDCP typically lower than power
4. Prior sample size and DDCP
5. What triggers a DDCP update?
   - Update DDCP with external information
   - Update DDCP after an interim analysis
6. Choice of prior
7. What about single-arm studies?
8. Resources
9. Take home messages
10. Backup

DDCP typically lower than power

Question from decision-makers: "DDCP is smaller than power?"

DDCP smaller than Power if Power ≥ 0.5 for commonly used priors.



Normal prior: use explicit formulas to show.

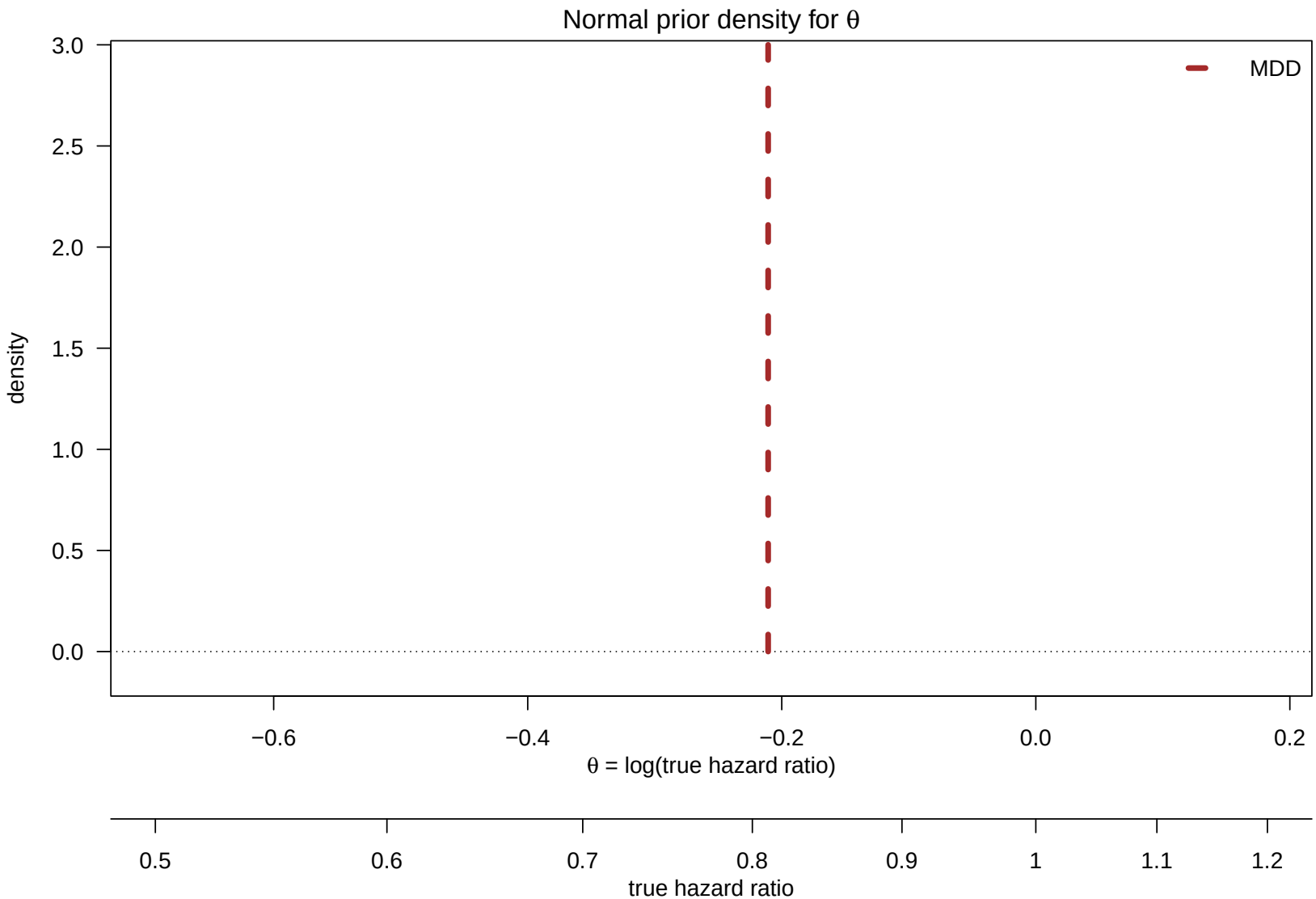
Rufibach et al. (2016): any **symmetric and unimodal** prior.

Agenda

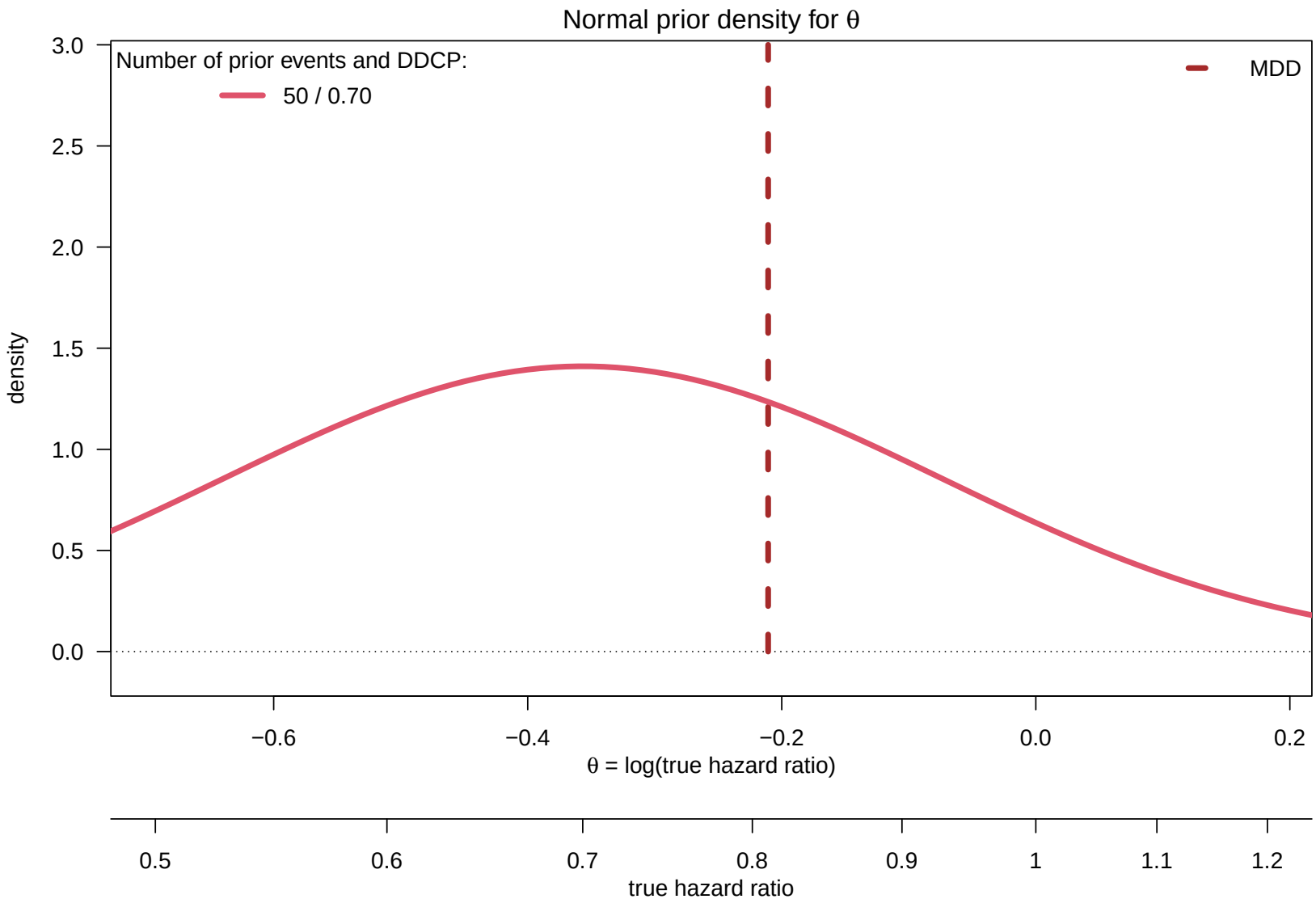
1. What is PTS?
2. How can we quantify probability of technical success?
3. DDCP typically lower than power
4. Prior sample size and DDCP
5. What triggers a DDCP update?
   - Update DDCP with external information
   - Update DDCP after an interim analysis
6. Choice of prior
7. What about single-arm studies?
8. Resources
9. Take home messages
10. Backup

