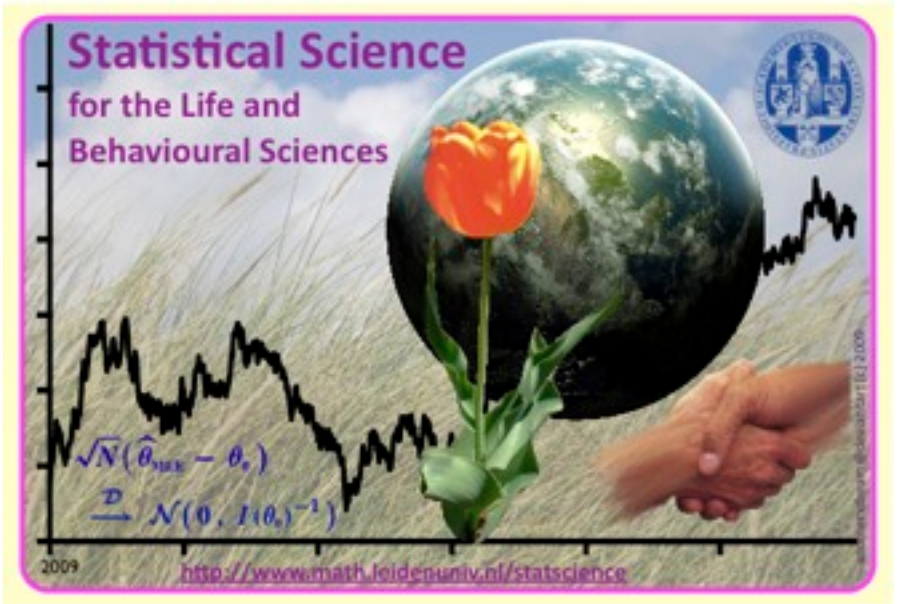
Bijdrage van de statistiek: Vroeg stoppen voor *futiliteit*

http://www.math.leidenuniv.nl/~gill/CCMO.pdf

Richard Gill
Universiteit Leiden



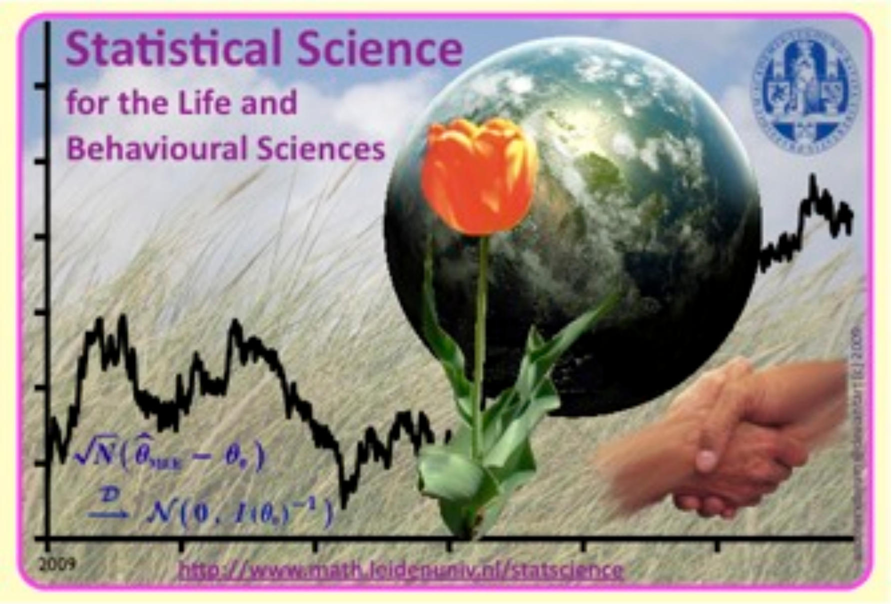
CCMO

Erasmus University
11 december 2009

Bijdrage van de statistiek: Vroeg stoppen voor *futiliteit*

Errors of 1st and 2nd kind (1- *or* 2- sided),
are not enough!

