mODa 14
Model-Oriented Data Analysis and Optimum Design

Integrating Bayesian and promising zone approaches in a Phase 3 clinical trial

Sergei Leonov (Statistical Innovation, CSL Behring)

Drongen Abbey, Belgium

June 16, 2026

CSL Behring

Outline

- Trial background, model, design
- IA1, simulations, decision grids
- Conditional power, promising zone approach
- IA2, combined simulations

Joint work with Aparna Raychaudhuri and Vaibhav Mahajan

To appear in “*Statistical Papers*”

Clinical trial: Phase 3, rare disease

Primary efficacy endpoint: time to a predefined event within follow-up period $[0, t_F]$

- Two planned interim analyses (IA)
  1. IA1: Bayesian futility, after sufficient number of participants have completed follow-up time t_1, $t_1 < t_F$
  2. IA2: futility & sample size re-estimation (conditional power/promising zone approach)
- Bayesian futility & promising zone: each studied extensively
- Our focus: develop a unified hybrid design

Goal: reduction of event rate on active arm

Paper/talk: some settings are modified compared to the actual trial

T_c, T_a: times to event for control and active arms,

$$Pr\{T_c < t\} = 1 - e^{-\lambda_c t}, \quad Pr\{T_a < t\} = 1 - e^{-\lambda_a t},$$

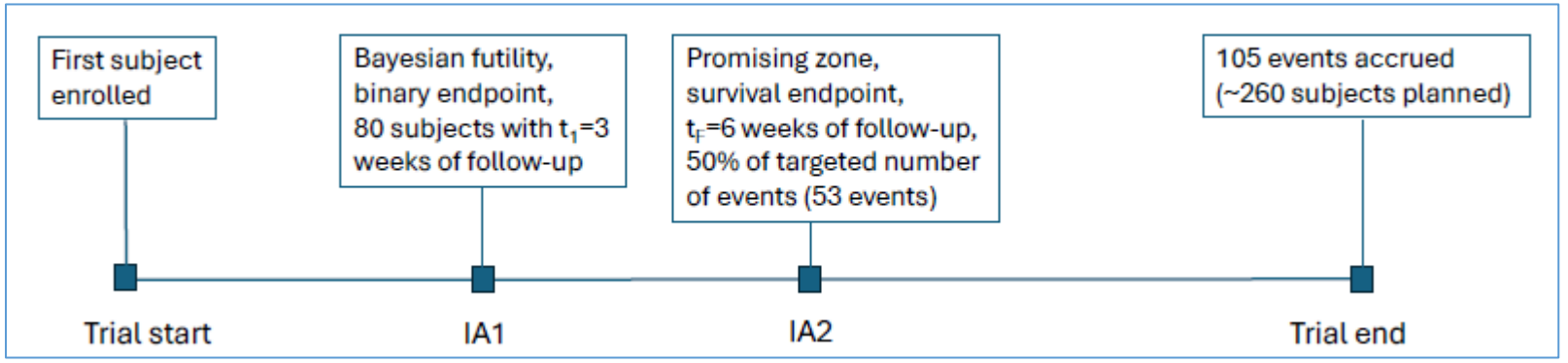
p_c, p_a: probabilities of event at the end of follow-up period,

$$\lambda_c = -\frac{\log(1-p_c)}{t_F}, \ \lambda_a = -\frac{\log(1-p_a)}{t_F} \ - \text{ hazard rates,}$$

Sample size calculation: $p_c = 0.5$, $p_a = 0.3$, $t_F = 6$ weeks.

Trial design (cont.)

- Sample size, GSD, event-driven: 105 events required, approximately 260 subjects
  - Power 90%, two-sided type I error rate of 5%
  - One interim analysis, information fraction of 0.5
    - No stopping for efficacy, futility boundary of 10% for conditional power (CP)
- New treatment: limited knowledge, substantial investment → Additional interim analysis (IA1) added based on Bayesian stopping criterion, after ~80 subjects complete t_1 = 3 weeks of follow-up
- Same exponential model for both endpoints: binary endpoint for IA1, survival endpoint for IA2



CSL Behring

IA1, futility: Bayesian posterior probabilities, beta-binomial model

Let $p_c^* = Pr\{T_c < t_1\},\ p_a^* = Pr\{T_a < t_1\}$ be event rates at time t_1.

- Prior: uninformative $Beta(1,1)$ for both arms $\Longrightarrow$
- Posterior: $Beta(r+1, n-r+1)$ with r events for n subjects.

Let $(\widetilde{p}_c, \widetilde{p}_a)$ be realizations from the posterior distributions $(p_c^*,\ p_a^*)$.

