Breaking Boundaries in Drug Development: How large should the Type I Error be?

Carl-Fredrik Burman[1,2], Sebastian Jobjörnsson[2]

Advanced Analytics Centre, AstraZeneca R&D[1]
Applied Mathematics &Statistics, Chalmers University[2]

PSI Annual Conference, Amsterdam, 3-6 June 2018

Acknowledgments

We wish to thank professors Sören Christensen (Hamburg), Stephen Senn (Luxembourg) and Chris Jennison (Bath) for contributions.

This work has received funding from the European Union Seventh Framework Programme [FP7 2007-2013] under grant agreement no. Health-F5-602552 (IDEAL project).

What is α?

General understanding

For confirmatory clinical trials,
regulatory authorities allow a
Type I Error of 5% two-sided, that is,
$\alpha = 2.5\%$ *one-sided.*

Regulatory guidance

EMA guidance (Multiplicity, 2016, EMA/CHMP/44762/2017):

- (Motivating that multiplicity is an issue)
  *"**if** statistical tests are performed on five subgroups, independently of each other and each at **a significance level of 2.5%** (one-sided directional hypotheses), the chance of finding at least one false positive statistically significant test increases to approximately 12%"*
- (About co-primary endpoints):
  *each null hypothesis on every primary variable has to be rejected at the same **significance level (e.g. 0.05)***

Regulatory guidance cont'd

FDA guidance (Clinical Evidence of Efficacy . . . , 1998):

- *"Even if all drugs tested in such trials were ineffective, one would expect **one in forty** of those trials to "demonstrate" efficacy by chance alone at **conventional levels**"*

ICH E9 (Statistical Principles, CPMP/ICH/363/96):

- *"**Conventionally** the probability of **type I error** is set at **5% or less** or as dictated by any adjustments made necessary for multiplicity considerations; the precise choice may be influenced by the prior plausibility of the hypothesis under test and the desired impact of the results."*

Summary

- $\alpha = 2.5\%$ (one-sided) is almost always used.
- This is purely conventional.
- Not a regulatory requirement!
- ICH says that α *"may be influenced by"*
  - *"prior plausibility"*
  - *"desired impact"*
- However, I haven't found any concrete rule for how to choose α (beyond the convention).

Let's try to find the optimal α!

For 2-trials at $\alpha = 0.025$, resulting Type I Error is $\alpha^2 = 0.000\,625$

Model

1. Objective: Maximise expected response in N patients (current & future).
2. Each patient may receive treatment A or B.
3. We may compare A and B in an equally randomised clinical trial (RCT).
4. After stopping RCT, choose one treatment for all remaining patients.

Model

1. Objective: Maximise expected response in N patients (current & future).
2. Each patient may receive treatment A or B.
3. We may compare A and B in an equally randomised clinical trial (RCT).
4. After stopping RCT, choose one treatment for all remaining patients.

Question to the audience:

How should we approach this from a **frequentist** perspecive?

- OK with a tiny experiment with $p = 2\%$ when the impact of the treatment choice is enormous?
- Randomise all patients with an ultra-rare disease and never get a conclusion?
- Is there a better frequentist approach?

"Give me a point of support, and I will move the Earth"

We need some kind of further "starting" assumption.
Bayesian and frequentist approaches are conceptually different.

Bayes
Assume a prior for the efficacy difference d
Bayes fits well for decision problems.

"Give me a point of support, and I will move the Earth"

We need some kind of further "starting" assumption.
Bayesian and frequentist approaches are conceptually different.

Bayes
Assume a prior for the efficacy difference d
Bayes fits well for decision problems.

Frequentist
We have a symmetry in the problem: no preference between A and B.
(Not assuming that one is Active and one standard-of-care or placebo.)
"Accepting" a null hypothesis of $d = 0$ is not very helpful.
Recall Wald's sequential test of $d = d_0$ vs. $d = d_A$.
Due to symmetry, we will test $d = +d'$ vs. $d = -d'$ (with equal type I and type II errors).

Model

1. Objective: Maximise expected response in N patients (current & future)
2. Each patient may receive treatment A or B.
3. We may compare A and B in an equally randomised clinical trial (RCT).
4. After stopping RCT, choose one treatment for all remaining patients.

5 (Bayesian). Assume a prior for efficacy difference, d.
5 (Frequentist). Assume $d = \pm d'$ for some fixed $d' > 0$

Set-up

Normal approximation usually works. Making it even easier, assume that responses are i.i.d. and patient i has response

$$X_i \sim N(m + \frac{d}{2}, \sigma^2) \text{ if patient i is given treatment A}$$
$$X_i \sim N(m - \frac{d}{2}, \sigma^2) \text{ if patient i is given treatment B}$$

Define utility

$$U = \sum_{i=1}^{N} (X_i - m)$$

Taken d as fixed,

$$E[U \mid d] = \frac{d}{2} E[N_A - N_B]$$

where N_A and N_B are the number of patients that we give treatment A and B, respectively. Thus, we are to optimise the expected number of patients receiving the best treatment.

