Statistics boost 9:
Trial success probabilities (PTS & DDCP)

Godwin Yung & Jianmei Wang
Statistical Methods, Collaboration & Outreach Group (MCO)
PD Data & Statistical Sciences
September, 2023

Monthly 1-hour sessions organized by the MCO group from PD DSS.

Stats boost PD connect posts: post 1, post 2

Slides & recordings @ R&D Academy: link

Topic	Date EU/China: 10-11am CET, NA : 11-12am PT
Basics of inferential statistics	Nov 7 (EU/China) & 8 (NA), 2022
Introduction to survival analysis	Dec 5 (EU/China) & 6 (NA), 2022
Sample size calculation	Jan 9, 2023
Statistical considerations for complex innovative designs and master protocols	Feb 13, 2023
Efficacy and futility interim analyses: Why, how, when?	March 27, 2023
Subgroup analyses in confirmatory trials	April 24, 2023
Bayesian vs. frequentist methodologies	June 5, 2023
Estimands	June 26, 2023
Trial success probabilities (DDCP & PTS)	Sep 4, 2023
Survival analysis - advanced topics	Oct 2, 2023
Multiple testing in clinical trials	Oct 30, 2023
Confirmatory adaptive designs	Nov 27, 2023
Non-inferiority and equivalence trials	Jan 15, 2024
Externally Controlled Trials	Feb 12, 2024

Acknowledgments

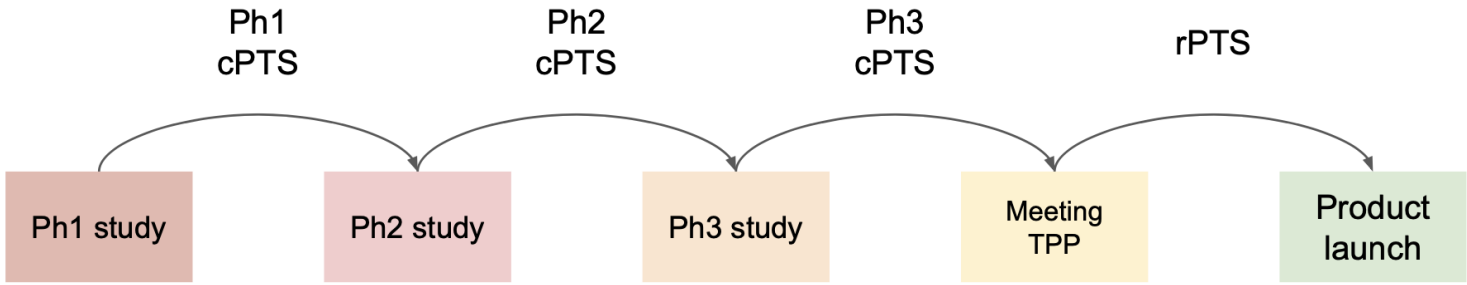
These slides are built on top of material previously created by Kaspar Rufibach. His contributions are gratefully acknowledged. We also thank other MCO members for their feedback.

Goal of today's session:

Understand the composition and use of PTS and DDCP.

What is PTS?

Probability of technical success (PTS)

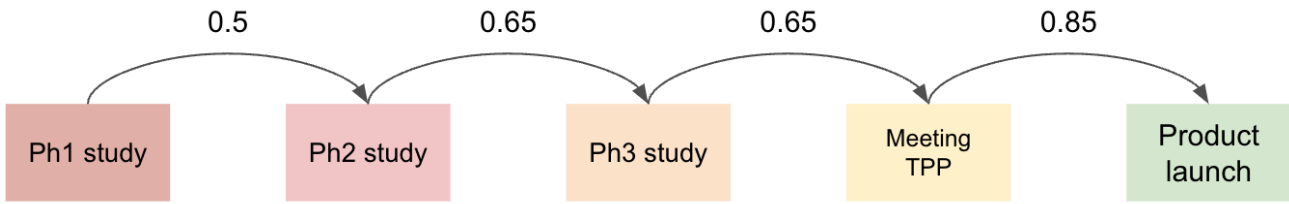


PTS late stage handbook (Slide 41):

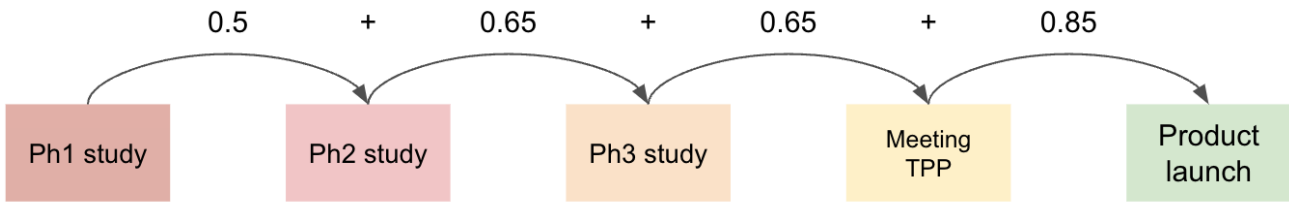
- **PTS**: Probability of successfully transitioning to the next phase.
  - **Clinical PTS (cPTS)**: Probability of meeting efficacy and safety requirements. **Ph3 cPTS** is more specifically the probability of achieving the target product profile (TPP).
  - **Registration PTS (rPTS)**: Probability of gaining approval with the submitted data package.
- **Probability of launch (POL)**

Exercise 1

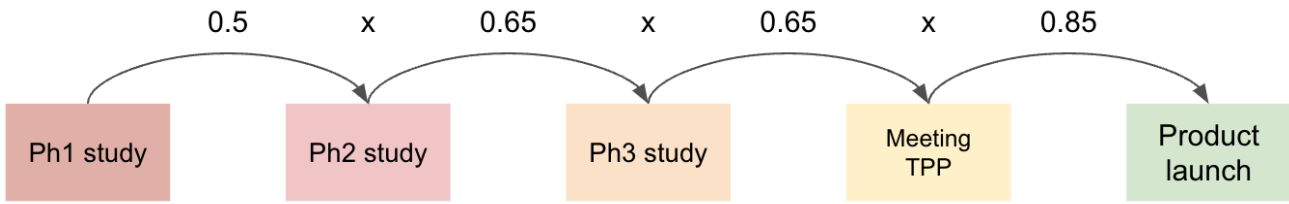
What is the probability of launch in this example?