Prior sample size and DDCP

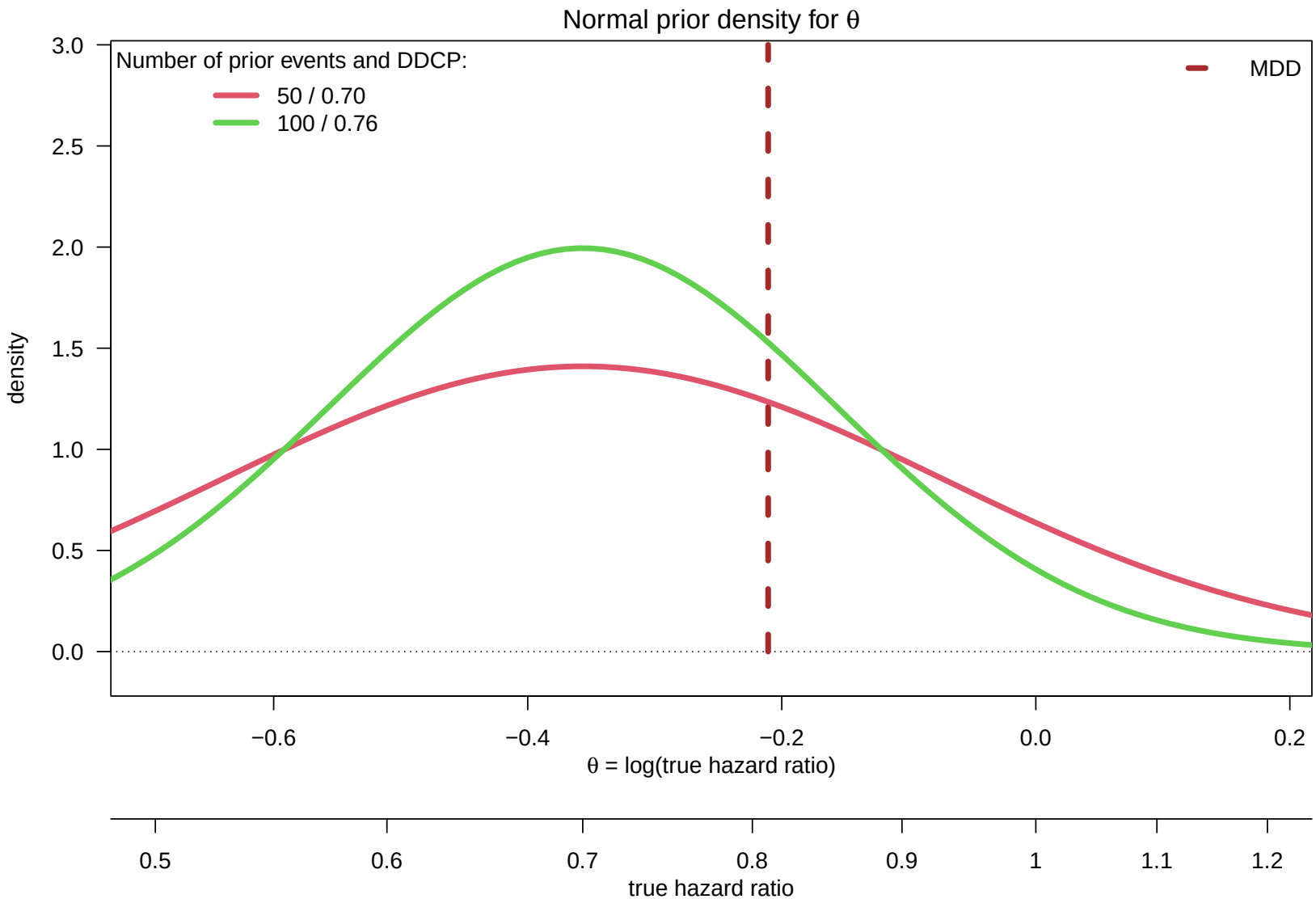
Prior sample size and DDCP



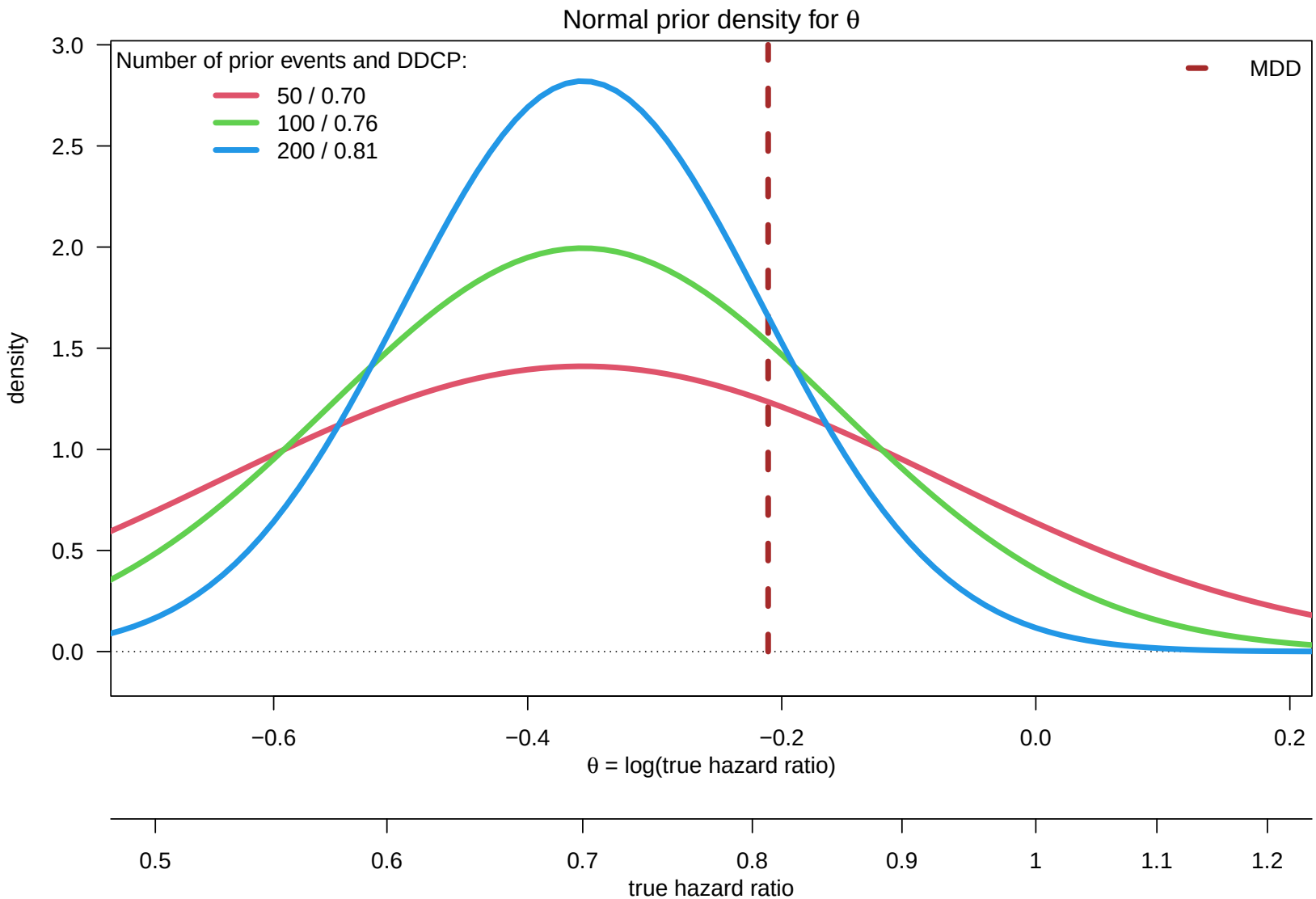
Prior sample size and DDCP



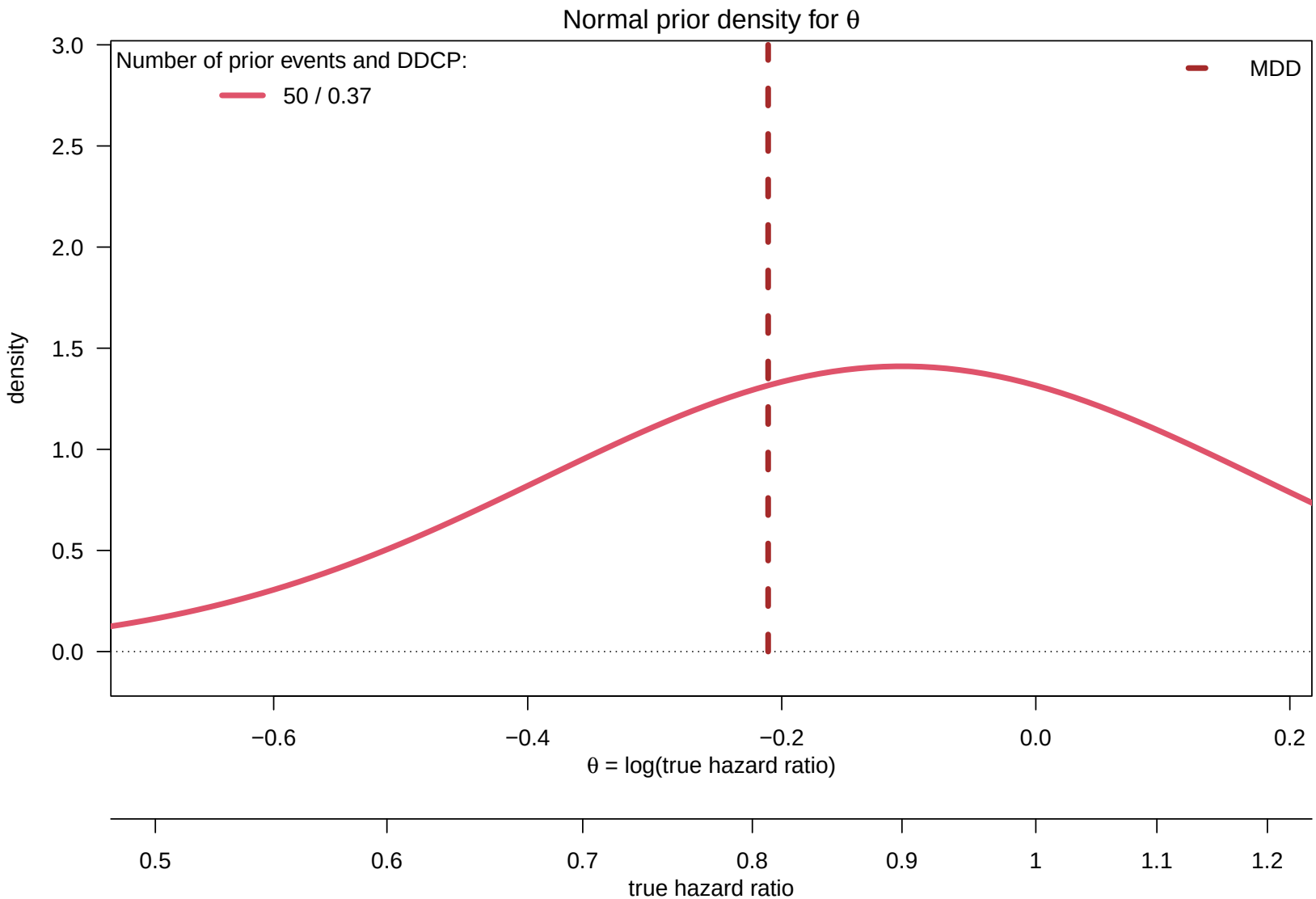
Prior sample size and DDCP



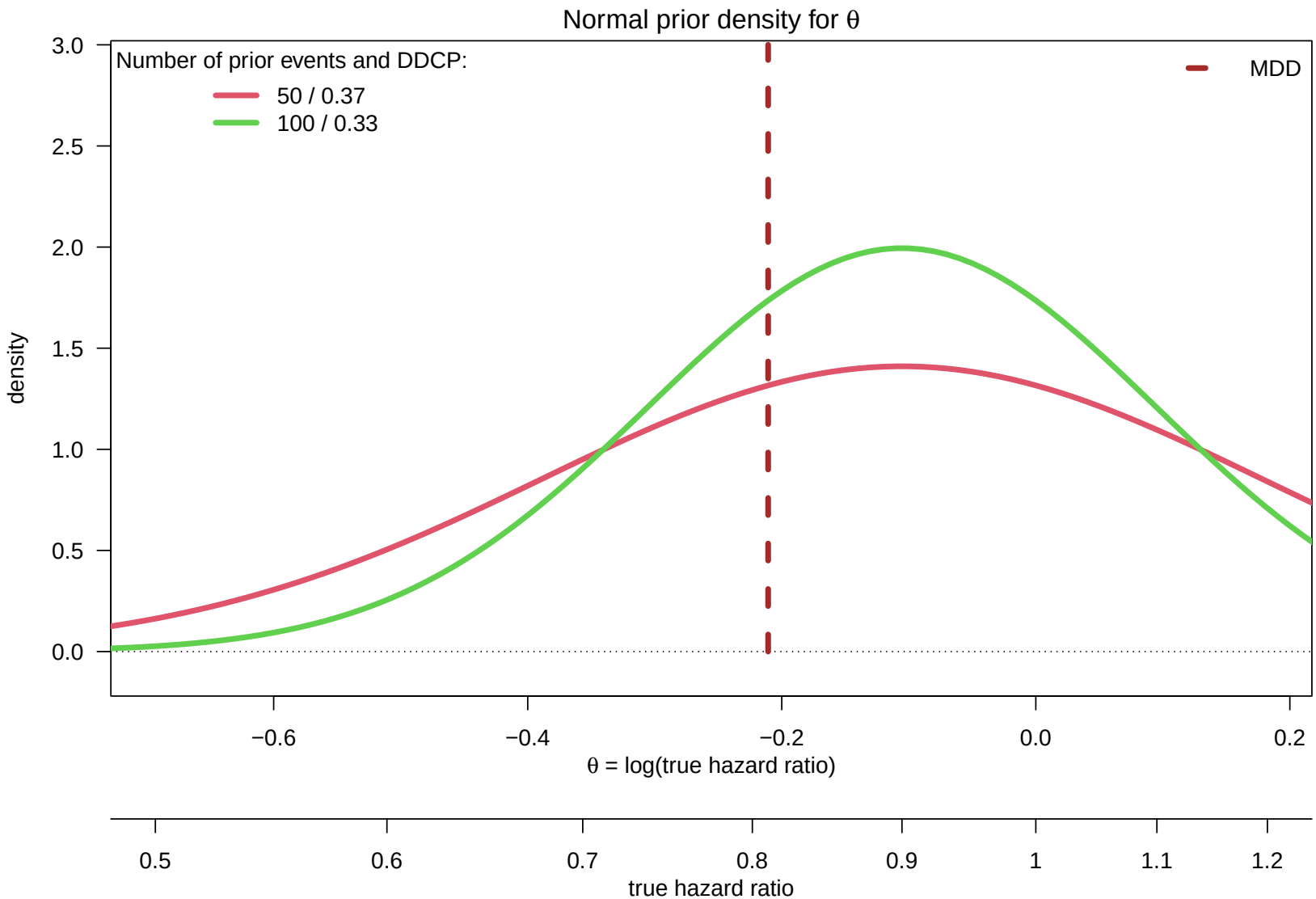
Prior sample size and DDCP



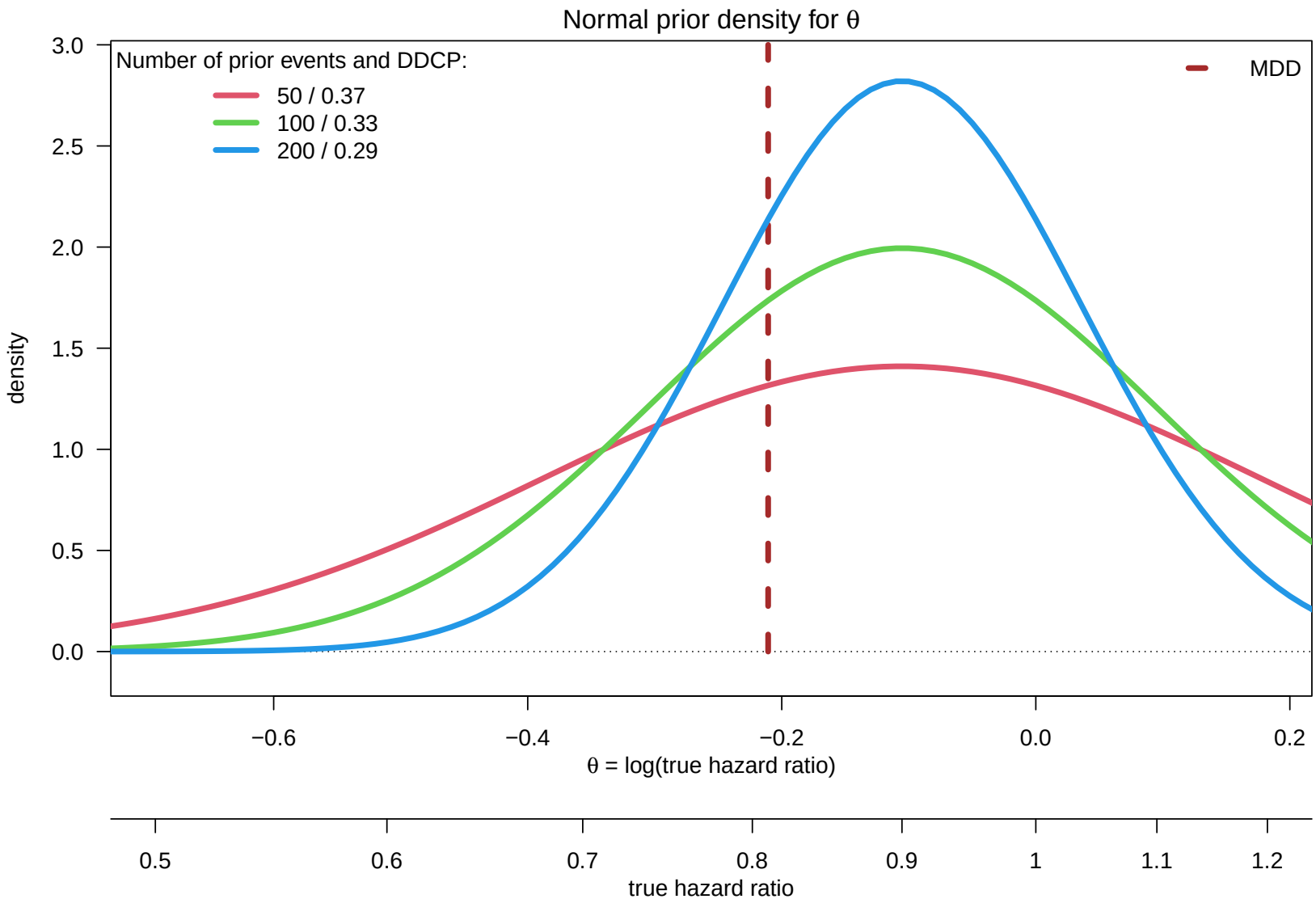
Prior sample size and DDCP



Prior sample size and DDCP



Prior sample size and DDCP



Be very careful using DDCP to chose P2 sample size!

Maybe better do not do it at all.

Old (?) LSPC requirement

Pivotal trial powered $\Rightarrow$ MDD.

Determine effect size in Phase 2 s.t. Phase 3 DDCP $= 0.6$.

Appears to only make sense if prior mean larger effect than Phase 3 MDD!

If prior mean smaller effect than Phase 3 MDD $\Rightarrow$ **decreasing sample size increases DDCP!**