Criterion: stop for futility if

$$Pr\{\widetilde{p}_c - \widetilde{p}_a > 0 \mid data\} < \eta, \text{ where } \eta \in (0, 1].$$

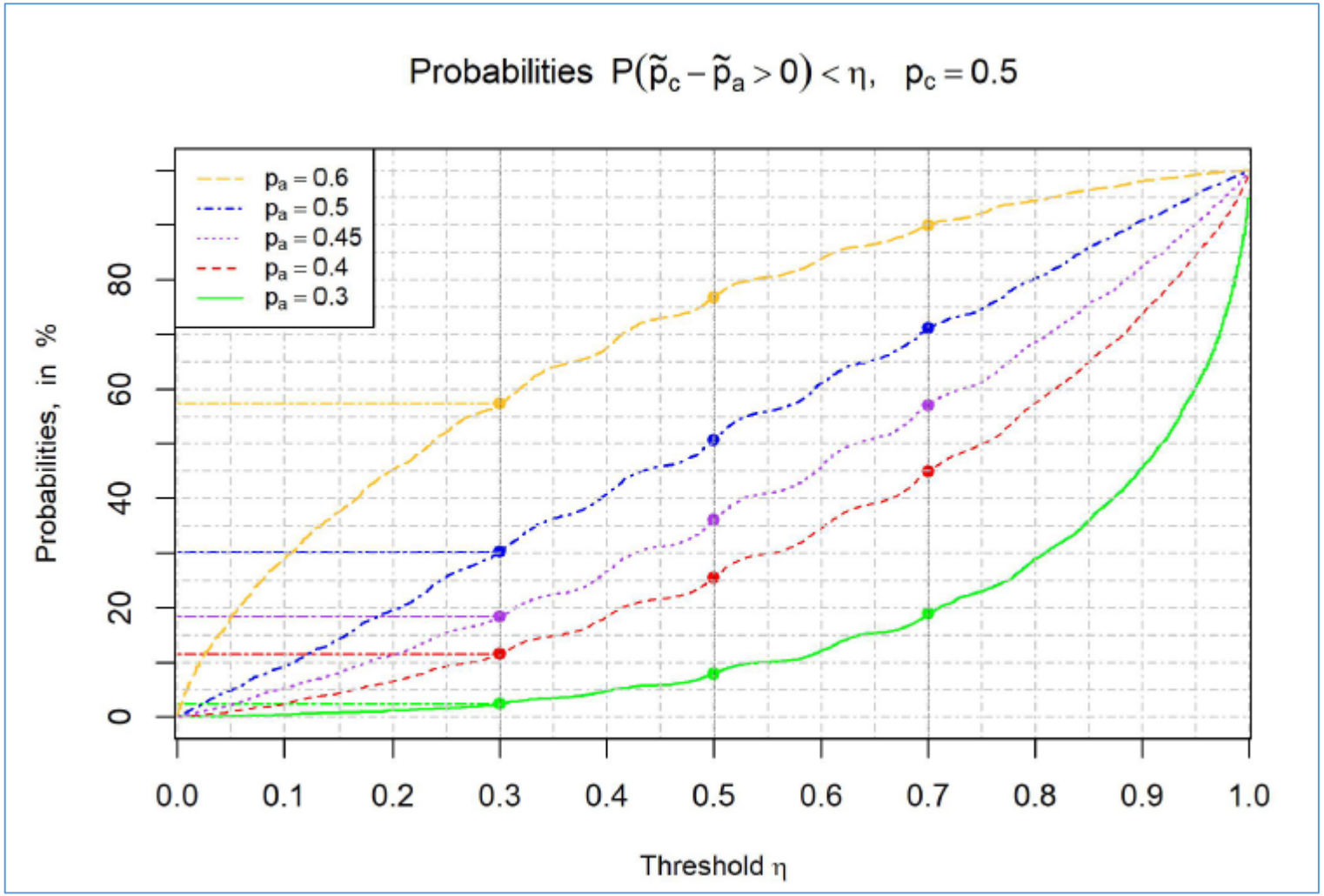
Why the above criterion is meaningful:

1. When drug "works" ($p_c - p_a$ is "large"), and t_1, n are large enough:
   (a) Difference $\Delta = \widetilde{p}_c - \widetilde{p}_a$ is expected to be "large".
   (b) Event $[\Delta > 0 \mid data]$ is likely true; $Pr\{\Delta > 0 \mid data\}$ is "large".
   (c) For "small" η, inequality $Pr\{\Delta > 0 \mid data\} < \eta$ will be unlikely.
2. When the drug does not work ($p_c - p_a$ is "small"/negative):
   (a) Difference $\Delta = \widetilde{p}_c - \widetilde{p}_a$ is likely "small" or negative.
   (b) Event $[\Delta > 0 \mid data]$ is likely false; $Pr\{\Delta > 0 \mid data\}$ is "small".
   (c) For "moderate" η, inequality $Pr\{\Delta > 0 \mid data\} < \eta$ will be likely.

