Assurance in drug development

Kaspar Rufibach
Methods, Collaboration, and Outreach Group, Roche Data Sciences, Basel
JSM Toronto, 9th August 2023

Acknowledgments

Markus Abt (Roche).

Hans Ulrich Burger (Roche).

Paul Jordan (Roche).

Kevin Kunzmann (Boehringer Ingelheim).

1. Define "success"!

2. Definition.

3. Do not compare to power!

4. Where are you centered at?

5. Update after not stopping at interim analysis.

6. Is the mean the right summary?

Definition

Assurance

Any endpoint type.

True effect δ.

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

What is ”success”?

Define success

δ_{suc}: can be

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

Assurance

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:

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

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

O'Hagan et al. (2001), O'Hagan et al. (2005).

Assurance vs. Bayesian predictive power

- Success: $\widehat{\delta}_{\text{final}} \leq \delta_{\text{suc}}$.
- δ_{mdd}: minimally detectable difference.
- δ_{MCID}: minimally clinically interesting difference. Make sure $\delta_{\text{MCID}} \approx \delta_{\text{mdd}}$.

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

- Assurance: significance $\Rightarrow$ irrelevant effects + type I errors are "success".
- Bayesian predictive power (BPP): relevant effects only, Spiegelhalter et al. (1986).
- Often, BPP(δ_{mdd}) $\approx$ assurance(δ_{mdd}), see Kunzmann et al. (2021).

Illustration: Time-to-event endpoint

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

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

- $\theta = \log(\text{HR})$: **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:

- Prior based on 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: $\theta_{\text{suc}} = \theta_{\text{mdd}} = \log(0.818)$.

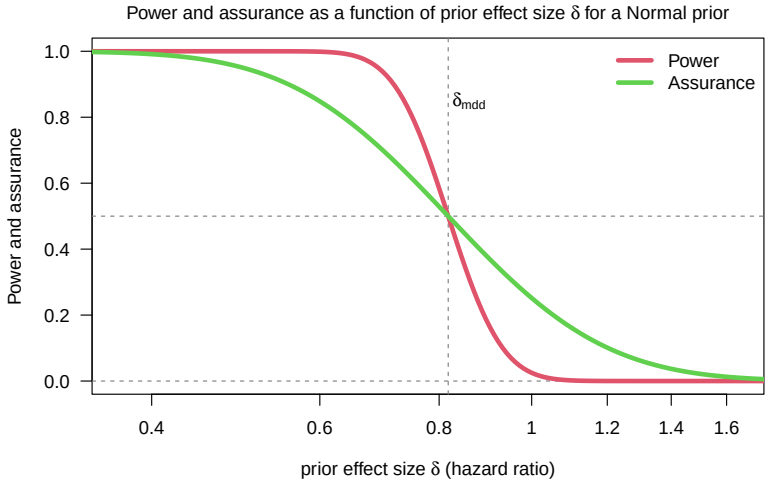
Assurance at start of Phase 3, assuming we know Phase 2 result:

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

Do not compare assurance to power!

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

Assurance smaller than Power if power ≥ 0.5 for commonly used priors.



Assurance is smaller than power

Normal prior: show using explicit formulas.

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

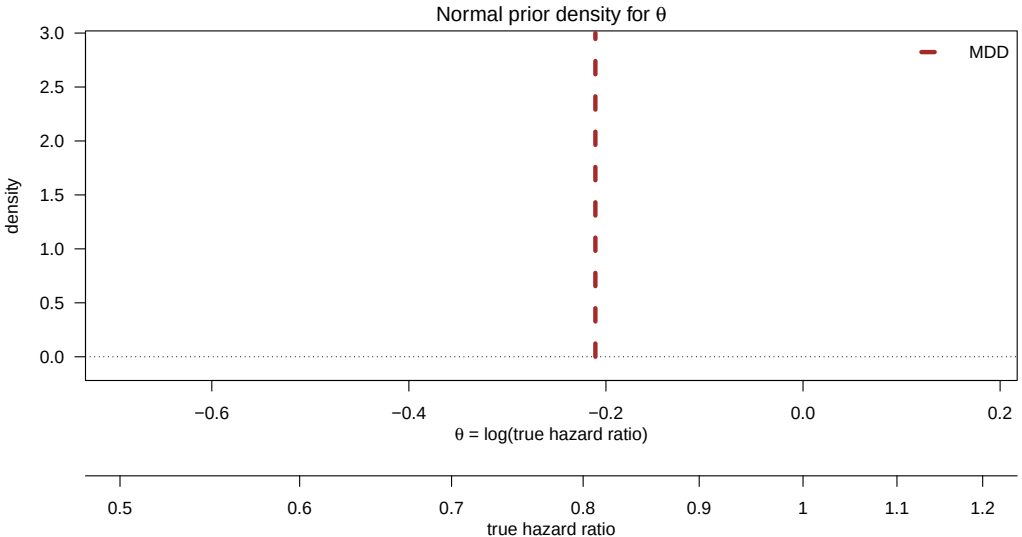
Dallow and Fina (2011).:

> **4.3. Observation 3: Irrespective of the magnitude of the final sample size, predictive power may not reach desired level (e.g. 80 or 90%)**

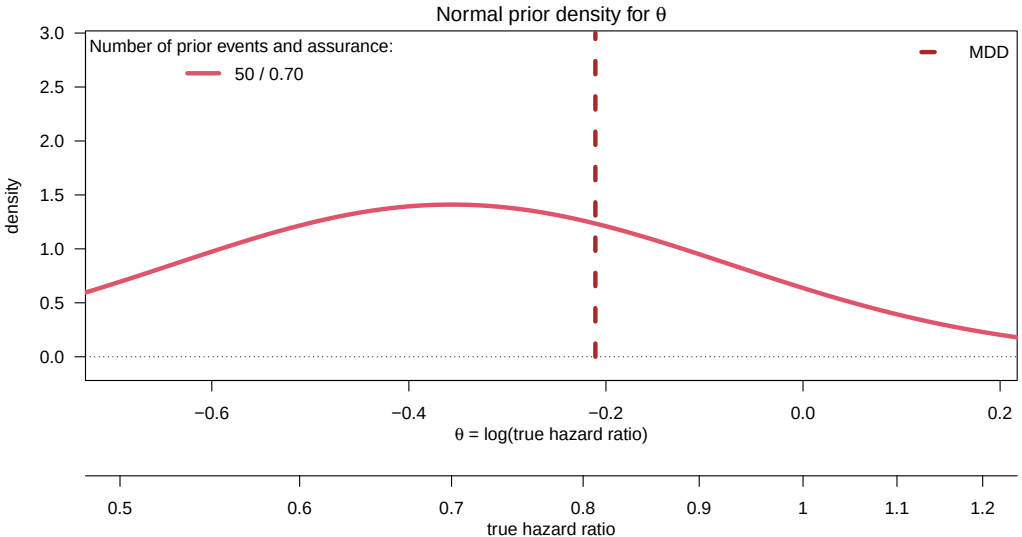
Make sure you calibrate decision-makers!

Where are you centered at?