Agenda

What triggers a DDCP update?

Update DDCP with external information

Assume we have **external** estimate $\widehat{\delta}_{\text{extern}}$ of treatment effect, with $\text{SE}(\widehat{\delta}_{\text{extern}})$:

- Competitor or collaborative group trial, ...
- Internal trial in same or related program, ...

Quantify knowledge with Normal density q_{data}, update prior q_{prior} to get $q_{\text{posterior}}$.

DDCP formula...

$$\text{DDCP} \quad = \quad \int_{-\infty}^{\infty} P_{\delta}(\widehat{\delta}_{\text{final}} \leq \delta_{\text{suc}}) q_{\text{prior}}(\delta) d\delta$$

...becomes:

$$\text{DDCP} \quad = \quad \int_{-\infty}^{\infty} P_{\delta}(\widehat{\delta}_{\text{final}} \leq \delta_{\text{suc}}) q_{\text{posterior}}(\delta) d\delta.$$

Simply update prior with external information, recompute DDCP.

Power part remains unaffected.

Update after interim (blinded or unblinded)

JOURNAL OF BIOPHARMACEUTICAL STATISTICS
2016, VOL. 26, NO. 2, 191–201
http://dx.doi.org/10.1080/10543406.2014.972508

Taylor & Francis
Taylor & Francis Group

Sequentially updating the likelihood of success of a Phase 3 pivotal time-to-event trial based on interim analyses or external information

Kaspar Rufibach, Paul Jordan, and Markus Abt

F. Hoffmann-La Roche Ltd., Product Development Biostatistics, Basel, Switzerland

ABSTRACT

When performing a pivotal clinical trial, it may be of interest to assess the probability of success (PoS) of that trial. Initially evaluated when the trial is designed, PoS can be updated as the trial progresses and