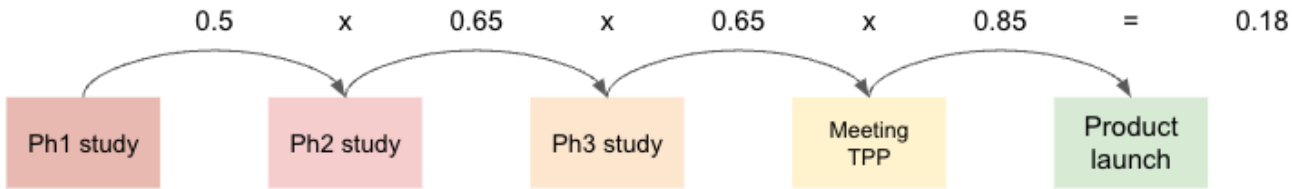
Is it (A)



Or is it (B)



Solution

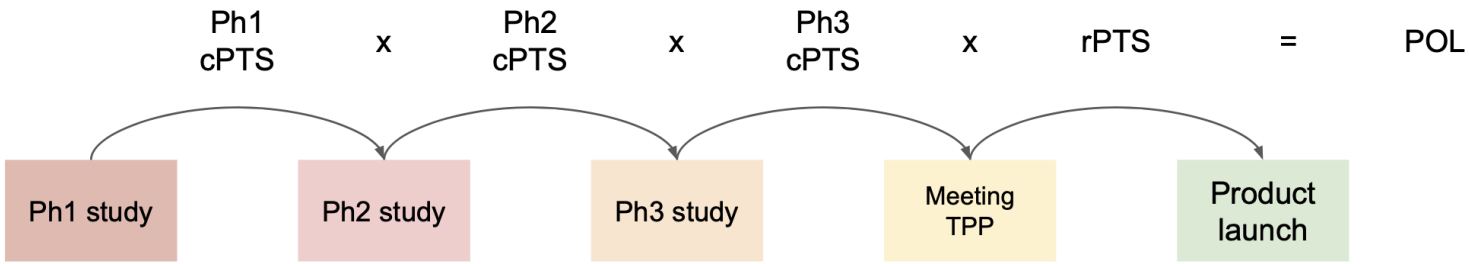


Solution is (B).

- Easy answer: Process of elimination. Probabilities add up to over 1, an impossibility.
- More appropriate answer: Chain rule of probability.

Probabiliy of launch (POL)

POL is the product of the individual phase probabilities and registration PTS.



- PTS for a prior phase becomes 100% when a decision is made to move to the next phase.

What is PTS used for?

- Calculating project **valuations**
- Supporting funding, trade-off, and **gating decisions** by senior management
- Informing due diligence for in-licensing
- Developing budgets and hiring plans
- Planning manufacturing capacity
- Informing probability of launch (POL) for affiliates
- Supporting transparency and consistency at the portfolio level

Use of PTS is only successful if calculations are accurate, consistent, and based on transparent assumptions.

Looking further underneath the hood ...

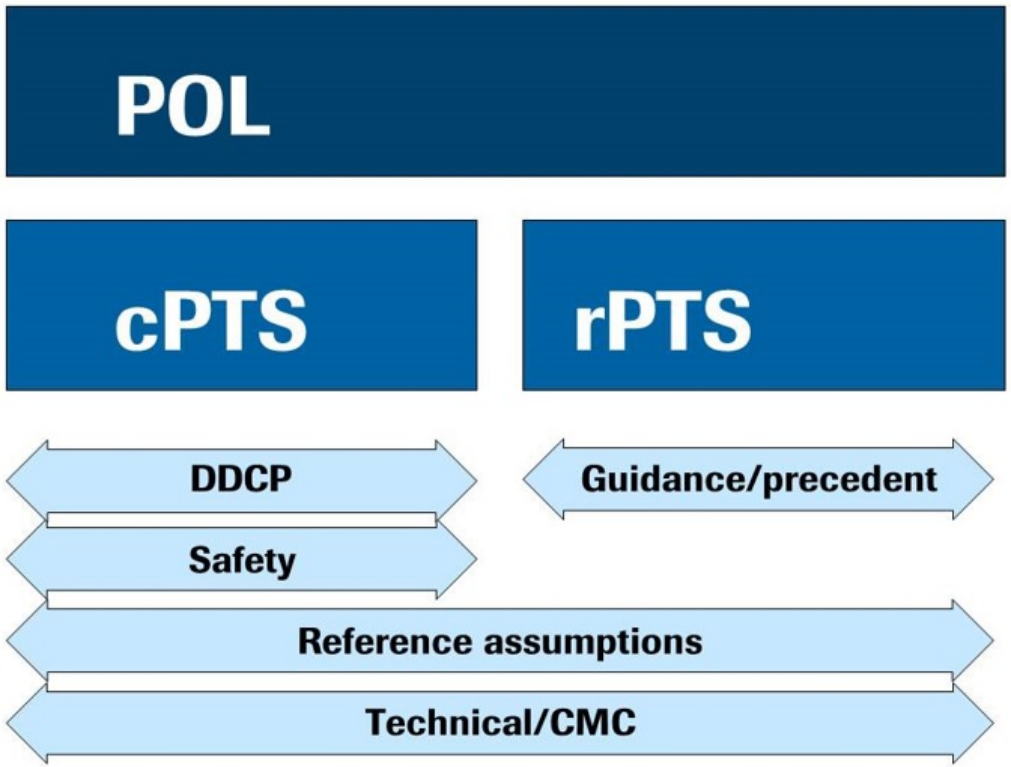


Figure: PTS handbook.

What is DDCP?

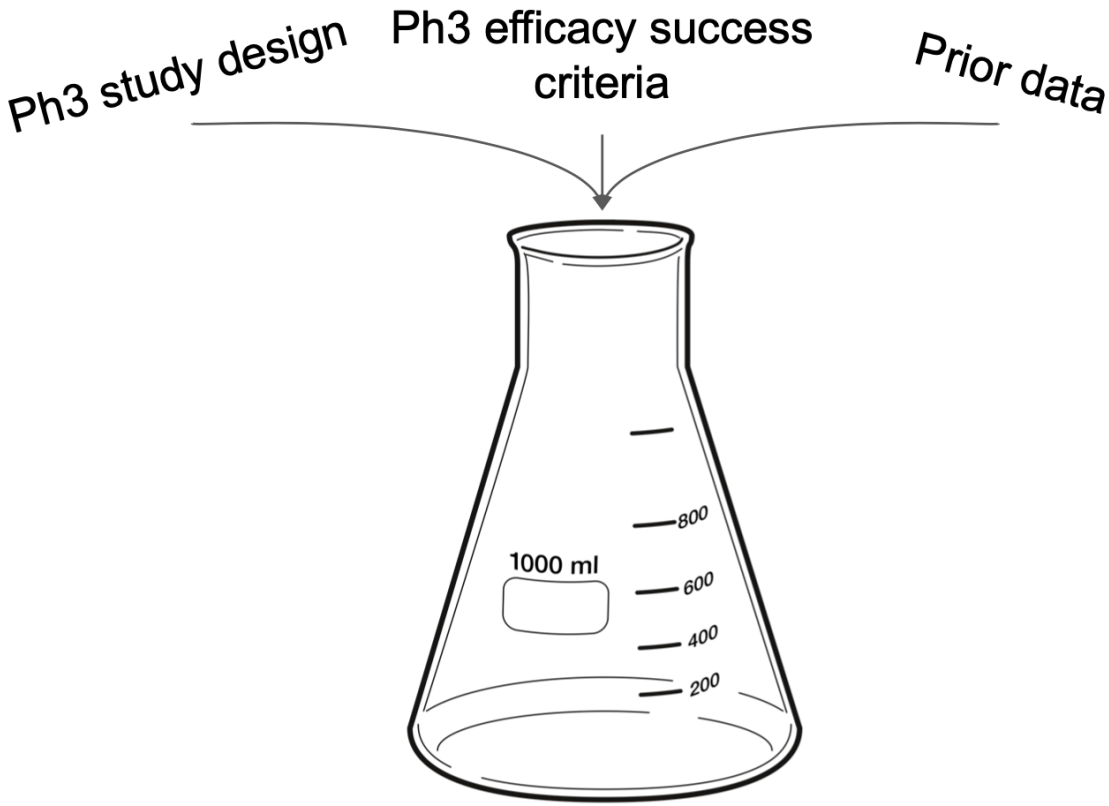
Data driven conditional probability (DDCP)

"**Probability** of a Ph3 study achieving its efficacy target **given prior data**"

Hence its Roche-internal name! Known as "assurance" in the literature.