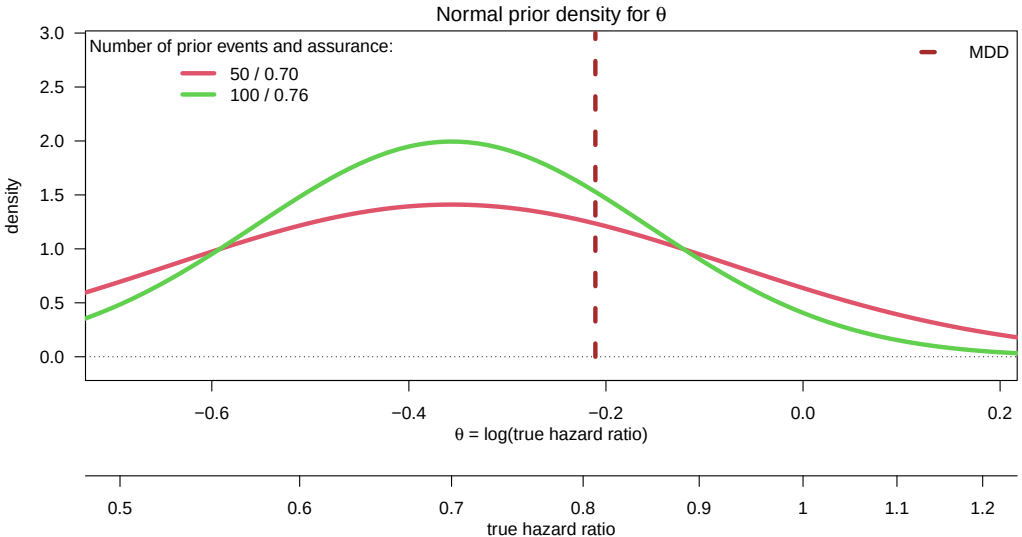
Prior sample size and assurance



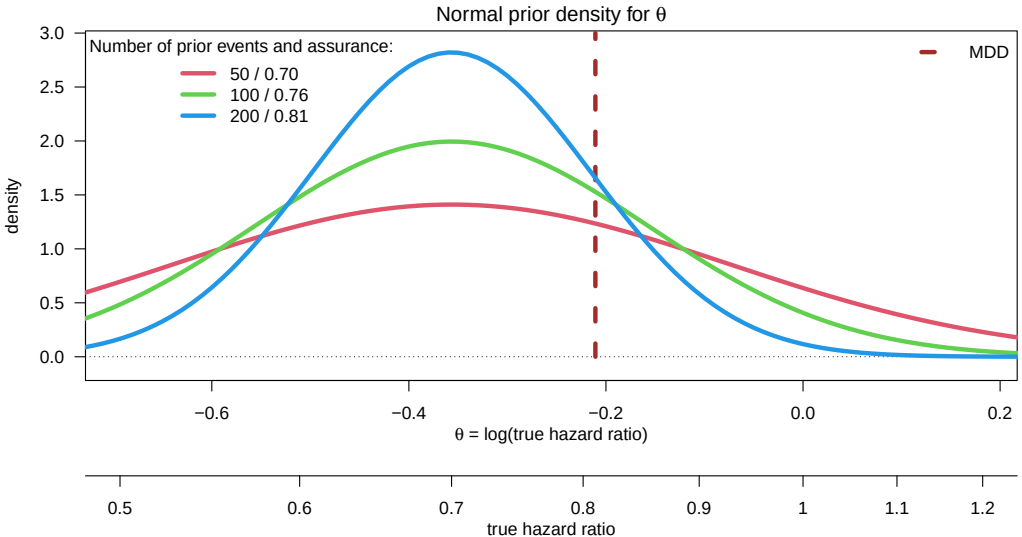
Prior sample size and assurance



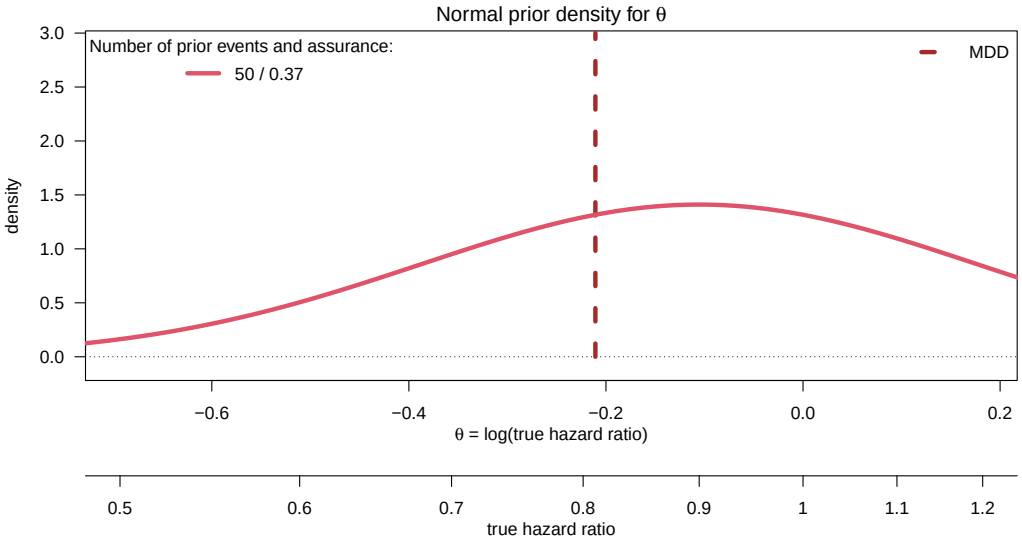
Prior sample size and assurance



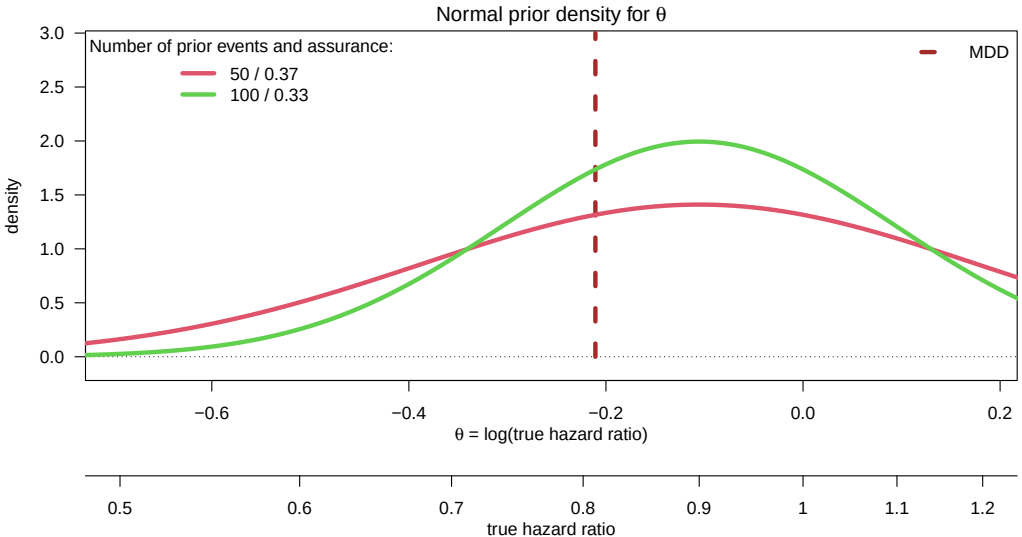
Prior sample size and assurance



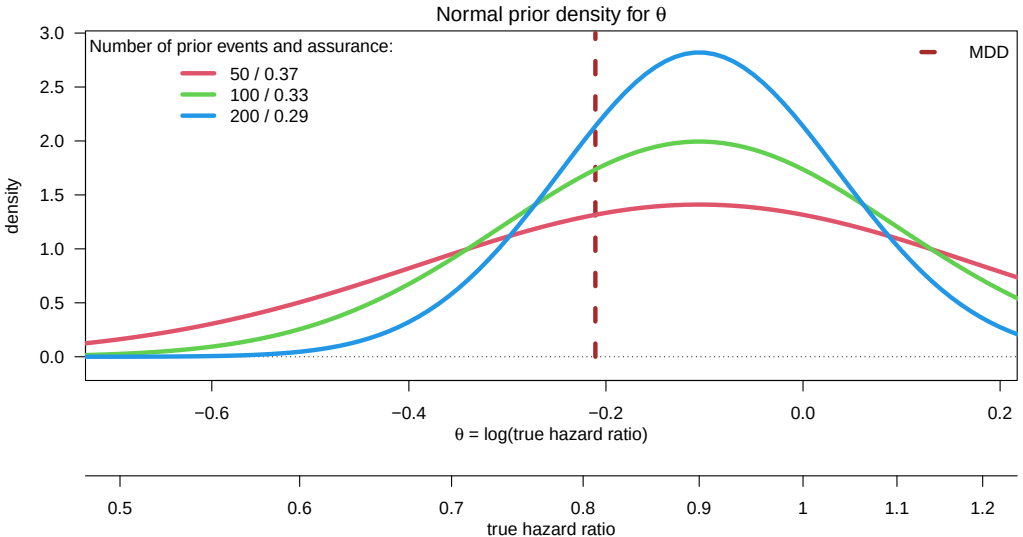
Prior sample size and assurance



Prior sample size and assurance



Prior sample size and assurance



If prior mean closer to H_0 than θ_{suc}
$\Rightarrow$ decreasing prior sample size increases assurance!

Be very careful using assurance
to chose P2 sample size!

Update assurance after not stopping at interim analysis

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

Rufibach et al. (2016b).