new information about the drug effect becomes available. Such information can be external to the trial, such as results from trials conducted in parallel, or internal, such as continuing after an interim analysis. We develop a framework to update PoS based on such internal and external information for a time-to-event endpoint and illustrate it using a realistic development program for a new molecule.

ARTICLE HISTORY

Received 8 May 2014
Accepted 30 September 2014

KEYWORDS

Bayesian predictive power; conditional power; interim analysis; prior distribution; probability of technical success

How does DDCP change if we do not stop at a futility interim?

Futility interim analysis only

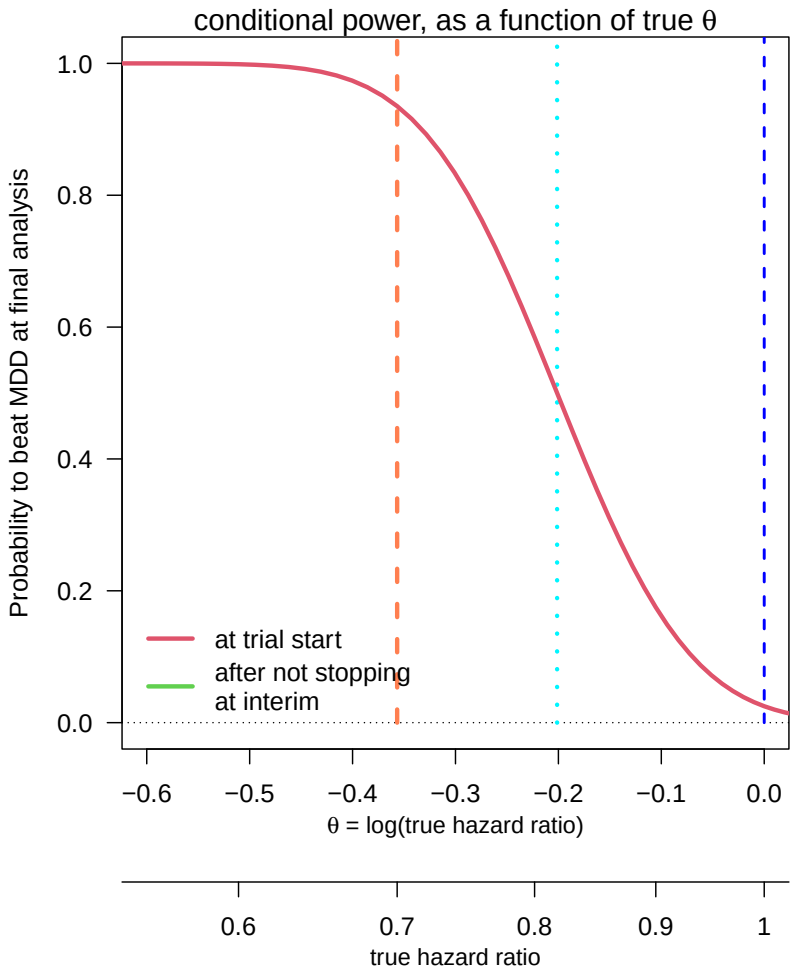
Blinded futility interim passed with boundary HR ≤ 1: we know that