Initially, we randomise patients 1:1 to A and B.
After some number, ν, of patients,
we choose A or B for the remaining $N - \nu$ patients.
ν depends on previous observations, and is therefore a stopping rule.
Given d, the RCT stopping time ν, and choosing treatment A, we get

$$N_A - N_B = (\frac{\nu}{2} - \frac{\nu}{2}) + ((N - \nu) - 0) = N - \nu.$$

Thus, the expected utility can be written as

$$E[U \mid d] = \frac{|d|}{2}(N - \nu)\Big(P(\text{Correct decision}) - P(\text{Incorrect decision})\Big).$$

A clear trade-off between fast decision (to increase number, $N - \nu$ post-RCT) and correct decision (where larger RCT gives larger security).

A frequentist approach

Test $d = +d'$ vs. $d = -d'$.
That is, we optimise the utility for a fixed value of $|d|$.
This doesn't mean we "believe" this value is true.

Let $S'_n = \sum_{i=1}^{n} T_i X_i$ be the cumulative sum of response differences between pairs of patients receiving treatment A and B.
We have $S'_n = \hat{d}_n \frac{n}{2}$.

Due to the symmetry of the problem, we consider symmetric stopping boundaries,
$\{b_n \geq 0\}_{n \leq N}$ and stop the RCT whenever $|S'_n| > b_n$.
The optimal boundary b can be found by backward induction.

High-level results (Frequentist)

The value of d', for which the procedure is optimised, is very important for when to stop.

- Locally, with $d' \searrow 0$, the boundary is really conservative in the beginning.
- When d' is extremely large, the RCT will typically stop almost directly.
- Intermediate values for d' gives more reasonable stopping rules, but the choice of d' is an important (subjective) input.
- Drawback: The method sticks to the pre-specified d' even if it is inconsistent with accumulating data.
- This frequentist approach coincides with a Bayesian one with equal-mass 2-point prior on $\pm d'$, although the interpretations may differ.

Simplifying asymptotics

Different solutions for every possible N. For reasonably large N,

$$S''_{(n/N)} = \frac{S'_n}{\sigma\sqrt{N}}$$

can be approximated by a Wiener process S_r in continues time, $r \in [0, 1]$. We have

$$S_r \sim N(r \cdot \delta, r)$$

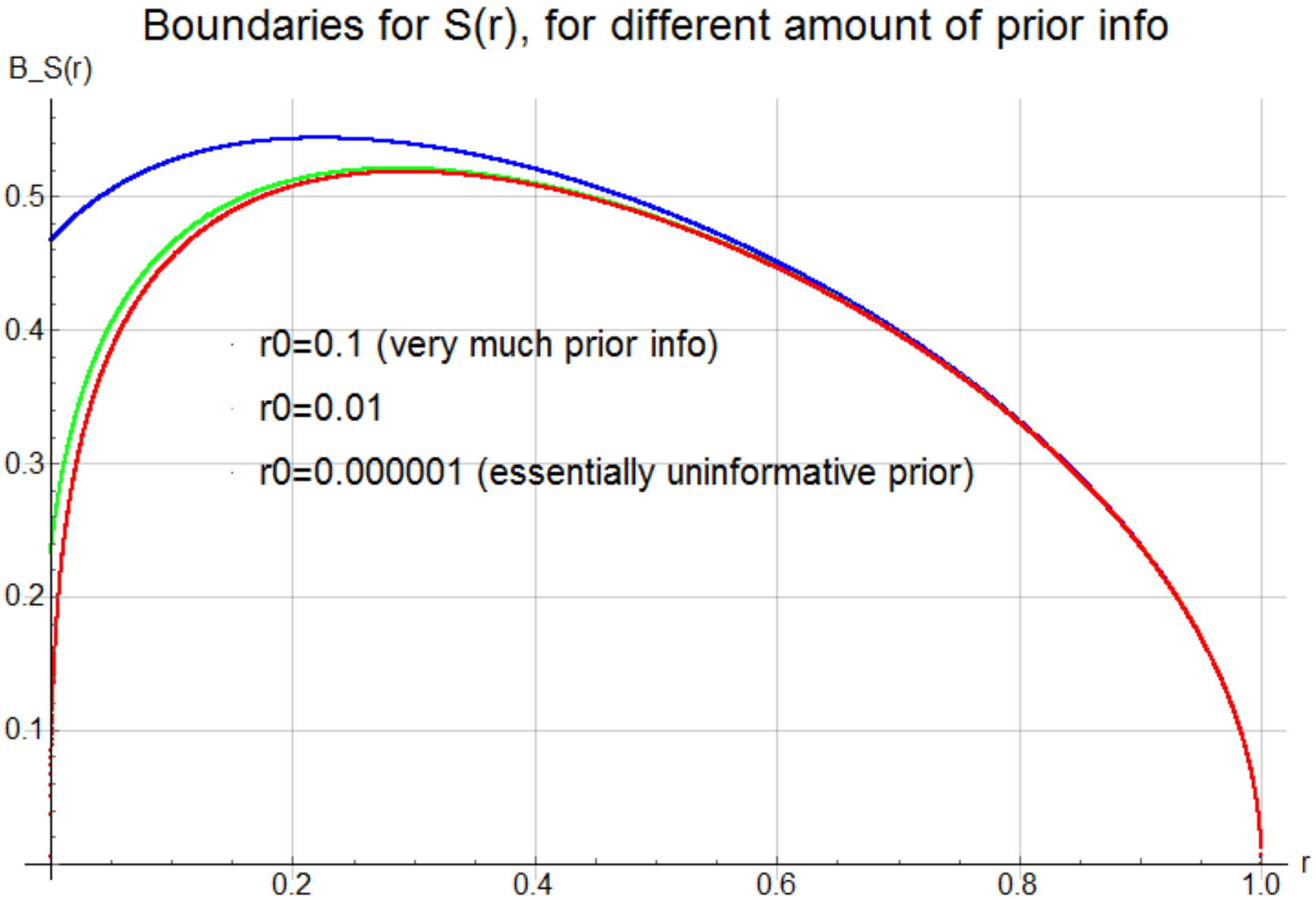
where $\delta = \frac{d}{2}\sqrt{N}/\sigma$ and r is the fraction of patients included in the RCT. The parameter δ is the expected Z score if all N patients are included in the RCT.