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

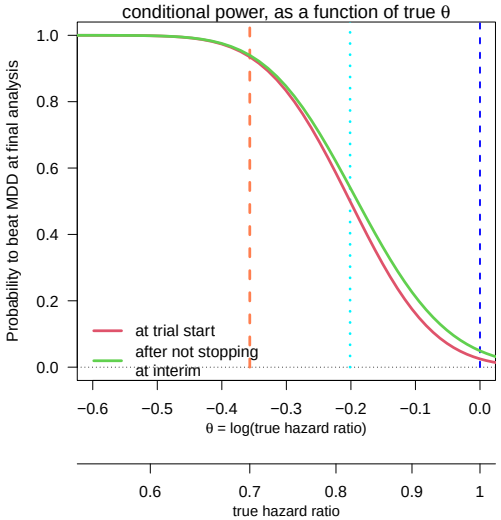
Futility interim analysis only

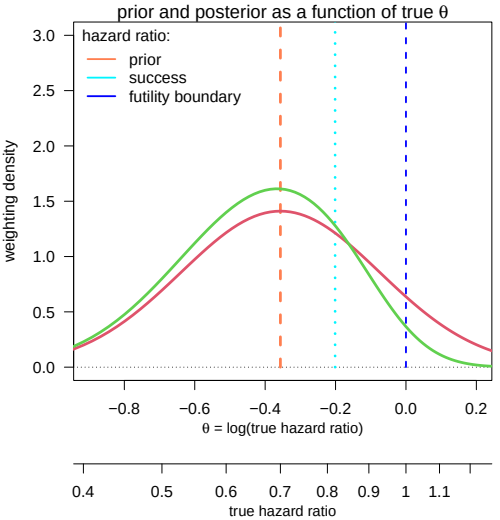
Blinded futility interim passed with boundary HR ≤ 1: we know that

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

How does assurance change after this interim?

Futility interim analysis: factors in assurance formula





Green density **not** a Normal density.

Futility interim analysis only - comments

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

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

- Prior with mean log(0.7) assigns weight to hazard ratios smaller than $\theta_{\text{suc}} = \log(0.818)$.
- **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 assurance** after not stopping.

Does assurance decrease or increase after not stopping?

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

Trial does not stop at **efficacy interim** $\Rightarrow$ assurance 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}.

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

Choice of prior

So far **Normal prior**.

Flat prior often associated with **non-informativeness**.

Not necessarily the case for assurance!

See Rufibach et al. (2016a) for details.

What is the problem?

Recall definitions and example

Power function:

$$T(\theta) \quad := \quad P_\theta(\widehat{\theta}_{\text{final}} \leq \theta_{\text{suc}}) \;=\; \Phi\Big(\frac{\theta_{\text{suc}} - \theta}{\sigma_{\text{final}}}\Big).$$

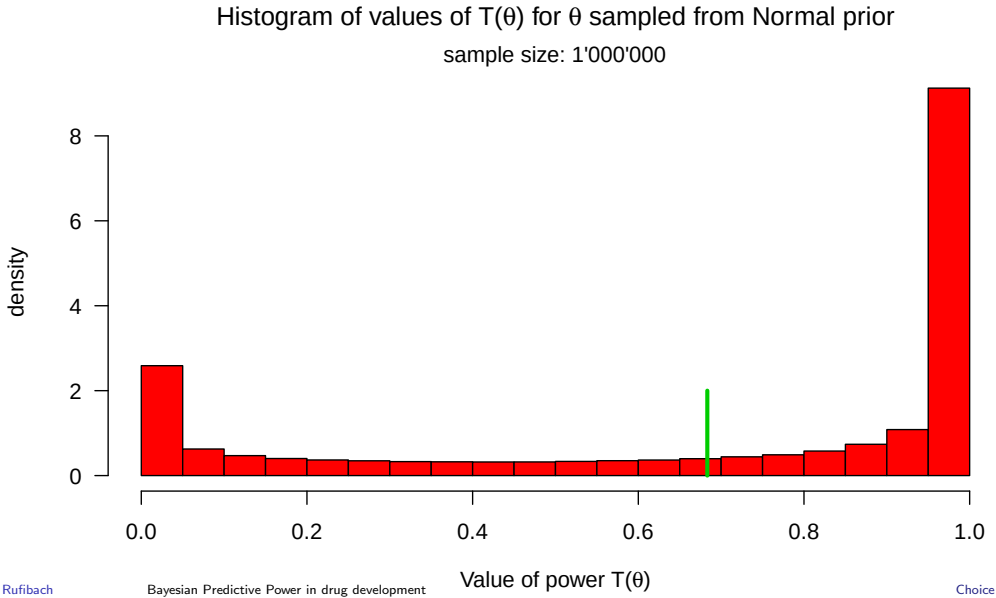
Assurance is averaged power:

$$\text{ASS} \;=\; \mathbb{E}_\theta\, T(\theta) \;=\; \int_{-\infty}^{\infty} P_\theta(\widehat{\theta}_{\text{final}} \leq \theta_{\text{suc}}) q_{\text{prior}}(\theta) d\theta.$$

Via simulation (law of large numbers):

- Draw a sample $(\widehat{\theta}_1, \ldots, \widehat{\theta}_M)$ from prior.
- Compute $T(\widehat{\theta}_1), \ldots, T(\widehat{\theta}_M)$.
- Assurance = average over these values.