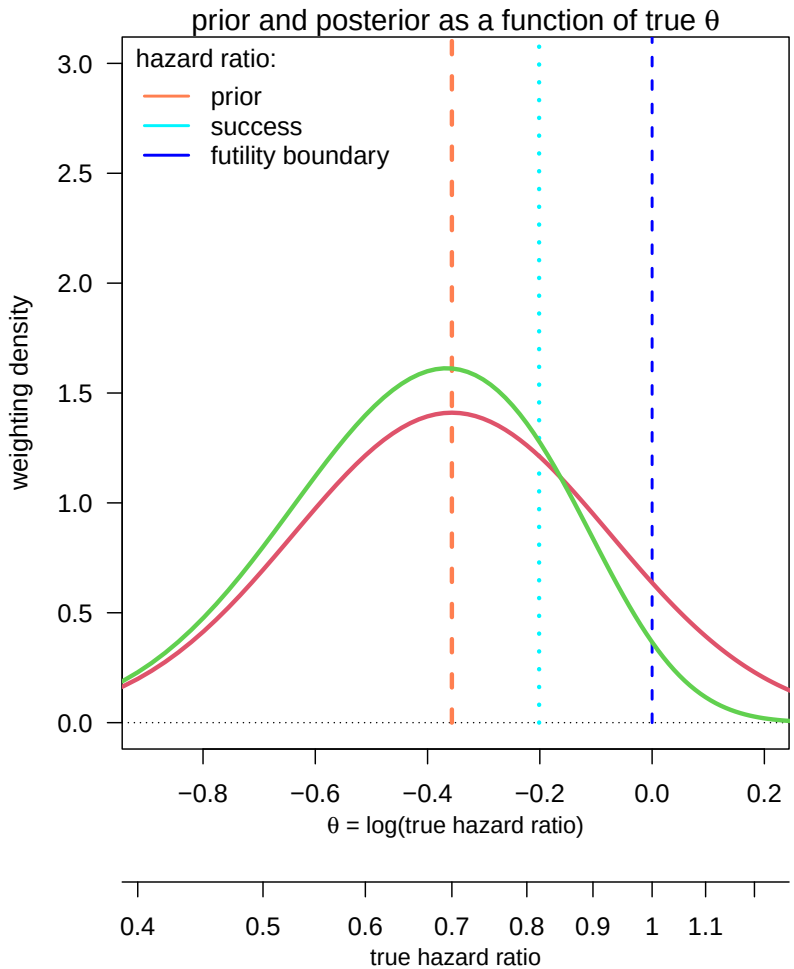
- $0 < \text{HR} \leq 1$ or
- $\widehat{\theta}_{\text{interim}} \in (-\infty, \log(1)]$.

How does DDCP change after this interim?

- DDCP unchanged?
- Increase it?
- Decrease it?

Futility interim analysis only - plot both factors in DDCP formula





Green density **not** a Normal density.

Futility interim analysis only - comments

After not stopping at interim, DDCP increases from **0.697** to **0.801**.

Why does DDCP **increase** after not stopping?

- Prior with mean log(0.7) assigns weight to hazard ratios smaller than hazard ratio to finally beat, $\theta_{\text{suc}} = \log(0.818)$.
- **Not stopping** shifts mass of prior q_{prior} to the left of 1 for $q_{\text{posterior}} \Rightarrow$ more weight on hazard ratios $\leq \theta_{\text{suc}}$.
- Together with small increase in conditional power accounts for **higher DDCP** after not stopping.

Does DDCP decrease or increase after not stopping?

Trial does not stop at **futility interim** $\Rightarrow$ DDCP increases.

Trial does not stop at **efficacy interim** $\Rightarrow$ DDCP decreases.

Extent depends on configuration of

- prior distribution,
- minimal detectable difference at final analysis θ_{suc},
- variability of final analysis estimate,
- efficacy interim boundary θ_{efficacy},
- futility interim boundary θ_{futility}.

go.roche.com/successR/cont_bpp_update.html
go.roche.com/successR/t2e_bpp_update.html

bpp: function for **two interim analyses** (Normal prior only).

Agenda

Choice of prior

Choice of prior

Bathtub effect!

MAIN PAPER

Pharmaceutical Statistics

(wileyonlinelibrary.com) DOI: 10.1002/pst.1764

Published online 21 July 2016 in Wiley Online Library

Bayesian predictive power: choice of prior and some recommendations for its use as probability of success in drug development

Kaspar Rufibach,* Hans Ulrich Burger, and Markus Abt

Bayesian predictive power, the expectation of the power function with respect to a prior distribution for the true underlying effect size, is routinely used in drug development to quantify the probability of success of a clinical trial. Choosing the prior is crucial for the properties and interpretability of Bayesian predictive power. We review recommendations on the choice of prior for Bayesian predictive power and explore its features as a function of the prior. The density of power values induced by a given prior is derived analytically and its shape characterized. We find that for a typical clinical trial scenario, this density has a u-shape very similar, but not equal, to a β-distribution. Alternative priors are discussed, and practical recommendations to assess the sensitivity of Bayesian predictive power to its input parameters are provided.

Keywords: Bayesian predictive power; conditional power; prior distribution; probability of technical success

go.roche.com/successR/prior_choice.html

Plan to have more discussions around prior choice in follow-up hands-on tutorials.

Agenda

What about single-arm studies?

DDCP vs. predictive probability

DDCP: average power function over assumed distribution of true effect:

$$\text{DDCP} \quad = \quad \int_{-\infty}^{\infty} \Phi\Big(\frac{\delta_{\text{suc}} - \delta}{\sigma_{\text{final}}}\Big) q(\delta) d\delta.$$

Predictive probability: posterior probability of rejecting H_0 if trial runs to maximum planned sample size:

$$\text{PP} \quad = \quad \mathbb{E}_{Y|X}\Big[1\Big\{P(P_E \geq p_0|x, Y) > \theta_T|x\Big\}\Big].$$

$P(P_E \geq p_0|x, Y = i)$: Probability that response rate $> p$ given x responses among n first patients and i responses among $N_{\text{max}} - n$ future patients.