Interim analysis IA1 (cont.)

Control arm rate $p_c = 0.5$.

- Ordering: for any given η and fixed p_c, probabilities corresponding to larger differences (p_c-p_a) are lower than those for smaller (p_c-p_a)
- Strong effect ($p_a = 0.3$): small probabilities for small/moderate η
- Weak effect ($p_a = 0.45$): probabilities are larger than for p_a=0.3
- Choice of η: would like "low" prob. of stopping when drug works



CSL Behring

Decision grids: to stop or not to stop at IA1?

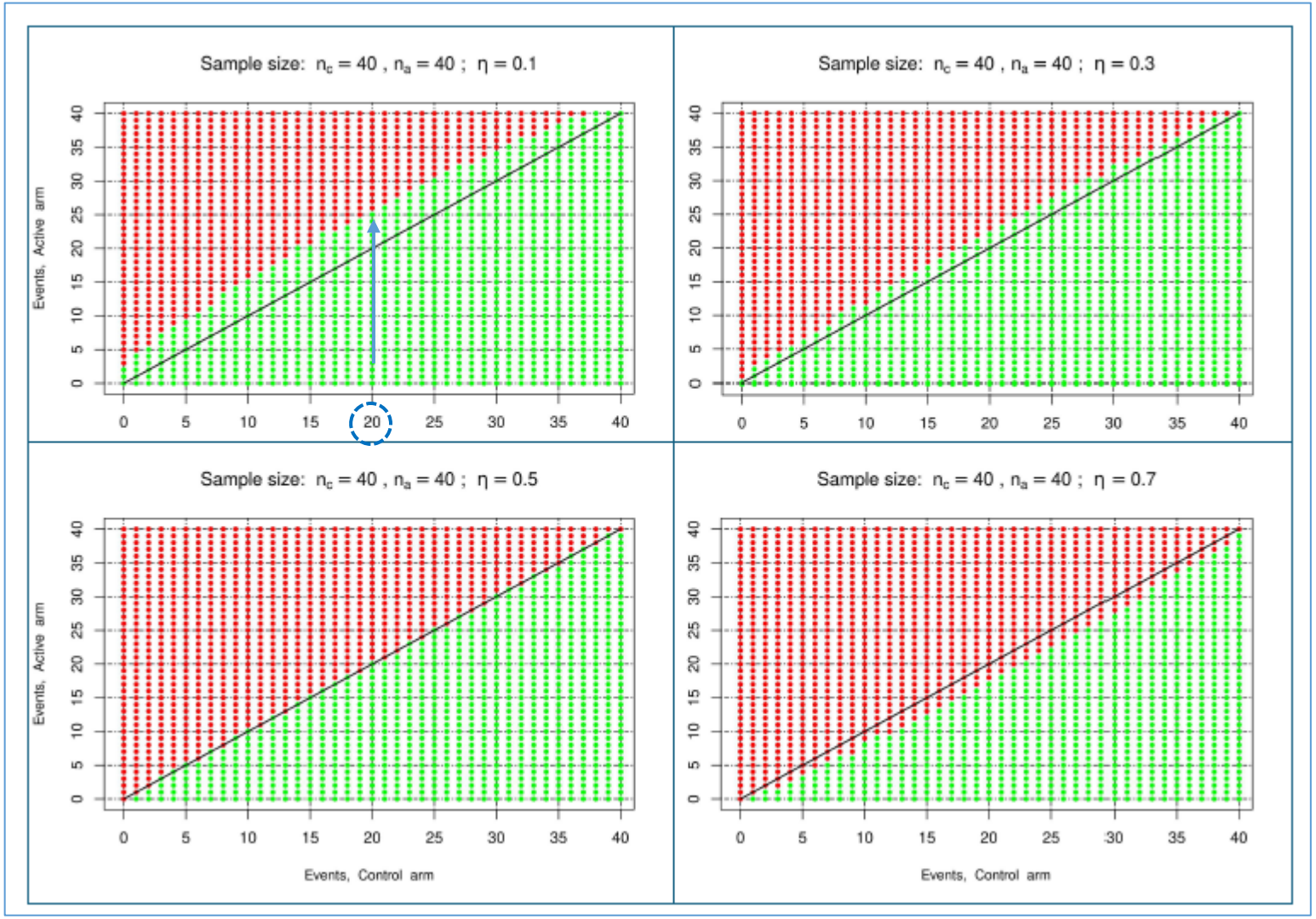
Assume $n_1 = 40$ per arm, calculate posterior probabilities for each combination (r_c, r_a) of no. of events (both numbers range from 0 to 40)

- If posterior probability for a given combination (r_c, r_a) is less than a particular threshold η, then a point is marked in red (stop for futility).
- If posterior probability is larger than η, then a point is marked in green (continue the trial).