Simulate assurance in example



Density of power $T(\Theta)$

Assume prior r.v. Θ with PDF q, CDF Q, and define $Y := T(\Theta)$ with PDF g, CDF G.

Use **transformation theorem** and **rule about derivative of an inverse** to get:

$$\begin{aligned} G(y) &= 1 - Q(\theta_{\text{suc}} - \sigma_{\text{final}} z), \\ g(y) &= q(\theta_{\text{suc}} - \sigma_{\text{final}} z) \frac{\sigma_{\text{final}}}{\phi(z)} \end{aligned}$$

with $z := \Phi^{-1}(y)$ and ϕ the standard Normal density function.

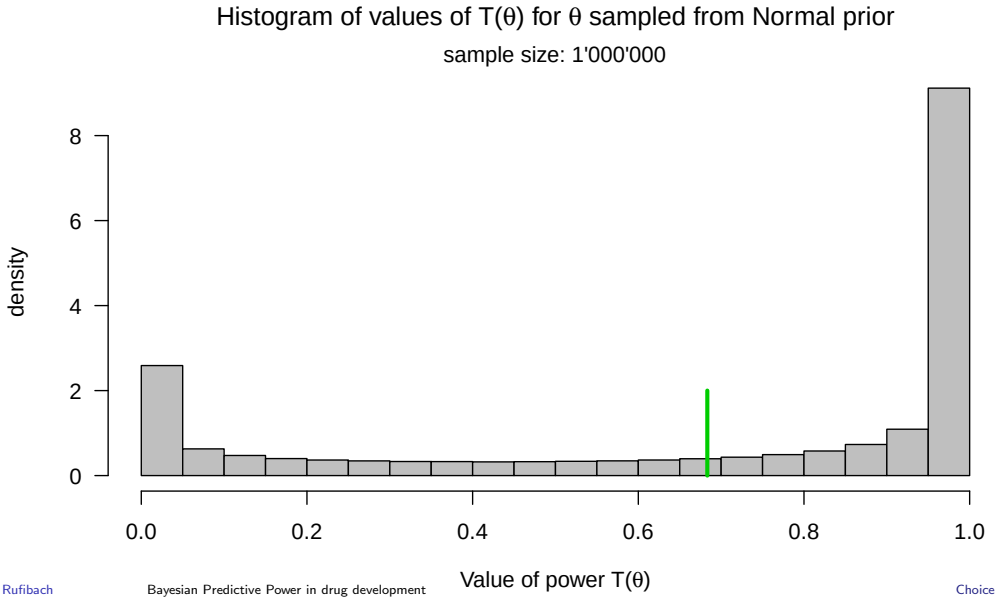
For Normal prior $\Theta \sim N(\theta_0, \sigma_0^2)$:

$$\begin{aligned} G(y) &= 1 - \Phi(\beta - \alpha z), \\ g(y) &= \alpha \phi(\beta - \alpha z) \Big[\phi(z)\Big]^{-1}, \end{aligned}$$