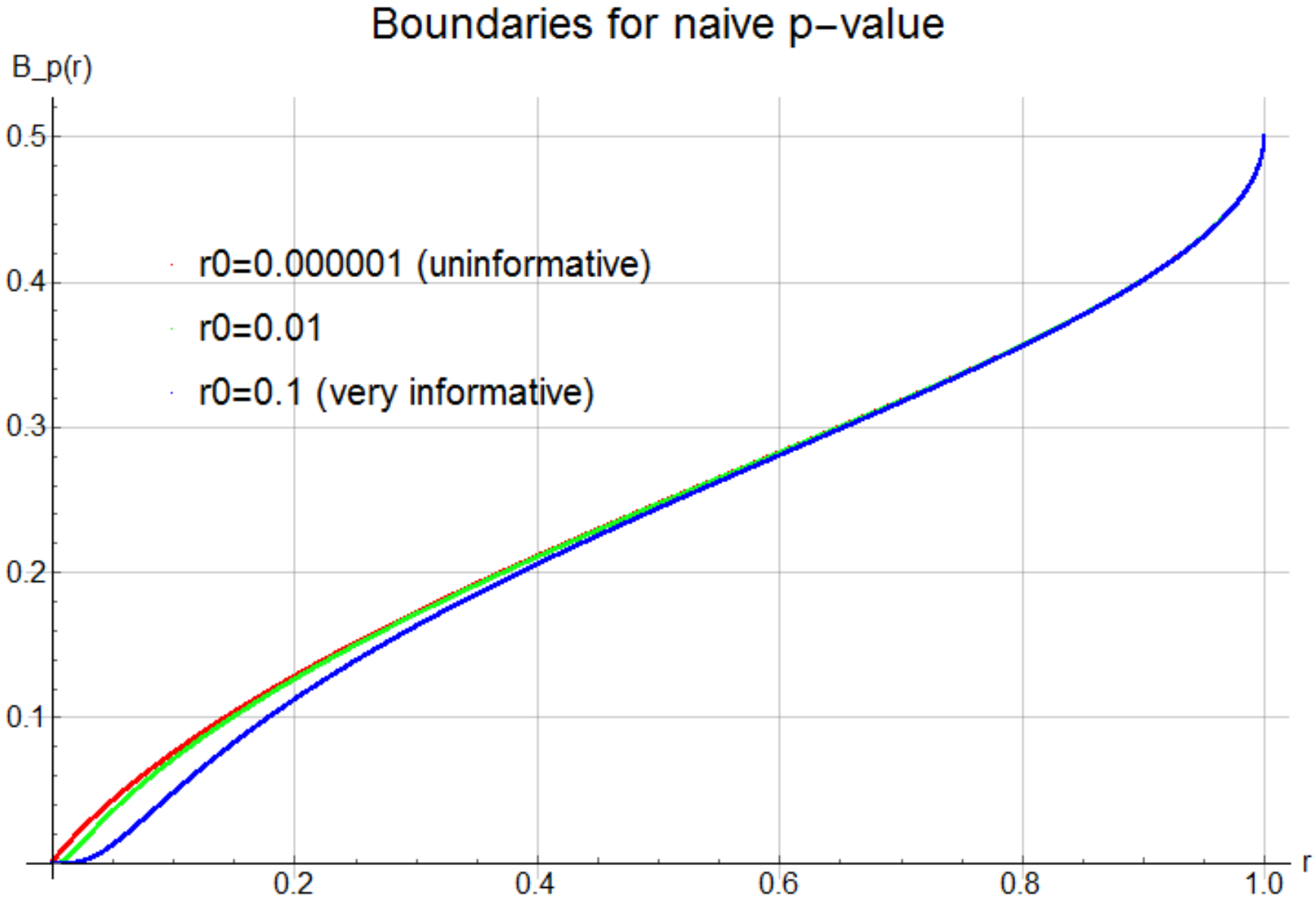
Bayesian approach

Assume a prior, π for δ. A normal prior corresponds to having observed n_0 out of N patients. The parameter r_0 indicates the relative information a priori. We take

$$\pi = N(m_0, 1/r_0).$$

It is sufficient to calculate the optimal boundary $B(r)$ for a non-informative prior (formally $N(0, \infty)$).
The optimal boundaries for all other normal priors are then simple transformations of $B(r)$.


Boundaries for S(r), for different amount of prior info
B_S(r)
0.5
0.4
0.3
0.2
0.1
r0=0.1 (very much prior info)
r0=0.01
r0=0.000001 (essentially uninformative prior)
0.2
0.4
0.6
0.8
1.0
r


Boundaries for naive p–value
B_p(r)
0.5
0.4
0.3
0.2
0.1