Results

- Values of η between 0.3 and 0.5 seem reasonable
- $\eta = 0.1$: various scenarios suggest continuing the trial when no. of events r_a in the active arm is "substantially" larger than r_c. E.g., with 20 events in the control arm, the trial will stop only when $r_a \geq 26$

Ultimate choice of η: after the analysis of both IA1 and IA2



CSL Behring

IA2, conditional power (CP): case of normal i.i.d. observations

$X_j \sim \mathcal{N}(\mu, \sigma^2)$, $Y_j \sim \mathcal{N}(0, \sigma^2)$, $j = 1, \ldots, n$; $\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$, $\overline{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$.

Z - Wald statistic,

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{2\sigma^2/n}} \sim \mathcal{N}\left(\Delta\sqrt{\frac{n}{2}}, 1\right), \ \Delta = \frac{\mu}{\sigma} \ - \ \text{effect size} \qquad (1)$$

Two-stage design, $n = n_1 + n_2$; n_i - sample size/arm for stage i, $i = 1, 2$:

cumulative Wald statistic is the weighted sum of stage-wise Wald statistics,

$$Z = Z_1\sqrt{t} + Z_2\sqrt{1-t}, \quad t = \frac{n_1}{n},$$

t is *information time*; Z_1, Z_2 are defined as in (1).

$$EZ_i = \Delta\sqrt{\frac{n_i}{2}}, \quad Z_i \sim \mathcal{N}(0,1) \text{ under } H_0.$$

Potential for sample size increase / sample size re-estimation (SSR)

CP: prob. of success at final analysis given $Z_1 = z_1$ and behavior of Z_2.

CP under *current trend*: assume that estimated $\widehat{\Delta}$ at the IA is the true effect,

$$EZ_2 = \widehat{\Delta}\sqrt{\frac{n_2}{2}} = z_1\sqrt{\frac{2}{n_1}}\sqrt{\frac{n_2}{2}} = z_1\sqrt{\frac{1-t}{t}} \implies$$

$$\begin{aligned} CP_{\widehat{\Delta}} &= Pr_{\widehat{\Delta}}\{Z_1\sqrt{t} + Z_2\sqrt{1-t} > z_\alpha | Z_1 = z_1\} = \\ &= Pr_{\widehat{\Delta}}\left\{Z_2 - EZ_2 > \frac{z_\alpha}{\sqrt{1-t}} - EZ_2 - z_1\sqrt{\frac{t}{1-t}}\right\} = \\ &= \Phi\left(\frac{z_1}{\sqrt{t(1-t)}} - \frac{z_\alpha}{\sqrt{1-t}}\right) \end{aligned}$$

(Chen et al. (2004), Mehta and Pocock(2011))

IA2, SSR/Promising zone

Question 1: If CP is "low", can one increase sample size, $n_2 \to \tilde{n}_2$ (SSR), use $\tilde{n}_2$ at final analysis, increase CP /overall power and preserve type I error?

Answer: in general, NO. Type I error may be inflated dramatically.

Question 2: Any other methods that may work with SSR?

Answer: Yes, for example, weighted inverse normal method.

- Pros: type I error rate controlled.
- Cons: for Z_2, use weights based on originally planned n_2 (not $\tilde{n}_2$);
  observations from 2nd stage are downweighted;
  sufficient statistic is not used.

Promising zone approach: find conditions such that standard cumulative Wald statistic Z can be used with $\tilde{n}_2$, while controlling type I error rate

- Series of papers, from Chen et al. (2004) to Mehta and Pocock (2011).
- Use CP under the current trend.
- Range of Z_1 (CP estimates) divided into 3 'zones', closed-form

  - *Unfavorable*, [0,CP_u]: no sample size increase
  - *Promising*, [CP_u, CP_p]: increase from n_2 to $\tilde{n}_2$ based on information time t, target CP and maximal sample size increase.
  - *Favorable*, [CP_p, 1]: no sample size increase.

Additionally, futility zone may be added: stop trial at IA if CP $<$ CP_f.

IA1 & IA2, Operating characteristics

Not much different

Futility: CP ≤ 0.1, the trial

Unfavorable: 0.1 < CP ≤ 0.