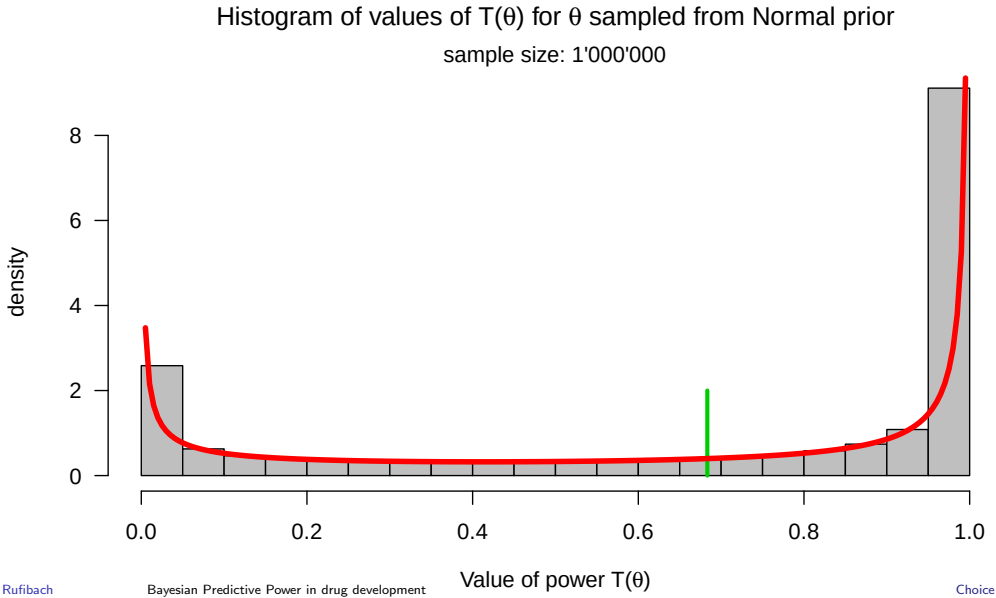
with

$$\begin{aligned} \alpha &= \sigma_{\text{final}}/\sigma_0 > 0, \\ \beta &= (\theta_{\text{suc}} - \theta_0)/\sigma_0. \end{aligned}$$