r0=0.000001 (uninformative)
r0=0.01
r0=0.1 (very informative)
0.2
0.4
0.6
0.8
1.0
r

Which parameters are of interest?

How large proportion, r, of all patients receiving a treatment was involved in clinical trials of that treatment?
(For continuous treatments, perhaps rather the proportion of patient-years in clinical trials.)

Typical values???

- $r \sim 0.1$ for ultra-rare disease
- $r \sim 0.01$ for rare disease
- $r \sim 0.001$ for common disease
- $r \sim 0.0001$ for very common disease

Very, very rough! What do you think?

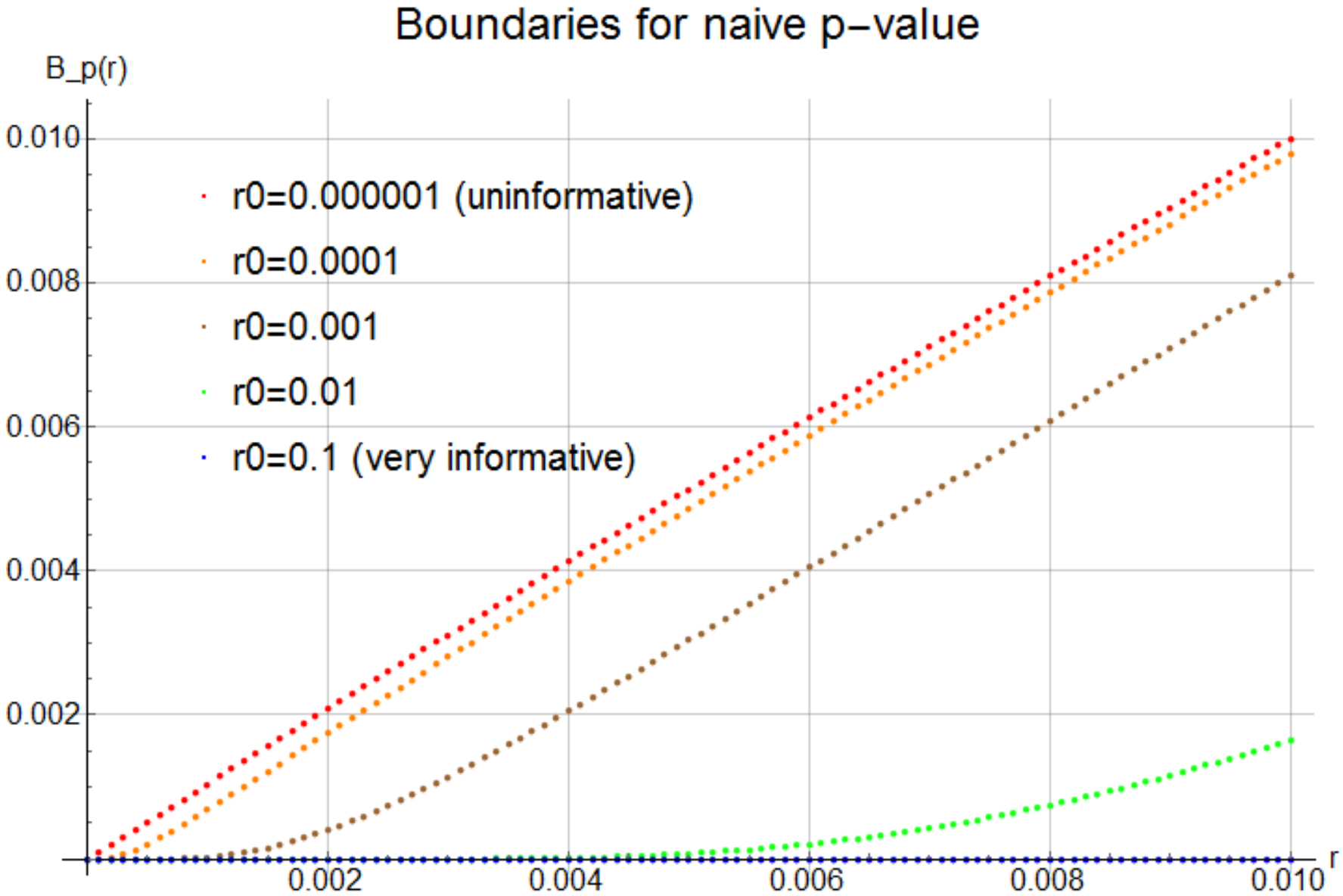
δ is the expected Z-score if all N patients were in the trial.
The expected Z-score for a smaller trial is $\delta\sqrt{r}$.