Promising: 0.41< CP ≤ 0.9,

increase up to 50% (

Favorable: CP > 0.9, no ch

Main scenario, η = 0.3

Type I error, one-sided

Rates	η	Stop	Stop	Power	Fail	Unfavorable		Promising			Favorable		Power
p_c, p_a		IA1	IA2	SSR	End	All	Success	All	Success	Success No SSR	All	Success	No SSR
0.6,0.4	0.3	2.6	5.6	87.9	3.9	9.4	7.4	24.6	23.8	21.8	57.8	56.7	85.9
	0.5	7.8	4.1	84.7	3.4	8	6.4	22.8	22.1	20.2	57.2	56.2	82.8
	0.7	20.5	2	74.9	2.5	5.2	4.2	18	17.4	16.1	54.3	53.3	73.6
0.5,0.4	0.3	11.2	25.1	35.9	27.8	18.8	5.1	25.5	15.3	12.8	19.4	15.4	33.4
	0.5	24.9	17	33.5	24.6	15.6	4.3	23.4	14.1	11.8	19	15.1	31.2
	0.7	44.4	8.7	28.2	18.7	10.8	3	18.7	11.3	9.5	17.4	13.9	26.4
0.5,0.3	0.3	2.8	4	90.1	3.2	7.3	5.6	23.1	22.4	20.5	62.8	62.1	88.1
	0.5	8.2	3.1	86	2.8	6	4.6	21	20.5	18.8	61.7	60.9	84.3
	0.7	18.7	1.8	77.3	2.2	4.4	3.3	17.2	16.7	15.4	58	57.3	76
0.45,0.35	0.3	11.4	24.2	37.6	26.8	19.6	5.8	25.6	16.1	13.1	19.2	15.7	34.6
	0.5	24.6	17.4	34.6	23.4	16.3	4.9	23.1	14.5	11.9	18.6	15.2	32.1
	0.7	43.3	9.9	29.1	17.7	11.4	3.4	18.5	11.7	9.8	16.9	14	27.2
0.6,0.3	0.3	0.4	0.2	99.3	0	1.1	1.1	6.3	6.3	6.3	91.9	91.9	99.3
0.6,0.35		1	1.7	96.8	0.6	3.9	3.5	15.3	15.2	14.9	78.2	78	96.4
0.6,0.4		2.6	5.6	87.9	3.9	9.4	7.4	24.6	23.8	21.8	57.8	56.7	85.9
0.6,0.45		5	13.7	67.3	14	15.1	8.3	28.7	25.1	21.1	37.6	34	63.4
0.6,0.5		10	24.1	36.5	29.4	18.7	5.4	26	15.3	12.4	21.2	15.7	33.6
0.6,0.6		28.5	42.3	**2.2**	27	13.6	0.5	11.1	0.9	0.8	4.4	0.8	2.1
0.55,0.25		0.3	0.2	99.4	0	0.6	0.6	4.6	4.6	4.6	94.3	94.3	99.4
0.55,0.3		1.3	1.1	97.1	0.5	2.9	2.6	13.8	13.7	13.4	80.9	80.7	96.7
0.55,0.35		2.7	4.6	88.9	3.7	8.4	6.4	23.6	22.9	21	60.6	59.6	87
0.55,0.4		5.3	12.9	67.6	14.2	15.1	7.9	29.4	25.2	21.4	37.3	34.5	63.8
0.55,0.45		10	24.6	37.1	28.3	19.1	5.8	25.9	15.5	12.5	20.4	15.8	34.1
0.55,0.55		28.2	44.8	**1.8**	25.2	13.5	0.4	10.4	0.8	0.8	3.1	0.6	1.8
0.5,0.2		0.4	0	99.5	0	0.3	0.2	2.8	2.8	2.8	96.5	96.5	99.5
0.5,0.25		1.1	0.9	97.7	0.4	2.2	2	11.1	11.1	10.9	84.7	84.7	97.5
0.5,0.3		2.8	4	90.1	3.2	7.3	5.6	23.1	22.4	20.5	62.8	62.1	88.1
0.5,0.35		5.3	12	69.3	13.4	14.7	7.4	30	26.5	22.4	38	35.4	65.2
0.5,0.4		11.2	25.1	35.9	27.8	18.8	5.1	25.5	15.3	12.8	19.4	15.4	33.4
0.5,0.5		28.6	45.4	**2.1**	23.8	13.3	0.5	9.7	0.8	0.8	2.9	0.8	2.1