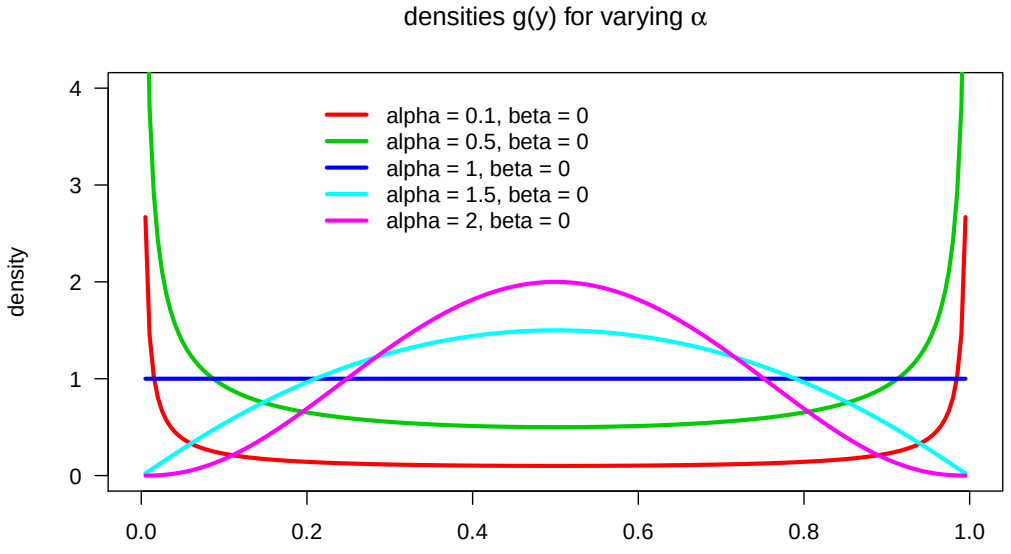
Simulate assurance in example



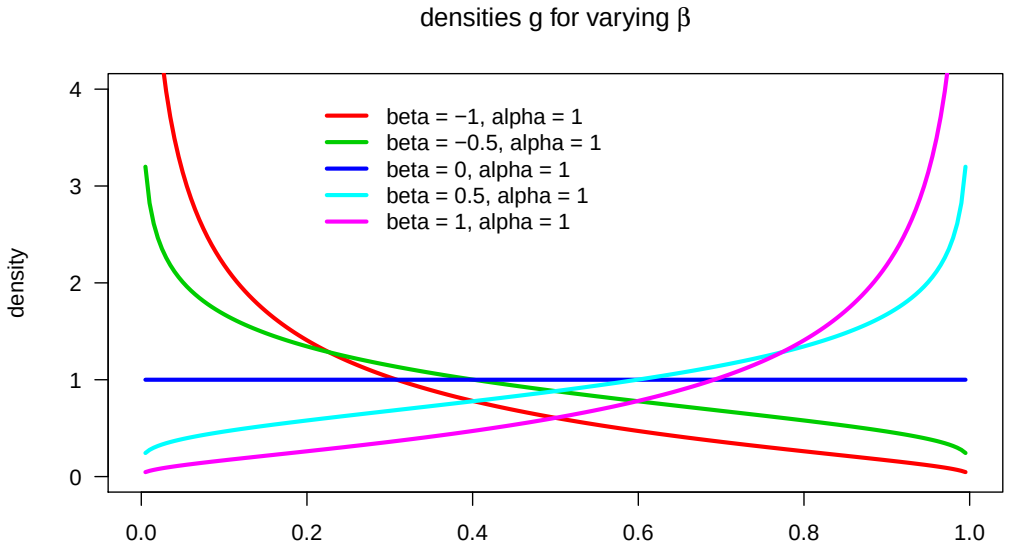
Simulate assurance in example



Density g as a function of α, for $\beta = 0$



Density g as a function of β, for $\alpha = 1$



Qualitative features of g

Theorem (Qualitative features of g)

We have the following statements:

1. *If* $\alpha = 1$, *then* g *is*

$$\begin{cases} \text{strictly decreasing for} & \beta < 0, \\ \text{constant for} & \beta = 0, \\ \text{strictly increasing for} & \beta > 0. \end{cases}$$

   on $[0, 1]$. *Minima and maxima of* g *are accordingly either at 0 or 1.*

2. *If* $\alpha \neq 1$ *then* g

$$\begin{cases} \text{has a minimum at } y_m \text{ if} & \alpha < 1, \\ \text{has a maximum at } y_m \text{ if} & \alpha > 1, \end{cases}$$

   for $y_m = \Phi(\alpha\beta/(\alpha^2 - 1))$. *Furthermore,* g

$$\begin{cases} \text{is decreasing for } y < y_m \text{ and increasing for } y > y_m \text{ if} & \alpha < 1, \\ \text{is increasing for } y < y_m \text{ and decreasing for } y > y_m \text{ if} & \alpha > 1. \end{cases}$$

Proof: Compute g', g'', discuss these.

Why? And what does it mean?

Simplest case: $\alpha = 1, \beta = 0 \Rightarrow d_{\text{prior}} = d_{\text{final}}, \theta_0 = \theta_{\text{suc}} \Rightarrow g$ uniform.

Prior and distribution of pivotal effect size have same variance $\Rightarrow$ power becomes uniform, either you beat θ_{suc} with $\widehat{\theta}_{\text{final}}$ or not, with equal probability.

Why P(extreme assurance values) so high if $\alpha < 1$? $d_0 < d_{\text{final}} \Rightarrow$ high variance of prior $\Rightarrow$ high probability to have extreme HRs $\Rightarrow$ power for these is either almost 0 or 1.

g unimodal if $\alpha > 1 \Rightarrow \sigma_{\text{final}} > \sigma_0 \Rightarrow d_{\text{final}} < d_0 \Rightarrow$ prior number of events **larger** than Phase 3 events.

Unrealistic in clinical development.

Priors explored

How should we choose prior to get **unimodal** power distribution?

Explored priors:

- truncated Normal,
- Uniform,
- Uniform prior with Normal tails.

None of them provides a unimodal density of power values under realistic assumptions.

Prior potentially informs assurance substantially.

Rufibach et al. (2016a).

1. Define "success"!

2. Definition.

3. Do not compare to power!

4. Where are you centered at?

5. Update after not stopping at interim analysis.

6. Is the mean the right summary?

Discussion

- Be clear about **definitions**.
- Assurance $\neq$ power $\Rightarrow$ **recalibrate** stakeholders.
- Update assurance **after not stopping at interim analysis**. Extension to >1 interims straightforward.
- Density of power values **bathtub-shaped** for typical development scenario.
  - **Sensible** to summarize this distribution in one number which we call assurance?
  - Prior with large variance **not necessarily** uninformative!
- R package bpp on CRAN: Rufibach et al. (2022).

Thank you for your attention.

kaspar.rufibach@roche.com

http://www.kasparrufibach.ch

numbersman77

References I

- Dallow, N. and Fina, P. (2011). The perils with the misuse of predictive power. *Pharm. Stat.* **10** 311–317.
- 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.
- Rufibach, K., Burger, H. and Abt, M. (2016a). 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. (2016b). 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. (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.

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 / bpp / mvtnorm / reporttools / xtable

This document was generated on 2023-08-10 at 00:22:14.