Say that traditional trials are designed to have an expected Z-score of 3, $E[Z] = 3$, for a certain d' to get around 85% power at conventional alpha.

In many cases, it is reasonable to have $r_0 << r$. A prior with info r_0 gives prior SD$[\delta] = 1/\sqrt{r_0}$.
Thus, for the prior to be consistent with $\delta = \delta'$, we should have something like $r_0 \leq 1/\delta'^2$ (assuming prior mean zero).

	r	$\delta(>)$	$r_0(<)$
"Ultra rare"	~ 0.1	~ 10	~ 0.01
"Rare"	~ 0.01	~ 30	~ 0.001
"Common"	~ 0.001	~ 100	~ 0.0001
"Very common"	~ 0.0001	~ 300	~ 0.00001

Obviously, the efficacy may be much better than indicated in the table.
However, much smaller values of δ may be difficult to detect.


Boundaries for naive p−value
B_p(r)
0.010
0.008
0.006
0.004
0.002
r0=0.000001 (uninformative)
r0=0.0001
r0=0.001
r0=0.01
r0=0.1 (very informative)
0.002
0.004
0.006
0.008
0.010
r

Implications

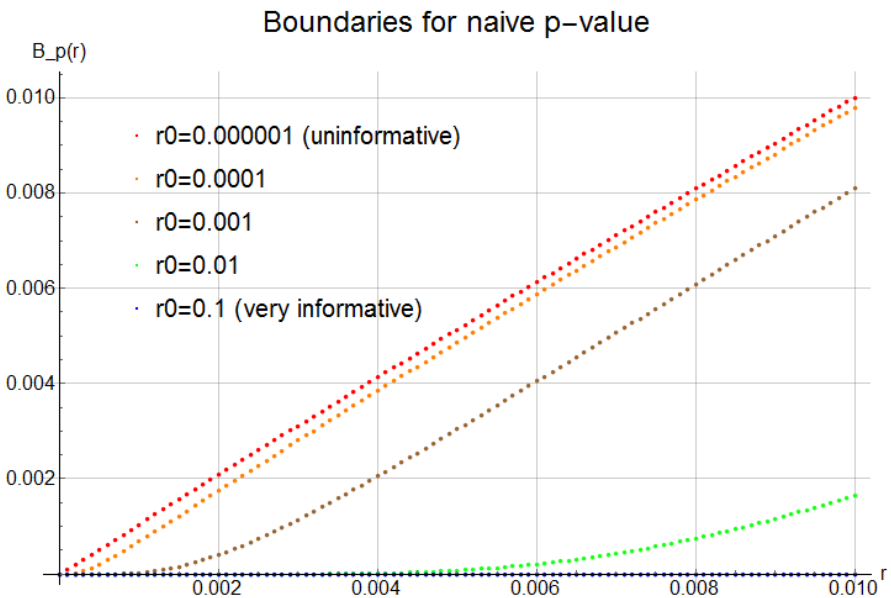
Regulatory implication

If we agree on the assumptions (maximise total efficacy in population; size of the population; (e.g.) non-informative prior; safety included; no costs), a regulatory agency should approve a new treatment whenever the p-value is less than the optimal boundary.