Operating characteristics, plots

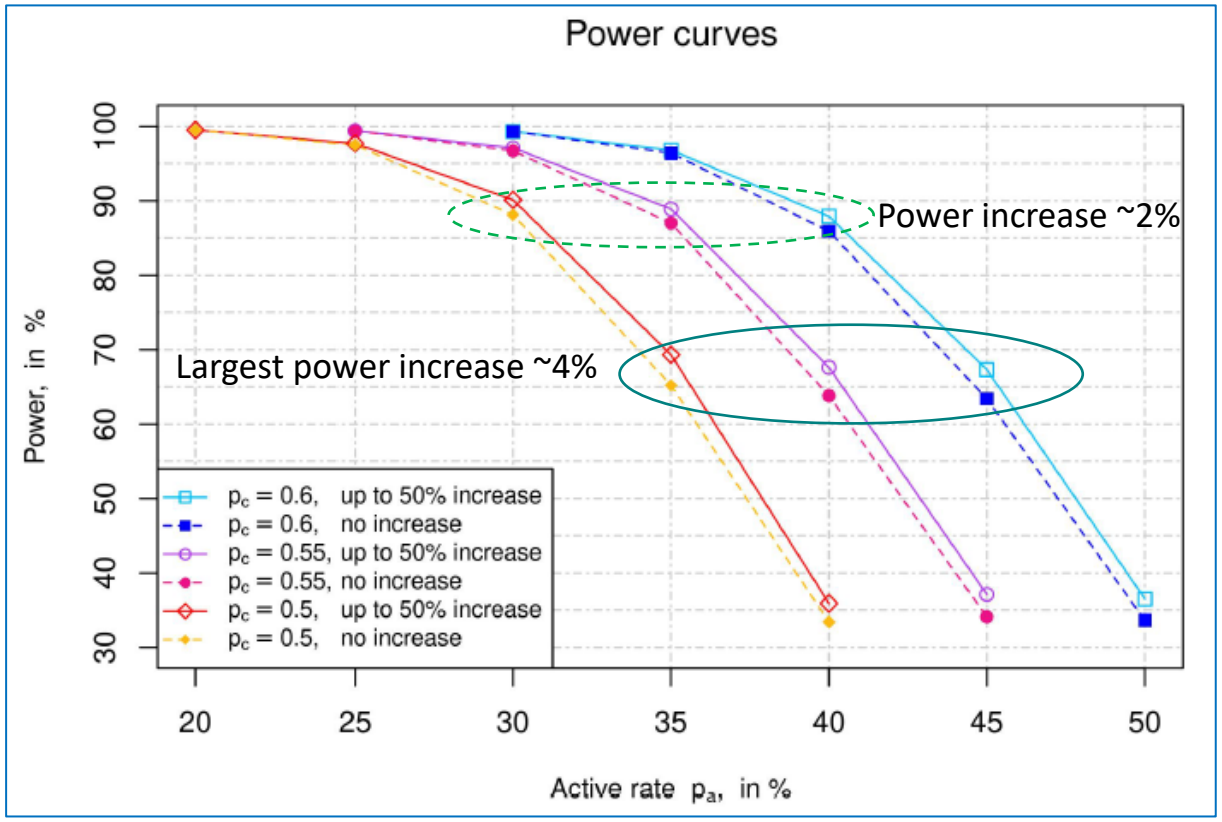
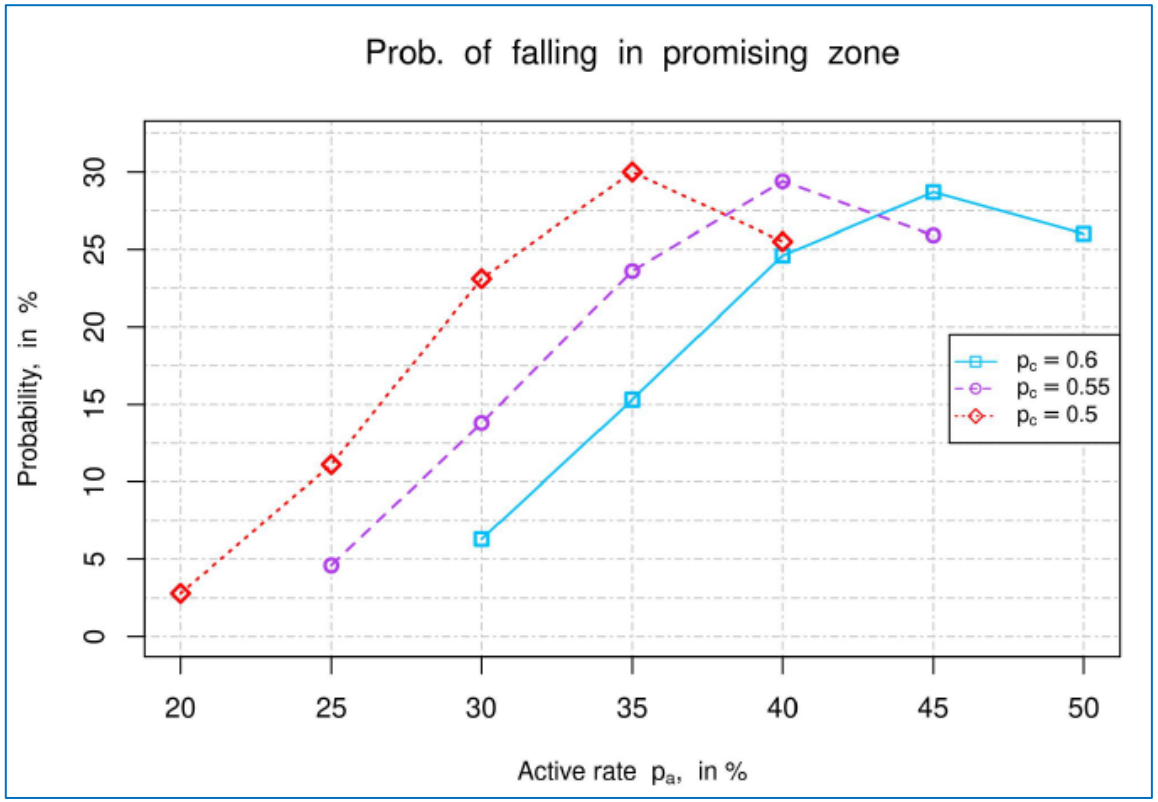