Ingredients in DDCP

"Probability of a **Ph3 study** achieving **efficacy target** given **prior data**"



"Efficacy target"

Efficacy target could be ...

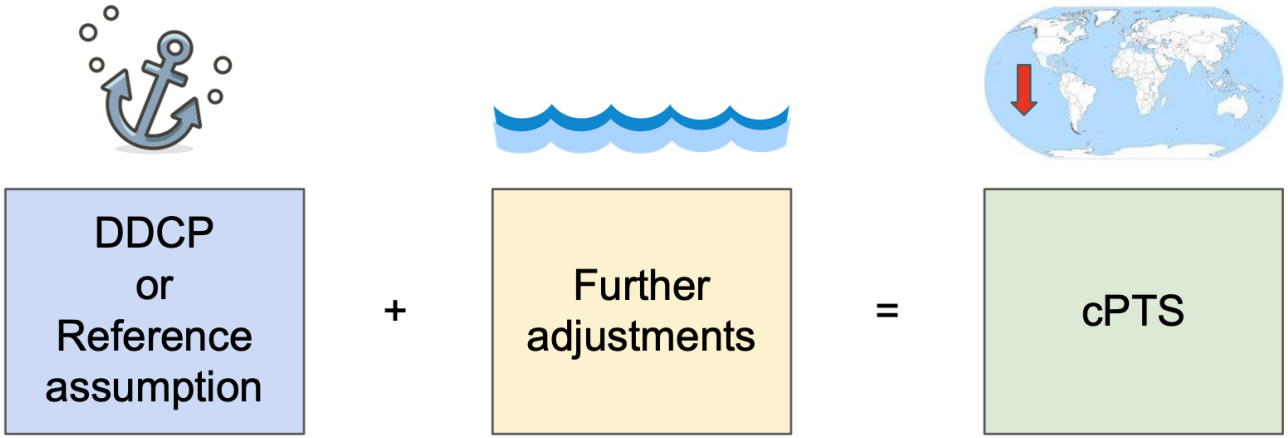
- Target product profile (TPP): desired characteristics of a product
  - **Target TPP**: optimal; typically aligned with power calculations
  - **Minimum TPP**: minimal but still sufficient; typically aligned with the minimal detectable difference (MDD), the smallest difference that still produces a statistically significant result
- Any other value that is **fixed** and **pre-specified**

"Prior data"

"Prior" is a term that is meant to capture the existing knowledge we have of the Ph3 treatment effect prior to starting the Ph3.

- **Ph1/2 data**
  - Single arm or randomized?
  - Same endpoint or different endpoint?
- **External information** (e.g., similar MOA)
  - Differences in estimand (population, treatment, outcome variable, etc.)?
  - Effective sample size?

DDCP serves as an anchor for cPTS.



Further adjustments are made based on additional, qualitative information. E.g.,

- safety
- similar molecules
- uncertainty around assumptions
- differences between Ph2 and Ph3 design (population, endpoint, etc.)
- CMC

Exercise 2

Consider a Ph3 trial design with a survival endpoint:

1. Compute sample size based on assumed effect size, power, and α.
   - Assuming the truth is HR=0.75, 380 events are needed to have 80% power to reject H0 (HR=1) with 2-sided α=0.05.
2. Determine MDD based on α and sample size.
   - MDD = 0.82. Reject H0 if observed HR $\leq$ 0.82; otherwise, fail to reject H0.

(Refer back to Sessions 1 and 3 for a refresher on α, power, MDD, and TPP.)

Exercise 2

In this example, our **Ph3 study was designed** to have 80% power to detect HR of 0.75 at $\alpha = 0.05$. MDD is 0.82. What is the approximate value of DDCP for the Ph3 study, if we have the following success criteria and based on earlier data (including Ph2), the team is convinced that the true Ph3 HR is 0.75?

- **Success criteria**: observed HR ≤ 0.82
- **Prior data**: HR $= 0.75$

Is it

A. $1 - \alpha = 95\%$?

B. $1 - \beta = 80\%$?

C. 50%?

D. 0%?

Solution

Solution is (B).

The success criteria "observed HR $\leq$ 0.82" is aligned with the MDD. Based on prior data, the team believes the true HR is 0.75 with very little uncertainty. The Ph3 trial is successful if H0 is rejected. In the Ph3 design, there was 80% power to reject H0 if the true HR=0.75. Therefore, the DDCP is $\sim$80%.

Now what if we change the success criteria to "observed HR $\leq$ 0.75"?

For the success criteria "observed HR $\leq$ 0.75", since the true HR is around 0.75, there is approximately 50-50 chance that the Ph3 estimated HR is above or below 0.75. Therefore, the DDCP is approximately $\sim$50%.

Something doesn't seem right ...

In reality, we assume a treatment effect under H_1 for the purpose of calculating sample size and MDD at which we have 80% power.

However, **we are never 100% confident about that treatment effect assumption**. (If we were, we wouldn't conduct a Ph3 study!)

How do we calculate DDCP in practice?

We consider **multiple potential values** for the true treatment effect.

- What values do I think the true treatment effect might be and how likely do I think that they are indeed the truth?
- Under each potential treatment effect, what is the probability that I reject H_0?

(Potential values can follow a discrete or continuous distribution, which we call the **prior distribution**.)

Example

Revisiting the Ph3 study in Exercise 2 with 80% power to detect HR of 0.75 ...

Hazard ratio	Probability of being true	Power
0.70	0.25	0.935
0.75	0.50	0.800
0.80	0.25	0.584

Then

$$\text{DDCP} = 0.25 \times 0.935 + 0.50 \times 0.800 + 0.25 \times 0.584 = 0.780$$

Alternative definition of DDCP

"Probability of a Ph3 study achieving efficacy target given prior data."

"Power averaged over a range of potential effect sizes, with each effect size weighted by how likely we think they are to be true."

How does DDCP behave?