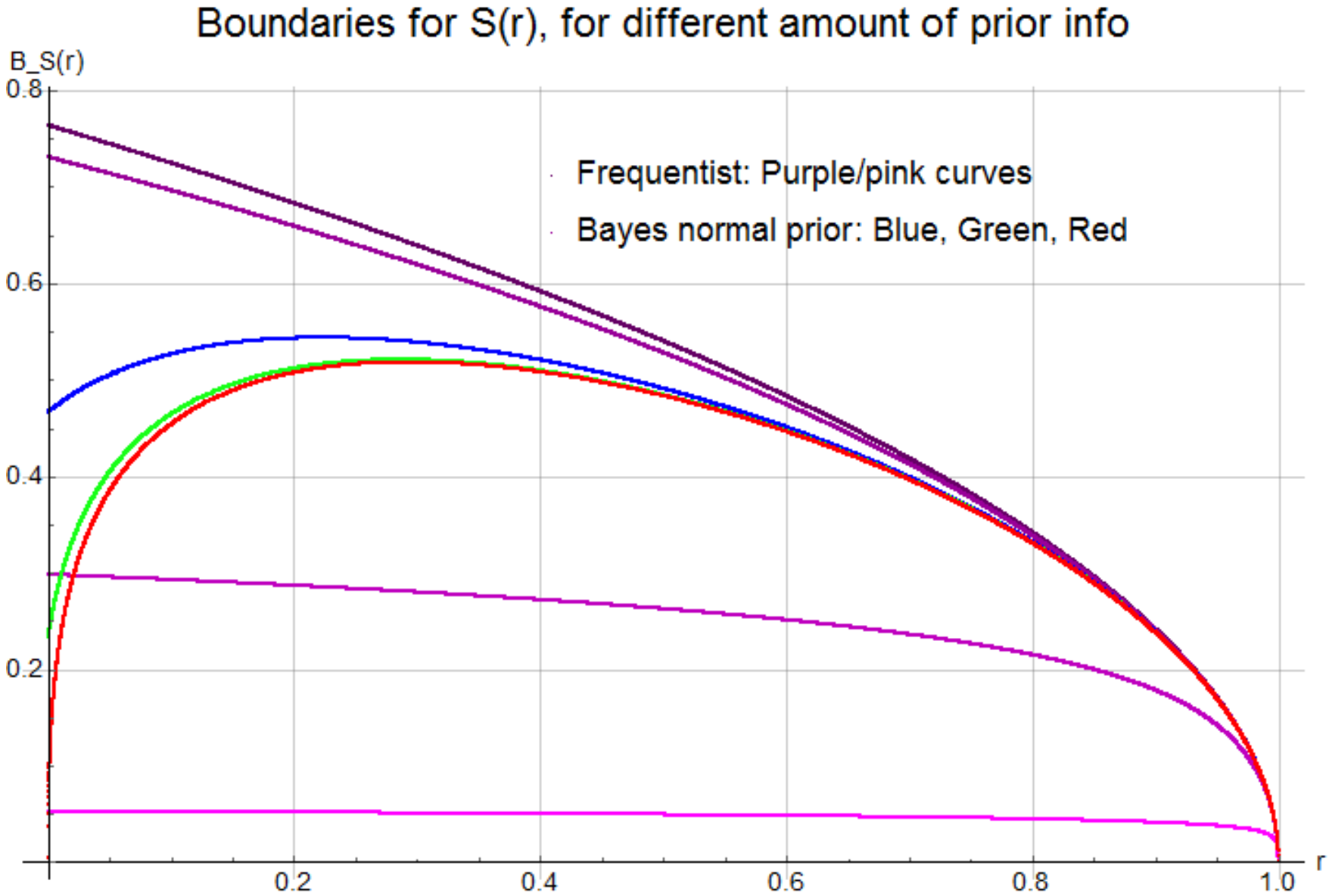
It *may* be optimal to be more liberal, approving also for somewhat larger p-values, for reasons like fully sequential trial is impossible, cost of trial, the trial only including fraction of available patients, etc. We'll discuss such aspects later on.