P_E: response proportion new treatment, random variable with posterior after first x patients.
x: #responses among first n patients.
Y: #responses among $N_{\text{max}} - n$ future patients.
θ_T: threshold for posterior probability of event $\{P_E \geq p_0\}$.
$\mathbb{E}$: w.r.t to bayesian posterior predictive distribution of y given x.

DDCP vs. predictive probability

	Single-arm studies - Unicycle	DDCP
Question to answer	quantify probability to meet "mini TPP" (designed for the ph1b trial) given prior knowledge about effect size	quantify probability to beat minimal or target product profile in Phase 2 or 3 given prior knowledge about effect size
Average over	$P(P_E \geq p_0 \mid x, Y = i)$ above θ_T weighted with the probability of seeing i responses among $N_{\max} - n$ future patients	distribution over effect size
Trial type of interest	single arm trials	typically RCTs, but in theory any normally distributed test statistic allows for use of framework
Endpoint types	binary	binary, continuous, time-to-event
Endpoints	ORR	any: response proportion difference, mean difference, PFS, OS, ...
Update after interim	Yes	Yes

go.roche.com/successR/phase1b.html

Agenda

Resources

Resources

go.roche.com/successR

Computations rely on **bpp** package: available on CRAN and pre-installed on BEE. Rufibach et al. (2022).

Technical details:

- **go.roche.com/successR/training_material.html**, especially BayesPharma Lyon slidedeck and papers referenced therein.
- **go.roche.com/successR/further_resources.html**.
    - Shiny app: `https://rsconnect.roche.com/content/4807/`. Not developed by successR. Strong recommendation: verify computations from shiny app using **bpp**, store code for **reproducibility**.

We need you!

Add further examples: trial examples for **continuous** and **binary** endpoint!

Add templates of slidedecks to share computations with teams.

You want to contribute a case study?

Download template
go.roche.com/successR/questions.html

We plan to organize follow-up sessions with code practicals.

Agenda

Take home messages

Take home messages

- If you want to compute probability of success first define **success**.
- Roche: **DDCP**. Literature: **assurance**. "Average probability to reject" ("marginal" in Kunzmann et al. (2021)).
- Initial DDCP: requires **prior knowledge** about effect of interest.
  - Act of synthesizing evidence already meaningful for teams.
  - Extrapolation from one endpoint type (binary) to another (T2E): project Endeavour. Gazyva example. Maybe more to come on successR.
- DDCP $\leq$ power $\Rightarrow$ recalibrate stakeholders.
- Do not stop at **futility interim** $\Rightarrow$ DDCP $\uparrow$.
- Do not stop at **efficacy interim** $\Rightarrow$ DDCP $\downarrow$.
- Do not use DDCP to chose Phase 2 sample size.
- Be clear on what you want to do - use appropriate quantity.

We plan follow-up sessions with **hands-on tutorials**.

Whether successR will be a success or become flopR

depends on everyone's contributions and input!

go.roche.com/successR

References I

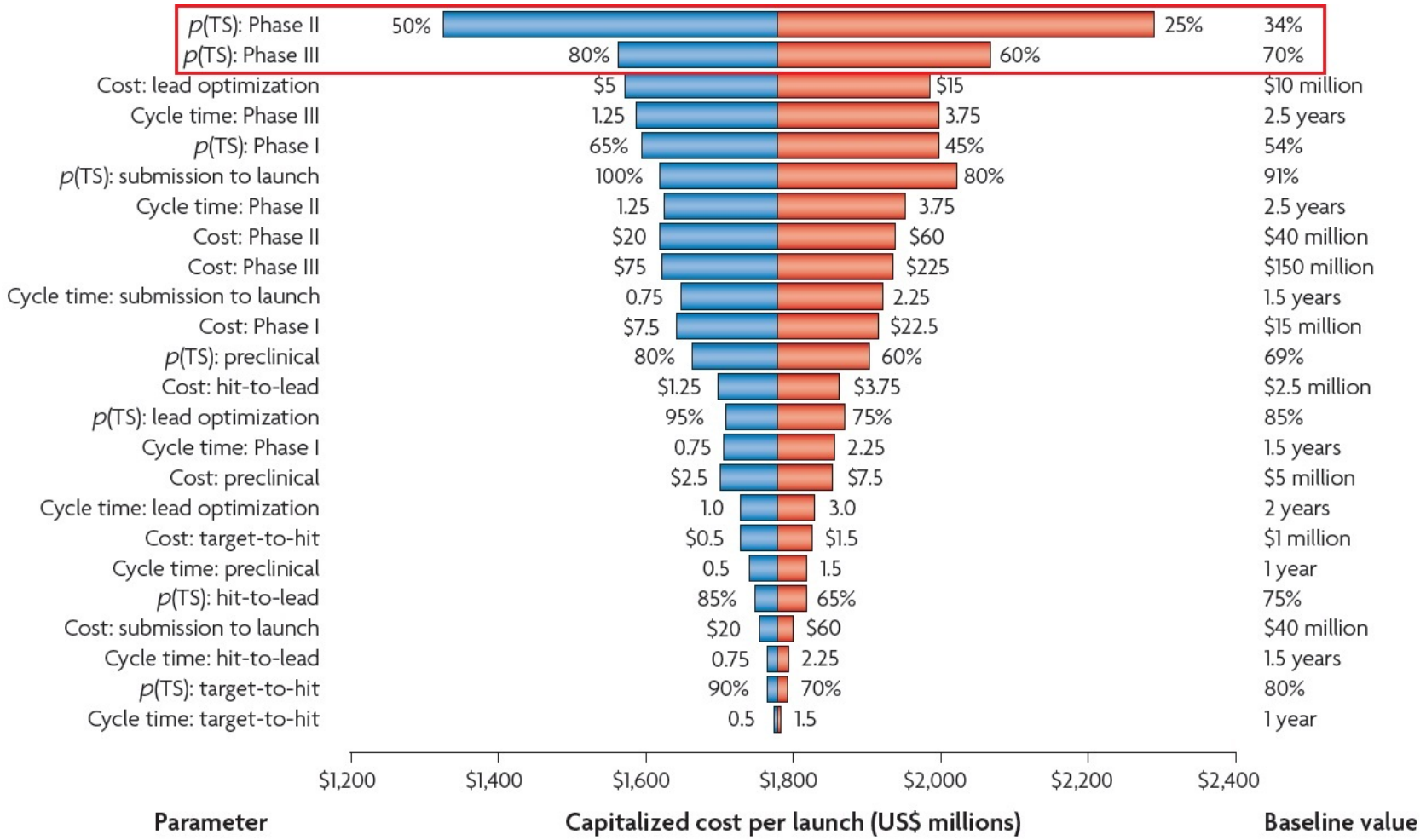
- Kunzmann, K., Grayling, M. J., Lee, K. M., Robertson, D. S., Rufibach, K. and Wason, J. M. S. (2021). A Review of Bayesian Perspectives on Sample Size Derivation for Confirmatory Trials. *Am Stat* **75** 424–432.
- O'Hagan, A., Stevens, J. W. and Campbell, M. J. (2005). Assurance in clinical trial design. *Pharm. Stat.* **4** 187–201.
- O'Hagan, A., Stevens, J. W. and Montmartin, J. (2001). Bayesian cost-effectiveness analysis from clinical trial data. *Stat Med* **20** 733–753.
- Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R. and Schacht, A. L. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat Rev Drug Discov* **9** 203–214.
- Ringel, M., Tollman, P., Hersch, G. and Schulze, U. (2013). Does size matter in R&D productivity? If not, what does? *Nat Rev Drug Discov* **12** 901–902.
- Rufibach, K., Burger, H. and Abt, M. (2016). Bayesian predictive power: Choice of prior and some recommendations for its use as probability of success in drug development. *Pharm. Stat.* **15** 438–446.
- Rufibach, K., Jordan, P. and Abt, M. (2022). *bpp: Computations Around Bayesian Predictive Power*. R package version 1.0.4. `https://CRAN.R-project.org/package=bpp`
- Spiegelhalter, D., Reedman, L. and Blackburn, P. (1986). Monitoring clinical trials - conditional power or predictive power. *Control Clin Trials* **7** 8–17.