DDCP is typically smaller than power

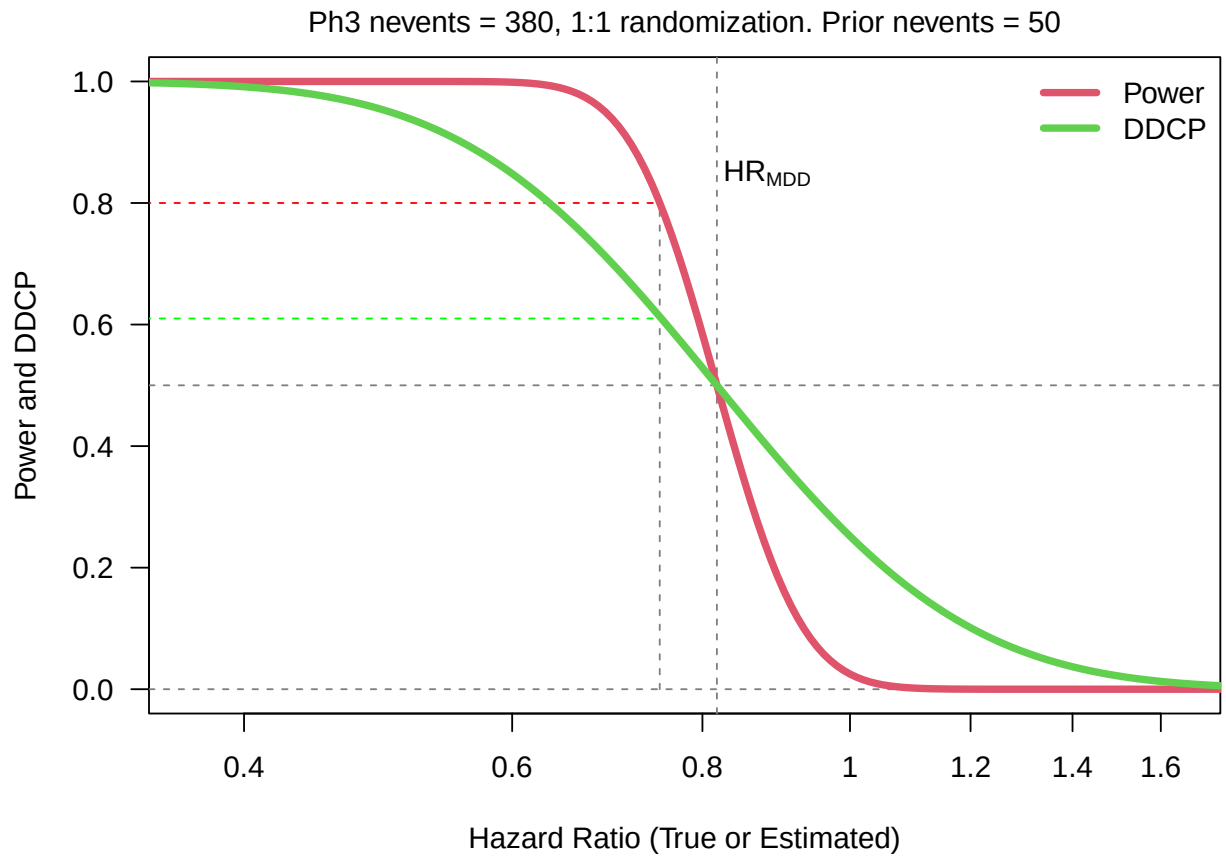
Revisit the Ph3 trial design in Exercises 2: 380 events, 2-sided α=0.05, 1:1 randomization. MDD for HR = 0.82.

Assume we have prior data from 50 events. Ph3 success criteria for DDCP is aligned with MDD: estimated HR $<$ 0.82.

- If **true** HR = 0.75, then **power = 80%**.
- If the **estimated** HR = 0.75 from prior data, then **DDCP = 61%**.

DDCP is typically smaller than power

In the previous example, consider other true / estimated HRs, DDCP is smaller than power if power ≥ 0.5.



Exercise 3: The better the Ph2 result, the _____ the DDCP.

A. higher

B. lower

C. it depends

Higher. The better you believe the true treatment effect to be, the more likely you are to achieve the Ph3 efficacy target (whatever it may be).

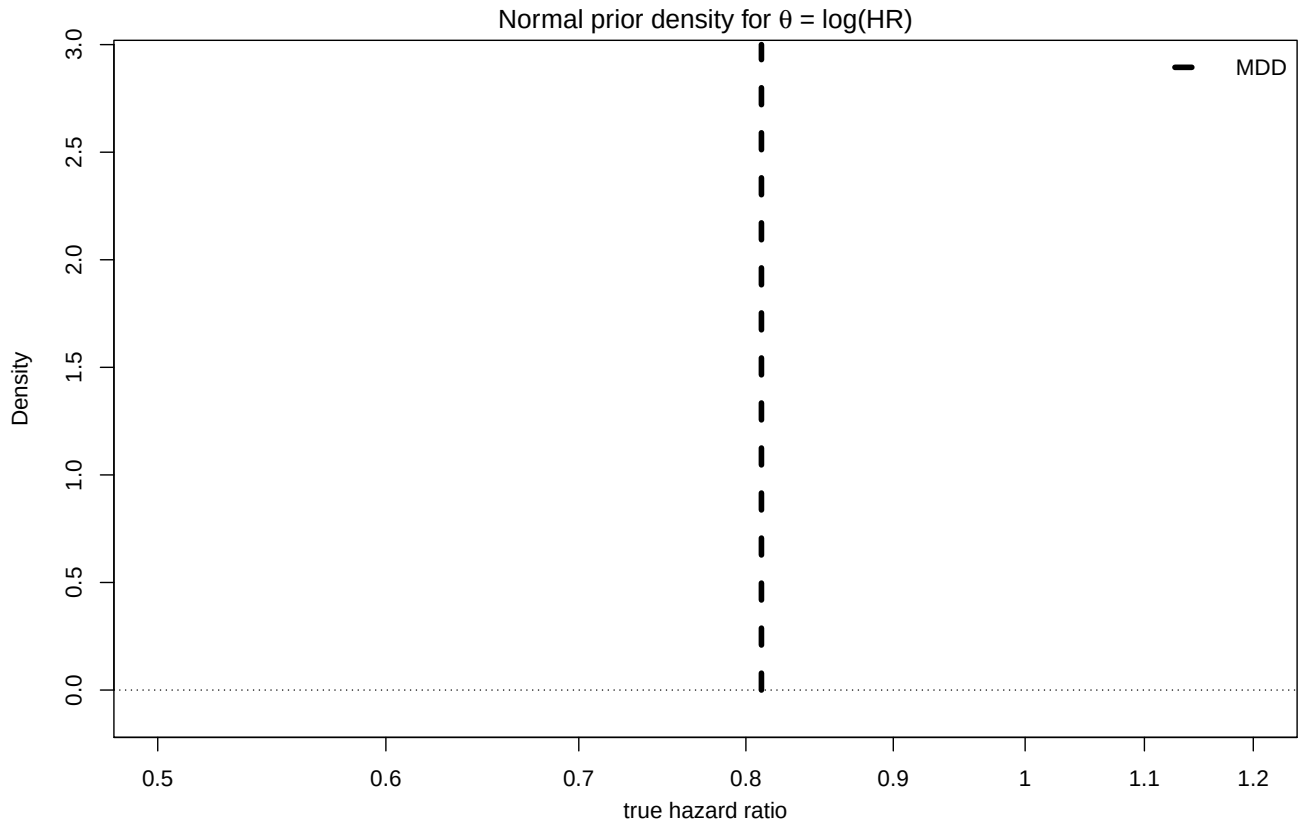


Figure: Revisiting the Ph3 study in Exercise 2: 380 events planned, 2-sided $\alpha = 0.05$, and **success criteria: HR $<$ MDD $= 0.82$**.