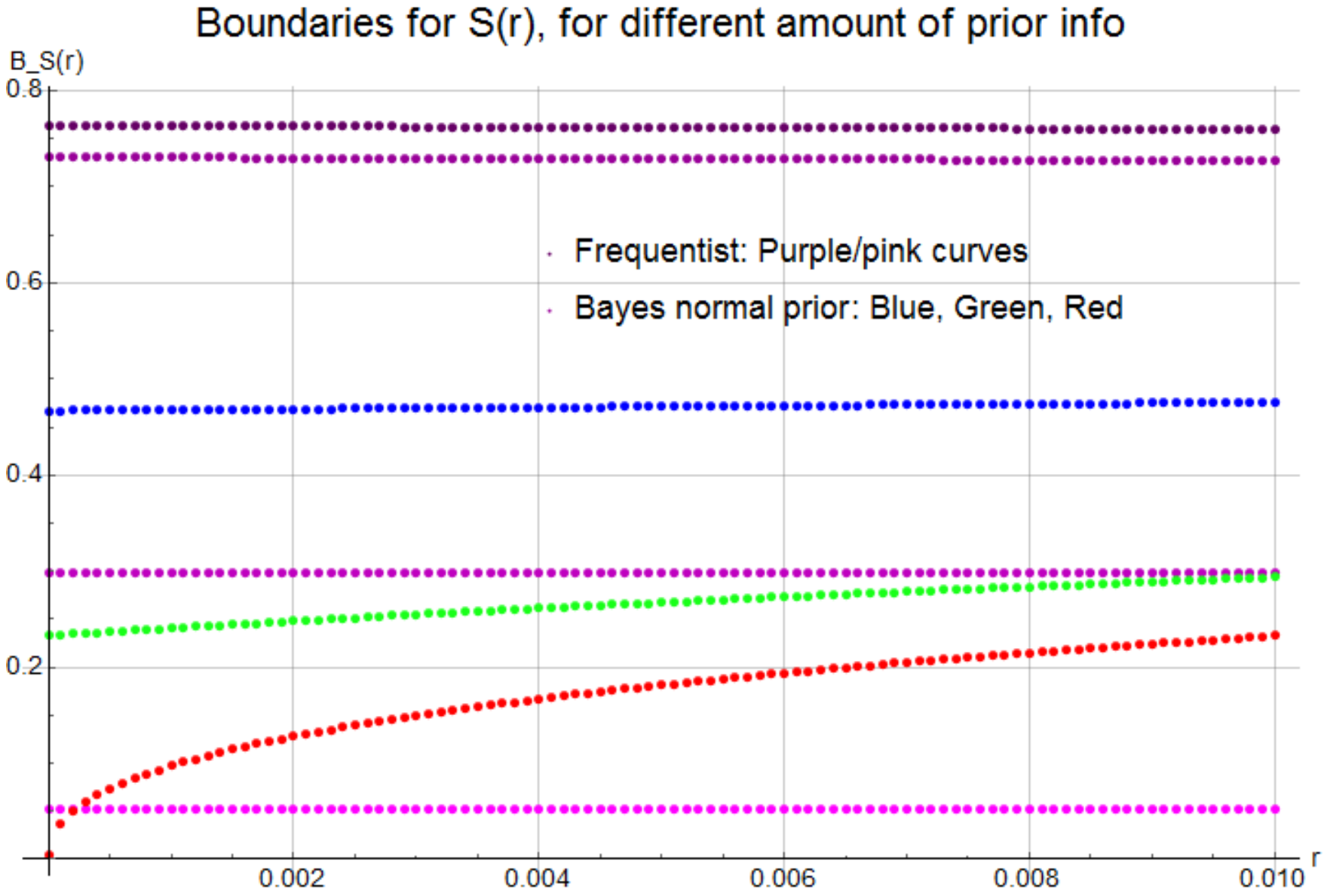
Rule of thumb

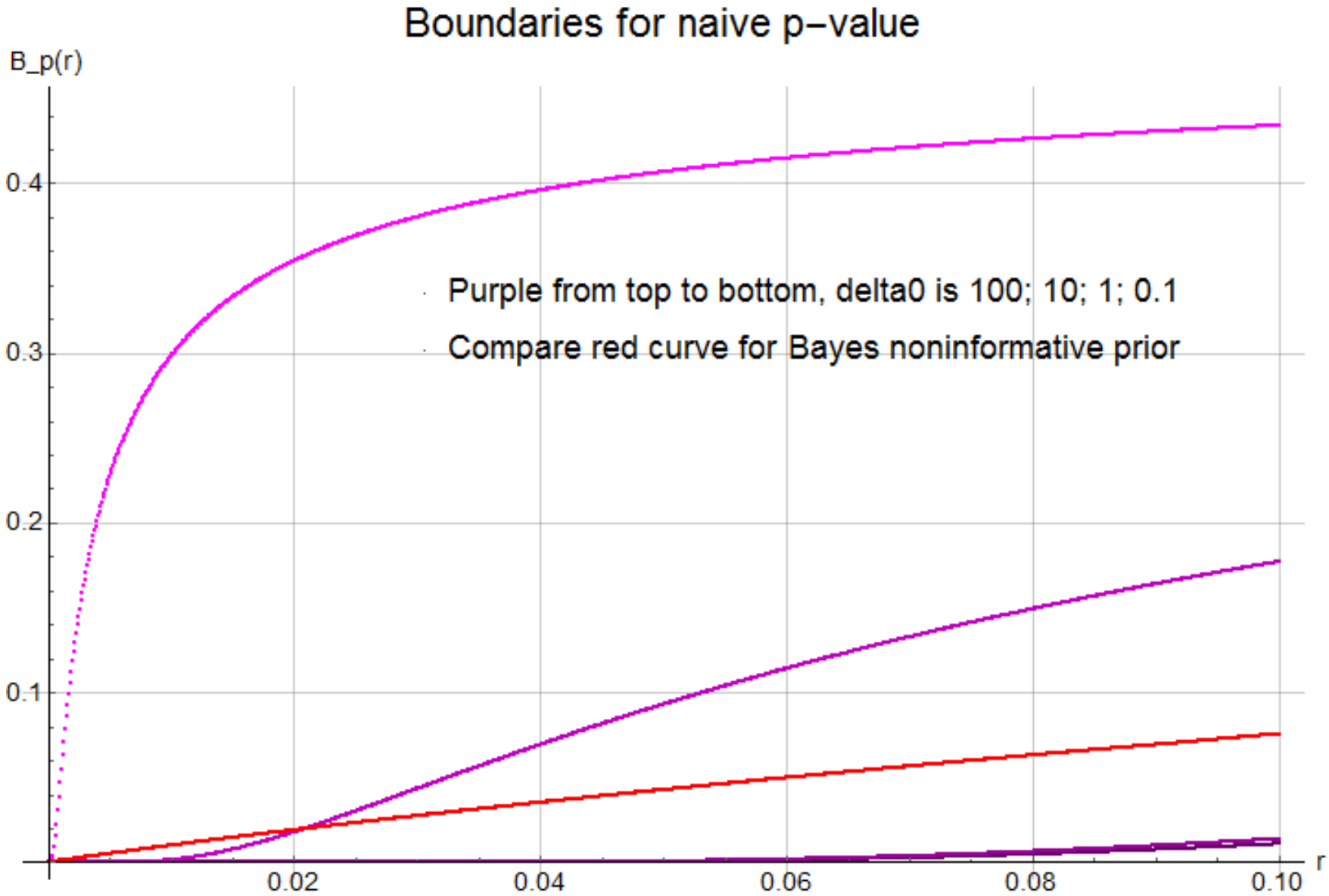
For non-informative prior, $B_p(r) \sim r$ when $r \downarrow 0$.
Approve a treatment (at least) when $p < r$.



Comment: May use a small r_0 to discourage very small trials.


Boundaries for S(r), for different amount of prior info
B_S(r)
0.8
0.6
0.4
0.2
Frequentist: Purple/pink curves
Bayes normal prior: Blue, Green, Red
0.2
0.4
0.6
0.8
1.0
r


Boundaries for S(r), for different amount of prior info
B_S(r)
0.8
0.6
0.4
0.2
0.002
0.004
0.006
0.008
0.010
r
Frequentist: Purple/pink curves
Bayes normal prior: Blue, Green, Red


Boundaries for naive p–value
B_p(r)
0.4
0.3
0.2
0.1
Purple from top to bottom, delta0 is 100; 10; 1; 0.1
Compare red curve for Bayes noninformative prior
0.02
0.04
0.06
0.08
0.10
r

Extensions

Current extensions (details not shown):

- For each patient in the RCT, q patient will receive treatment B.
  - Treatment B is standard-of-care
  - Introduces non-symmetry
  - p-value boundary increases from r to at most $2\,r$ (for non-informative prior)
- Non-fixed population size, N
  - Results depend slightly on distribution of N
  - However, expected value is often good enough
- Total Type I Error increases if there are several interim looks
  - E.g. $B_p(r) = r$ and Type I Error of 1.0% for a fixed design, gives Type I Error about 1.6% if there are four equidistant interim looks.

Conclusions

1. More(/less) evidence should be required for common(/rare) diseases.
2. In principle, α can be optimised for different concrete situations
3. Several factors to consider:
   - Choice of patient horizon, N. Optimal alpha is not robust in N.
   - Individual ethics, safe-guarding trial patients' autonomy and expected utilities.
   - How large part of the patients can be included in a RCT?
   - A fully sequential RCT is not feasible. How to adjust?
   - Cost of experimentation
   - One response variable, including efficacy, safety, convenience
   - Homogeneous population?
   - Can real-world data post-RCT sometime reverse decisions?
   - How will regulatory rules affect incentives for drug development?

References

Results and further references are given in:

Jobjörnsson S, Christensen S (2018).
Anscombe's model for sequential clinical trials revisited.
Sequential Analysis 37(1):115-144.