Richard Gill
Universiteit Leiden



ccMo

Erasmus University
11 december 2009

Apologia

- Ik ben (engels) wiskundige – statisticus – wetenschapper
- **Statistiek**: koningin, dienstmeisje of *femme fatale* ?
- Bemoeienis in affaires zoals: Lucia de B, Claudia Pechstein, **PROPATRIA** ... uit nieuwsgierigheid en uit overtuiging dat licht wordt geworpen door *wetenschappelijk* instelling
- Een professioneel statisticus met ervaring op relevante toepassingsgebied is bij uitstek gekwalificeerd !
  - [Proof of my pudding will be in your eating !]

Inleiding

Altman (1981): 50% of published medical statistics is wrong
Today: about 15%

80% of doctors ignore evidence based medicine

90% of all statistics are just made up

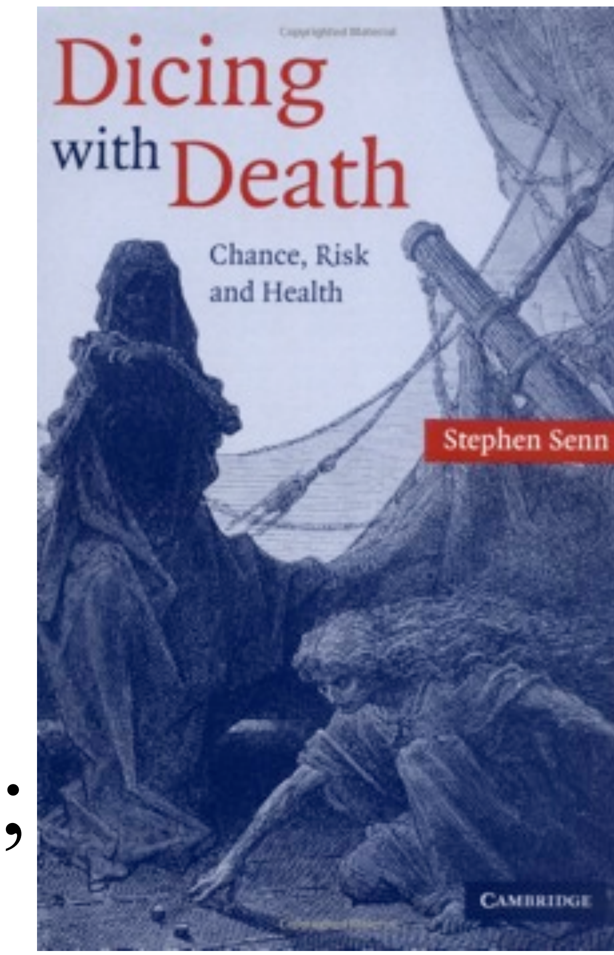
This talk:

- Theory: Snapinn (1992) *Statistics in Medicine*
  early stopping rule for randomized clinical trials
- Illustration: the Propatria trial
  (infectious complication in acute pancreatitis)

Dr. Steven M. Snapinn
Vice President, Global Biostatistics & Epidemiology at Amgen

Take home messages

Steve Snapinn,
Stephen Senn



- Early stopping for *futility* is a valuable statistical safety measure

open deur...

- Statistical issues in RCT's are complex; planning and implementation requires professional interdisciplinary collaboration

Special thanks to: Hein Gooszen & his team, the CCMO, certain journalists of NRC and NOVA, Maxim Kuil (Univ.Leiden), J.E.R.F., and *many* more colleagues, both at home and abroad

From: Stephen Senn <stephen@stats.gla.ac.uk>
Subject: Re: [Evidence] disastrous clinical trial
Date: February 24, 2008 11:36:52 CET
To: Richard Gill <gill@math.leidenuniv.nl>
Cc: evidence@casa.ucl.ac.uk

I have acted on a number of data safety monitoring committees and have the following points of view:

1) It is usually appropriate to have asymmetric stopping rules whereby one would stop a trial much earlier in favour of the standard than the new treatment. This is because if the standard is better, on stopping the trial all the patients in the world get the better treatment whereas if the new treatment is better most (and possibly even all patients in the short run) will continue to get the worse treatment.