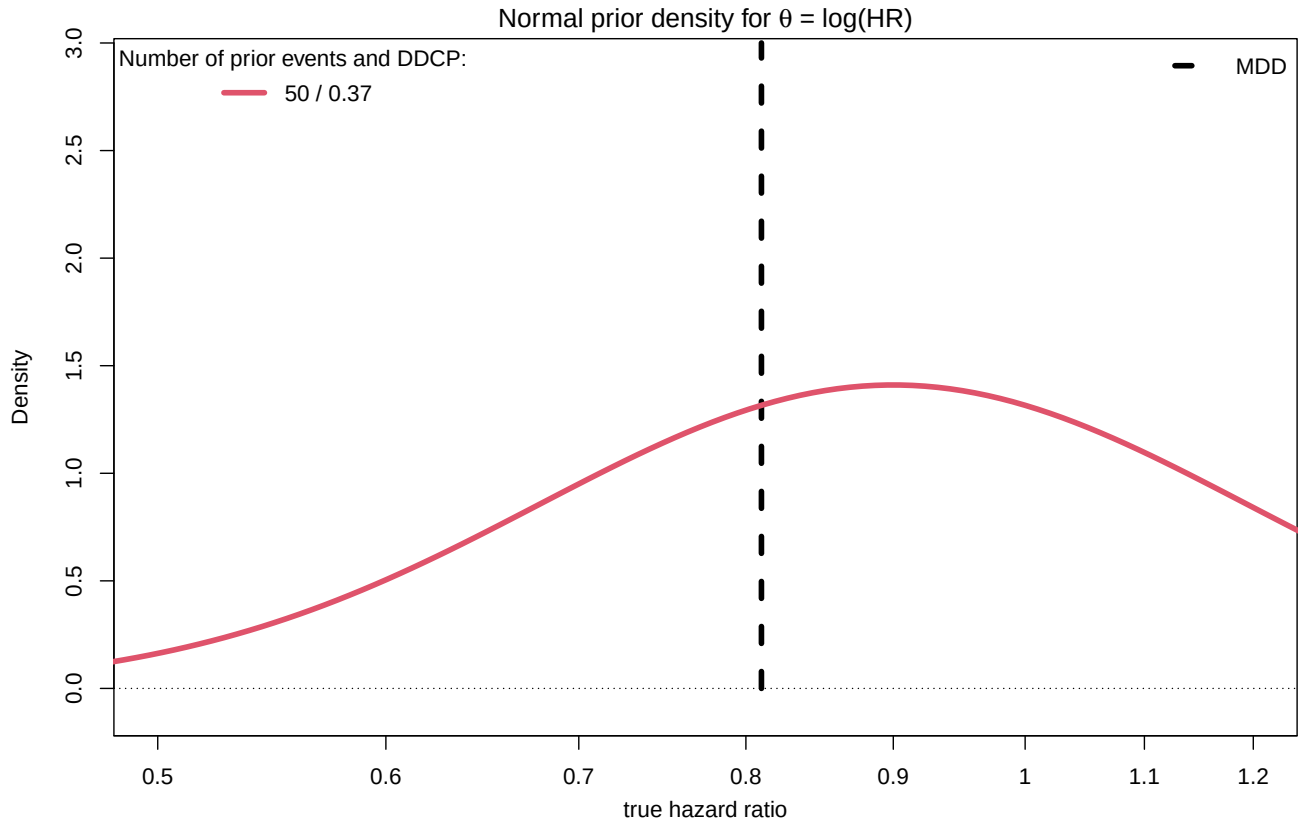


Figure: Prior distribution given **50 Ph2 events** and a **less favorable Ph2 outcome**. The resulting **DDCP=0.37** is calculated based on the Ph3 study design, success criteria = MDD, and this prior.

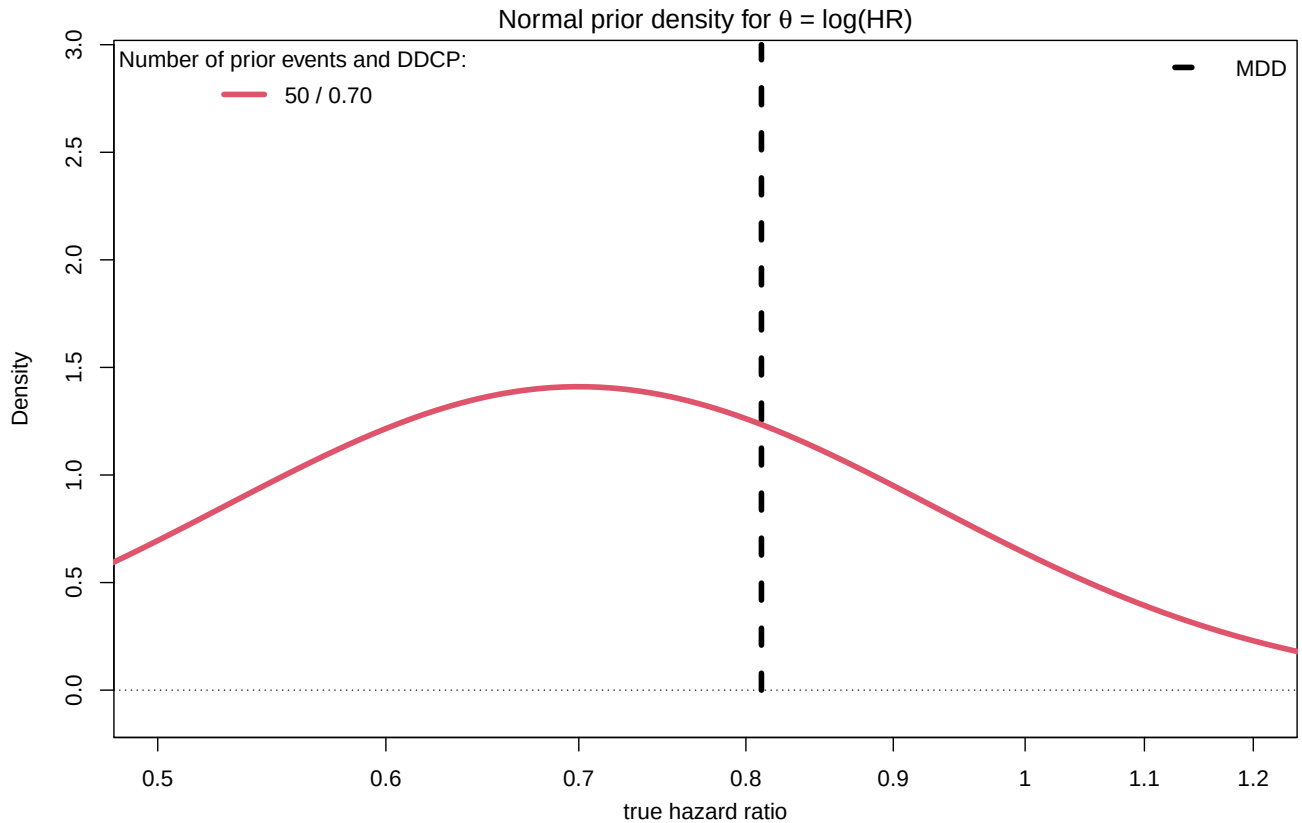


Figure: Prior distribution given **50 Ph2 events** and a **more favorable Ph2 outcome**. The resulting **DDCP=0.70**.