Fig. 6 Trial power, with and without sample size increase in the promising zone; futility threshold $\eta = 0.3$ for IA1. Solid lines with open squares correspond to power with SSR, column "Power, SSR" in Table 1; dashed lines with filled squares - to power without SSR, column "Power, No SSR". Light blue solid: $p_c = 0.6$ with SSR; blue dashed: $p_c = 0.6$, no SSR; purple solid: $p_c = 0.55$ with SSR; magenta dashed: $p_c = 0.55$, no SSR; red solid: $p_c = 0.5$ with SSR; orange dashed: $p_c = 0.5$, no SSR

Power, with/without SSR

Probability of "falling" in promising zone

Fig. 7 Probability of "falling" into promising zone; futility threshold $\eta = 0.3$ for IA1. Solid light blue line, open squares for $p_c = 0.6$; dashed purple line, open circles for $p_c = 0.55$; dotted red line, open diamonds for $p_c = 0.5$. Values from Table 1, column "Promising, All"



CSL Behring

Discussion

Promising zone

- Limited improvements due to SSR in scenarios of interest
- Greatest power gain may be achieved outside promising zone (Jennison, Turnbull, 2015) – but unlikely to achieve target power
- Large uncertainty for the team when in unfavorable or favorable zones (could be either rather bad or exceptionally good)

Hybrid design

- Early learning-oriented futility decisions based on short-term endpoints
- Standard software tools not available → Developed extensive simulation framework for IAs
- Combination/strengths of both Bayesian and frequentist approaches: learning-oriented futility assessment at IA1 and rigorously controlled sample-size adaptation at IA2

Promising zone, Mehta and Pocock (2011)

Emerson et al. (2011) review: harsh criticism on **17 printed pages**...

- We do not find that such an adaptive rule confers any advantage by the usual criteria for clinical trial design… From years of recreational spelunking, we have noted parallels between research and cave exploration. In both processes, explorers spend their time in the dark exploring the maze of potential leads, most often without a clear idea of where they will end up. Because the overwhelming majority of such leads are dead ends, the most useful companions to have along with you are the ones who will willingly explore the dead ends.
- … the problem we have faced in the statistical literature about adaptive designs versus GSD is the ever changing optimality criteria. It is like a game of `Whac-a-mole': As soon as one criticizes one aspect of an adaptive design or group sequential design, the proponents produce other optimality criteria.

Response of Mehta and Pocock to comments of Emerson et al. (2011): **single page**

- … Why have we (and many other authors) taken such an interest in adaptive sample size re-estimation? The key point is that fixed and group sequential designs require an up-front commitment to a prespecified maximum trial size. If the sponsor is prepared to do that then they have no need for an adaptive design, and hence the arguments of Emerson et al. are correct
- Emerson et al. are not alone in wishing to dismiss this `adaptive game' as not worth playing: indeed one of us (SJP) was amongst the skeptics only a few years ago. Furthermore, when trial planners and sponsors are attracted to the idea of adaptive sample-size re-estimation based on unblinded interim data, an initial gambit from a well informed biostatistical collaborator may well be to try and `talk them out of it' by encouraging an up-front commitment to a larger trial size, possibly with appropriate group sequential guidelines. However, the attraction of such statistical efficiency is of no use if the trialists simply will not (or cannot) agree to such a design of the type Emerson et al. elucidate.'

Scientific discussion: frequentists vs Bayesians

On the Future of Statistics—a Second Look

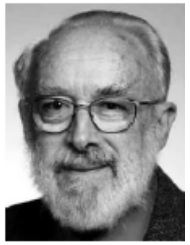
By M. G. KENDALL

JRSS (A), **131** (1968), 182-204

12. It may be that statisticians will learn a certain amount of restraint, but I doubt it. On other occasions, for example, I have lamented that Bayesian statisticians do not stick closely enough to the pattern laid down by Bayes himself: if they would only do as he did and publish posthumously we should all be saved a lot of trouble.

p.185

THE FUTURE OF STATISTICS — A BAYESIAN 21ST CENTURY



D. V. LINDLEY, *University College London and University of Iowa*

The thesis behind this talk is very simple: the only good statistics is Bayesian statistics. Bayesian statistics is not just another technique to be added to our repertoire alongside, for example, multivariate analysis; it is the only method that can produce sound inferences and decisions in multivariate, or any other branch of, statistics. It is not just another chapter to add to that elementary text you are writing; it is that text. It follows that the unique direction for mathematical statistics must be along the Bayesian road.