Agenda

Backup

R&D productivity model: influence of driving factors

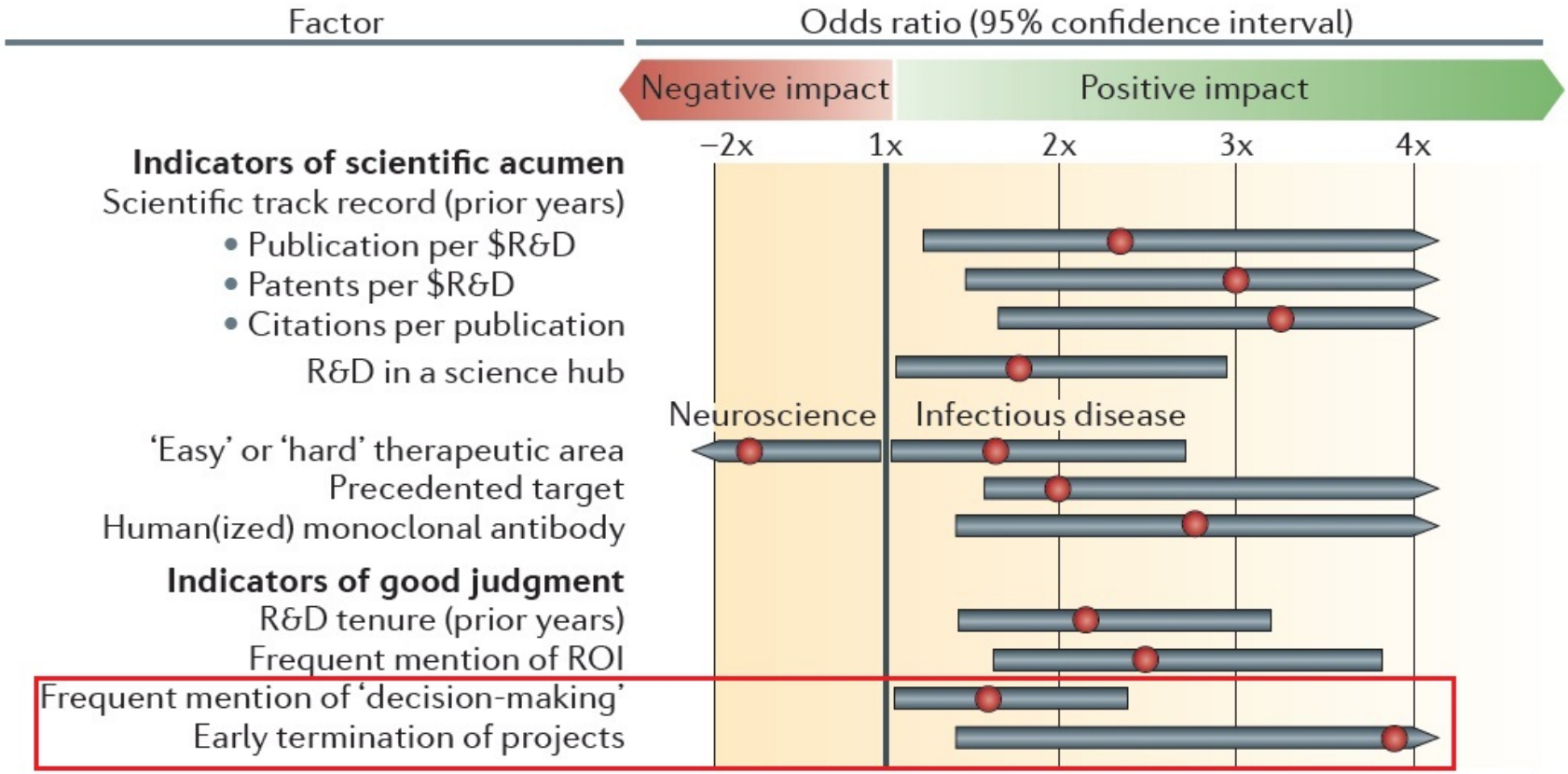
Increase p(TS) for Phase 3 from 70% (=baseline) to 80% ⇒ **save 12% / 213 Mio $**.



Parametric sensitivity analysis created from an R&D model. Paul et al. (2010).

Factors associated with success or failure in drug development

Dataset of **842** molecules, Ringel et al. (2013).



Variables with strong **univariate** relationship to success or failure.

High probability of termination in pre-clinical and Phase 1: **OR for success 3.9**.

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: bpp / mvtnorm / rpact / reporttools / xtable

This document was generated on 2023-11-13 at 13:45:13.