Exercises 4: The bigger the Ph2 sample size, the _____ the DDCP.

A. higher

B. lower

C. it depends

It depends. Large Ph2 sample size gives more confidence in the observed Ph2 treatment effect. If this treatment effect is strong, then you are more likely to achieve the Ph3 efficacy target. However, if this treatment effect is weak, then you are less likely to achieve the Ph3 efficacy target. (Note: What constitutes as strong vs. weak is relative to the efficacy target.)

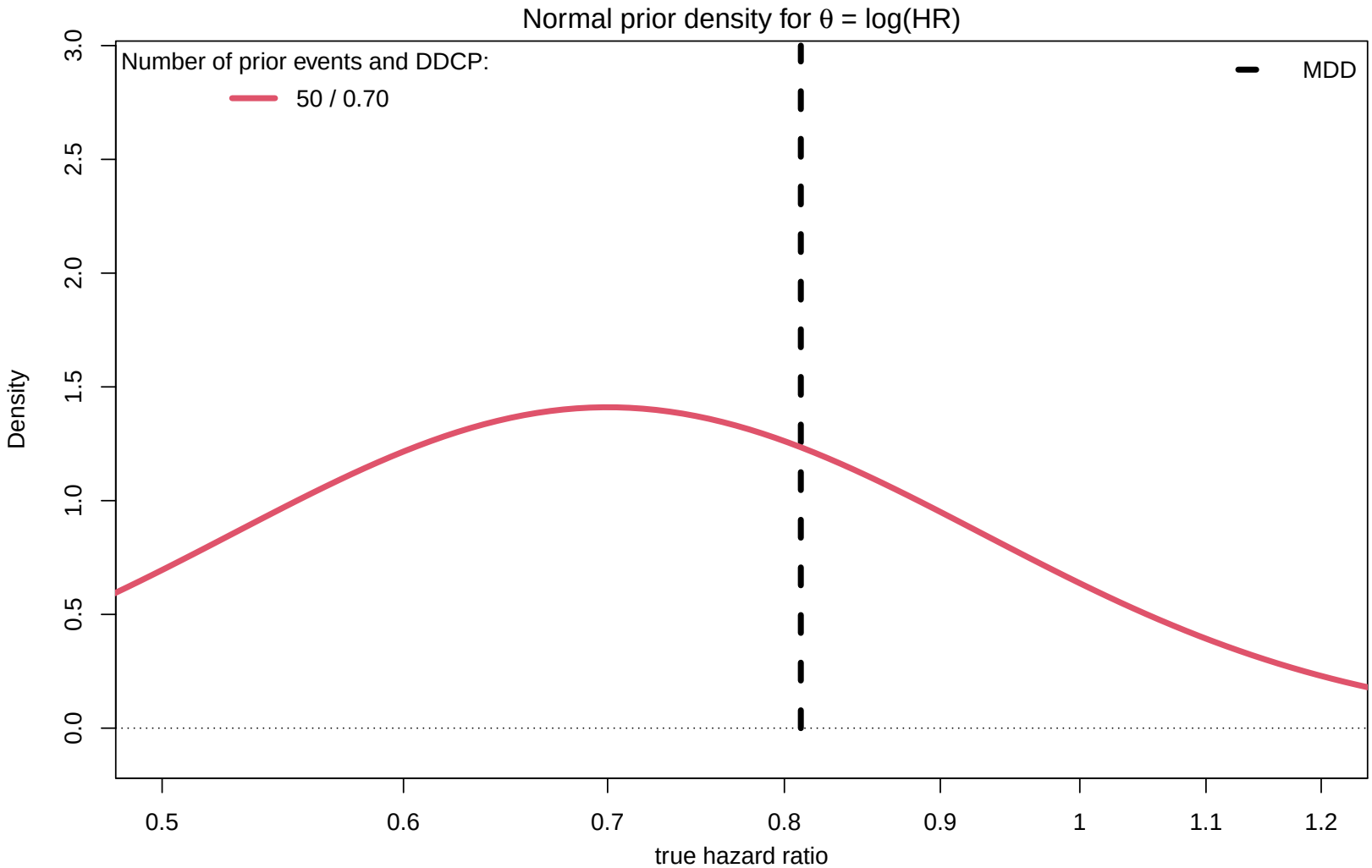


Figure: Prior distribution given **50 Ph2 events** and a **more favorable Ph2 outcome**. The resulting **DDCP=0.70**.