Supp. Adv. Appl. Prob. 7, 106–115 (1975)

In pursuing a sound path it is important to rescue those who are trying to cross the mountains by bad routes. What shall we do with sampling theory statistics, with significance tests, with confidence intervals; with all those methods that violate the likelihood principle? The answer is, let them die. Their role has been to provide valuable stepping-stones to the future and our appreciation of the originators of these ideas should not be diminished by this remark: for it is largely by the pursuit of the notions that we have reached the understanding that we have today. Each of these techniques has its Bayesian equivalent, which makes better practical sense, and I see no excuse for wasting our time on them except in a course on the history of our subject.* It is, I think,

p.112

CSL Behring

More on scientific discussion/development

Bayesian Statistics
(a very brief introduction)

Ken Rice

Epi 516, Biost 520

1.30pm, T478, April 4, 2018

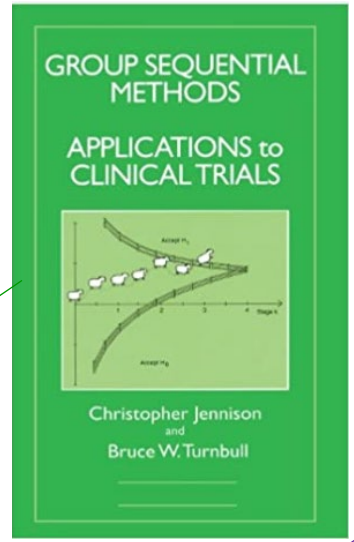
Dept of Biostatistics, Univ. of Washington

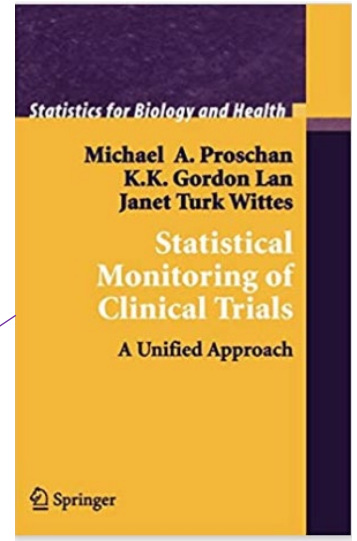
Using a spade for some jobs and shovel for others does *not* require you to sign up to a lifetime of using only Spadian or Shovelist philosophy, or to believing that *only* spades or *only* shovels represent the One True Path to garden neatness.

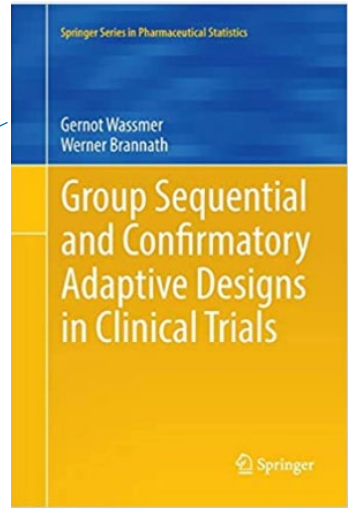
There are different ways of tackling statistical problems, too.

References

- Chen, Y.H.J., DeMets, D.L., Lan, K.K.G. (2004). Increasing the sample size when the unblinded interim result is promising. *Statist. Med*., **23**(7), 1023–1038.
- Emerson, S.S., Levin, G.P., Emerson, S.C. (2011). Comments on 'Adaptive increase in sample size when interim results are promising: A practical guide with examples'. *Statist. Med*., **30**, 3285–3301.
- Gao, P., Ware, J.H., Mehta, C. (2008). Sample size re-estimation for adaptive sequential design in clinical trials. *J. Biopharm. Statist*., **18**, 1184–1196.
- Jennison, C., Turnbull, B.W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. CRC/Chapman & Hall, Boca Raton, FL.
- Jennison, C., Turnbull, B.W. (2015). Adaptive sample size modification in clinical trials: start small then ask for more? *Statist. Med*., **34**, 3793–3810.
- Mehta, C.R., Pocock, S.J. (2011). Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statist. Med*., **30**, 3267–3284.
- Mehta, C.R., Pocock, S.J. (2011). Authors' response to 'Comment on adaptive increase in sample size when interim results are promising'. *Statist. Med*., **30**, 3302–3303.
- Proschan, M.A., Lan, K.K.G., Wittes, J.T. (2006). *Statistical Monitoring of Clinical Trials. A Unified Approach*. Springer, New York.
- Wassmer, G., Brannath, W. (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer, Switzerland.







CSL Behring