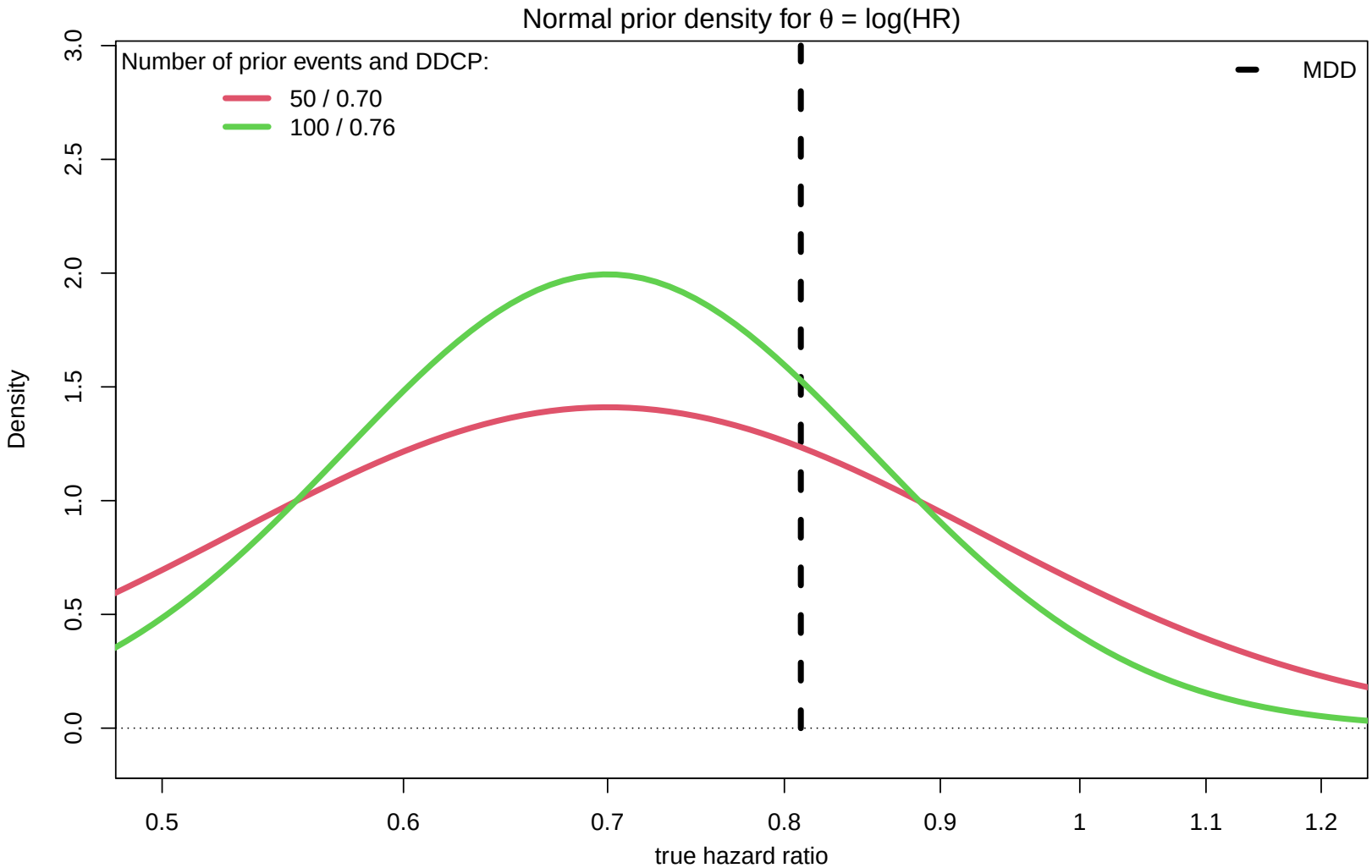


Figure: Prior distribution given **100 Ph2 events** but the **same, more favorable Ph2 outcome**. The resulting **DDCP=0.76**.

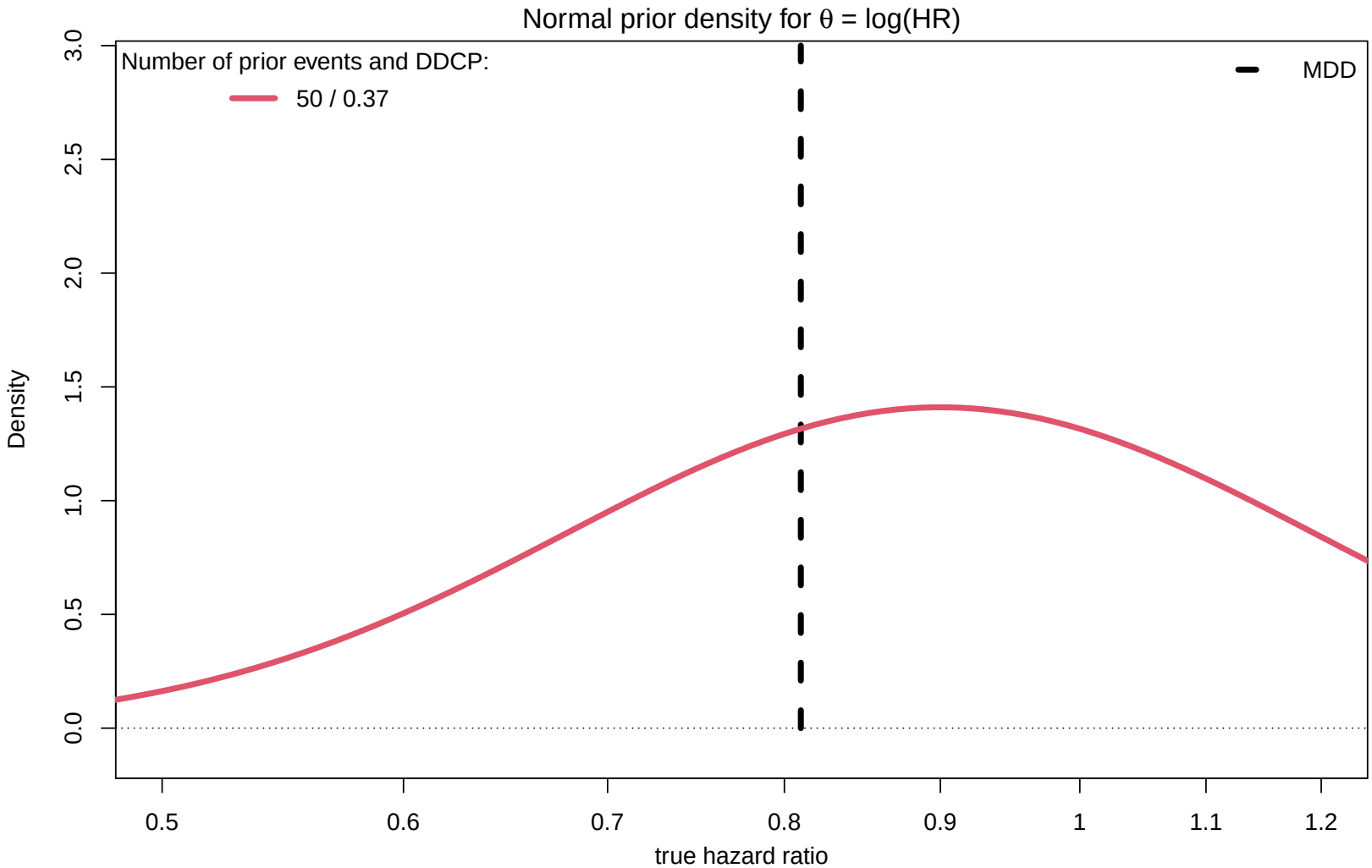


Figure: Prior distribution given **50 Ph2 events** and a **less favorable Ph2 outcome**. The resulting **DDCP=0.37**.

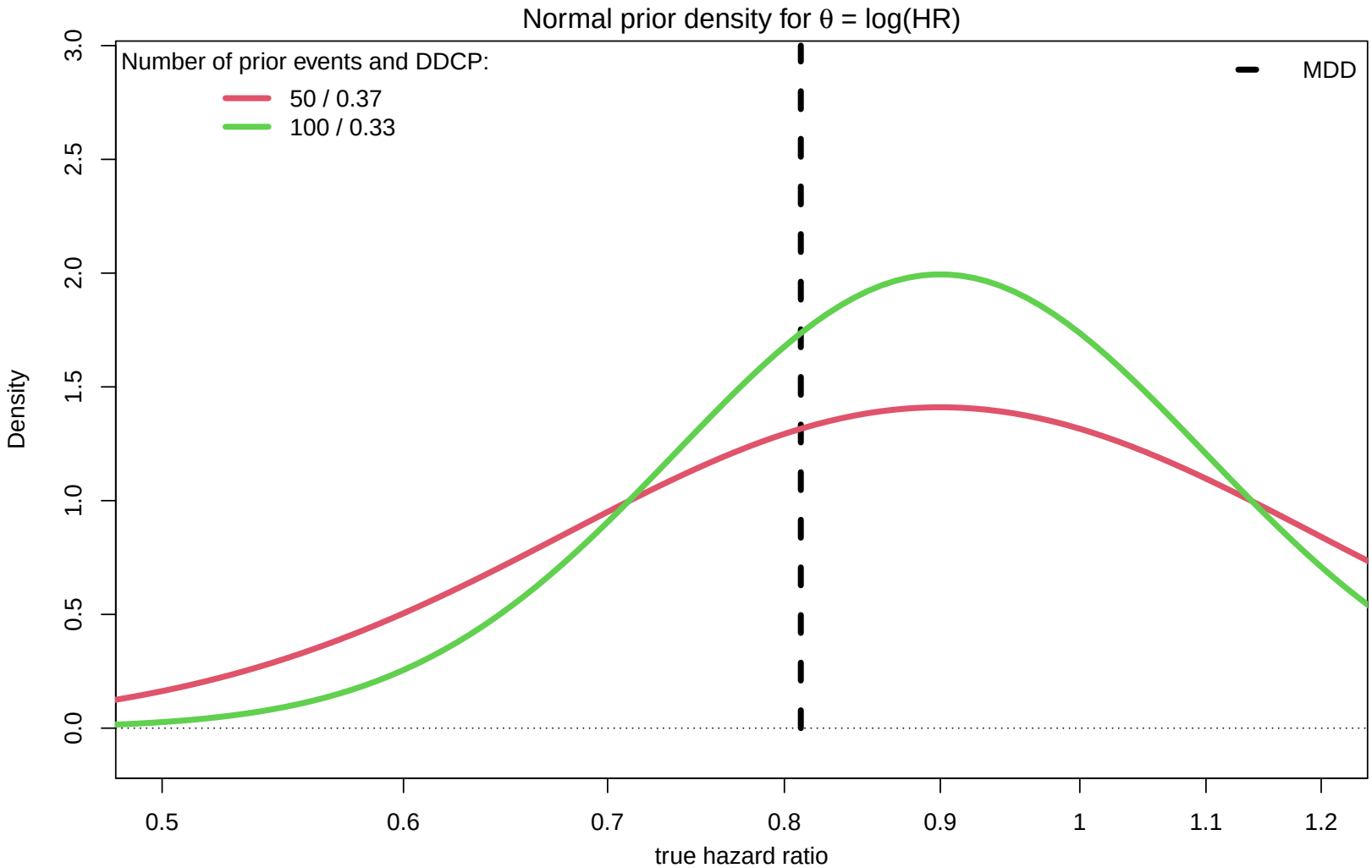


Figure: Prior distribution given **100 Ph2 events** but the **same, less favorable Ph2 outcome**. The resulting **DDCP=0.33**.

Take home messages

Take home messages

- **PTS** is the "probability of transitioning to the next phase".
  - Uses: valuations, gating, etc.
- **DDCP** is the "probability of a Ph3 study achieving its efficacy target given prior data".
  - Ph3 study design
  - Efficacy success criteria
  - Prior data
    - Ph1/2, external
    - Collecting and synthesizing all available prior evidence is critical and this is a team effort.
- DDCP serves as an **anchor** for clinical PTS (see PTS handbook).

Take home messages

- Think carefully.
  - **DDCP $\leq$ power**. Recalibrate stakeholders!
  - Understand how **prior affects DDCP**!
  - Do not use DDCP to chose **Ph2 sample size**!
  - Spell out, explore, and communicate your **assumptions**!
- successR offers excellent resources for DDCP (training material, exercises, tutorials, case studies, code).
- DDCP may be difficult to calculate in certain situations. Prior for Ph3 treatment effect is best determined based on randomized data on the Ph3 endpoint.
  - Project Endeavor and Project Providentia offer alternative approaches.
  - Ongoing conversations on how they fit and how to improve PTS calculation.

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.

Appendix

Project Endeavor

Background:

- Initiated in 2016 to enable robust decision making on large number of Non-Hodgkin lymphoma (NHL) combo studies (InHarmony)
- Expansion beyond NHL to acute myeloid leukemia (AML, 2019) and to EDIS multiple myeloma (MM)
- Strict Go/No-Go gating evolved into recommended cPTS adjustments as a way to more flexibly adapt to differences in strategic context

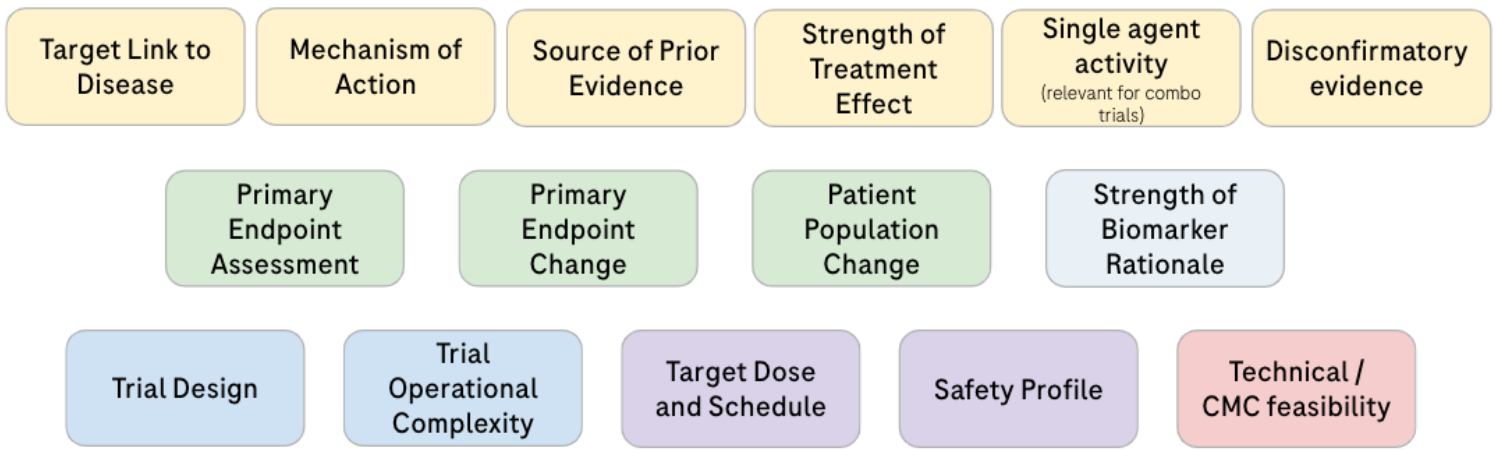
Approach:

- Current guidance is to not formally calculate DDCP for single arm trials or trials with different endpoints from the Ph3 study–which is how most of our Ph1b/2 evidence is currently generated in Heme.
- Project Endeavour uses a quantitative method for assessing cPTS in the situation where there is limited Ph1b/2 data by leveraging SuccessR with a particular construction of the prior distribution of treatment effect.
- Provides methodology for extrapolation between endpoints.

Appendix

Project Providentia

- Alternative approach/tool for calculating PTS at the transition from early stage into pivotal trials.
  - Success = study meets TPP
  - Logistic regression model **incorporating 15 risk factors simultaneously** (including anticipated treatment effect)
  - Risk factors are graded on an **ordinal scale** (e.g., suggestive, promising, compelling)
- Objective: consistency; transparency; dialogue; balance between simplicity, speed, and accuracy



Doing now what patients need next

R version and packages used to generate these slides:

R version: R version 4.2.2 (2022-10-31)

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

Other packages: bpp / mvtnorm / rpact / reporttools / xtable

This document was generated on 2023-09-11 at 10:28:38.