2) Hence stopping for futility is the most important reason to stop. In the case of this trial had that philosophy applied it would have been stopped earlier.

3) Asymmetric stopping rules make triple blind inappropriate.

4) The important decision to make in stopping a trial is not "is A better than B" but "are patients being harmed by continuing the trial".

5) Equipoise is irrelevant to the ethics of clinical trials, instead, a Rawlsian perspective is needed:
S. Senn (2002), Ethical considerations concerning treatment allocation in drug development trials, *Statistical Methods in Medical Research* **11** (5) 403–411
http://smm.sagepub.com/cgi/content/abstract/11/5/403

Stephen Senn

I agree with Niels. Clinical trials are not nice, but necessary. Today's extensive medical knowledge about which treatments actually work (in spite of limitations) would not at all be possible without clinical trials. Furthermore, it is very hard work to carry out a good clinical trial! Also, it has increasingly struck me how difficult it is to draw any certain conclusions about the value of treatments from basic medical (lab) research. Sure the ideas for new treatments usually come from the labs, but the sorting out of what really works (instead of say killing patients) is up to statistics.
This does not mean that Richard may not be right about his criticism.
Regards from Odd [Aalen - Univ Oslo]

Niels Keiding [Univ Copenhagen] wrote:
Native British speakers do not always recognize that their subtle ironies may get lost in the noise when aliens like me try to receive the message. So let me try to decode what I think Jane means, at the risk of sounding far from elegant. In my unironic view clinical trial methodology is at a very low end as regards entertainment value. And we have all heard about incompetence and outright fraud. BUT There Is No Alternative (TINA, as they said in the Thatcher days). Anything else is worse, and the majority of trials that I have met are run by conscientious and competent scientists & statisticians. Of course they should be criticized when relevant, but we need to stand guard around this tool, which is at least far better than subjective judgment.
Regards to all, Niels

Jane Hutton [Univ Warwick] skrev:
Dear Richard, and colleagues
It is interesting, and disappointing to see there can still be a summary of a trial, with _primary_ endpoints reported as showing no difference, and death ignored as an outcome. Yes, we can learn from this. The public and legislaturers might learn to avoid clinical trials...
Regards, Jane

Statistical Tests

- Null hypothesis: no effect
- Alternative: there is an effect
- Errors of two kinds, error probabilities
  - **α**, deciding there *is* an effect, when in fact there's *none*
  - **β**, *n*ot deciding there's an effect, when in fact there *is*

α a.k.a. *significance level*

β often used for complementary chance $1 - \beta$ a.k.a. *power*

Statistical Tests

- Null hypothesis: no difference
- Alternative: there is a difference
- Errors of two kinds, error probabilities
  - α, deciding there *is* an effect, when in fact there's *none*
  - β, *n*ot deciding there's an effect, when in fact there *is*
    - Limit risk of Type I error to 5%
    - Maximize *power* subject to *significance level*; i.e.,
    - Minimize Type II error probability β for given Type I error probability α

More realistically

- Null hypothesis: no difference
  alternative: *desired*: +ve difference
  alternative: *disaster*: –ve difference
- Errors of *six kinds*
  - deciding there is an +/– effect, when there is none
  - not deciding there is an effect, when there is one (+/–)
  - deciding there is a +/– effect when it is the other way round (–/+)

2+2+2=6

We want interim analyses!

We want to have our cake <u>and</u> eat it!

- Stop early in case of –ve effect (safety!)
- Stop early in case of +ve effect (ethics!)
- Stop early if no effect (money!)
- Conserve nominal *significance level* α
- Usual evaluation when the trial doesn't stop early
- Conserve *power*, i.e., same as power $1 - \beta$ without interim analysis

We want interim analyses!

NO FREE LUNCHES

- Stop early in case of –ve effect (safety!)
- Stop early in case of +ve effect (ethics!)
- Stop early if no effect (money!)
- Conserve nominal *significance level* α
- Don't alter evaluation when the trial doesn't stop early
- Conserve *power*, i.e., same as power $1 - \beta$ without interim analysis

We want interim analyses!

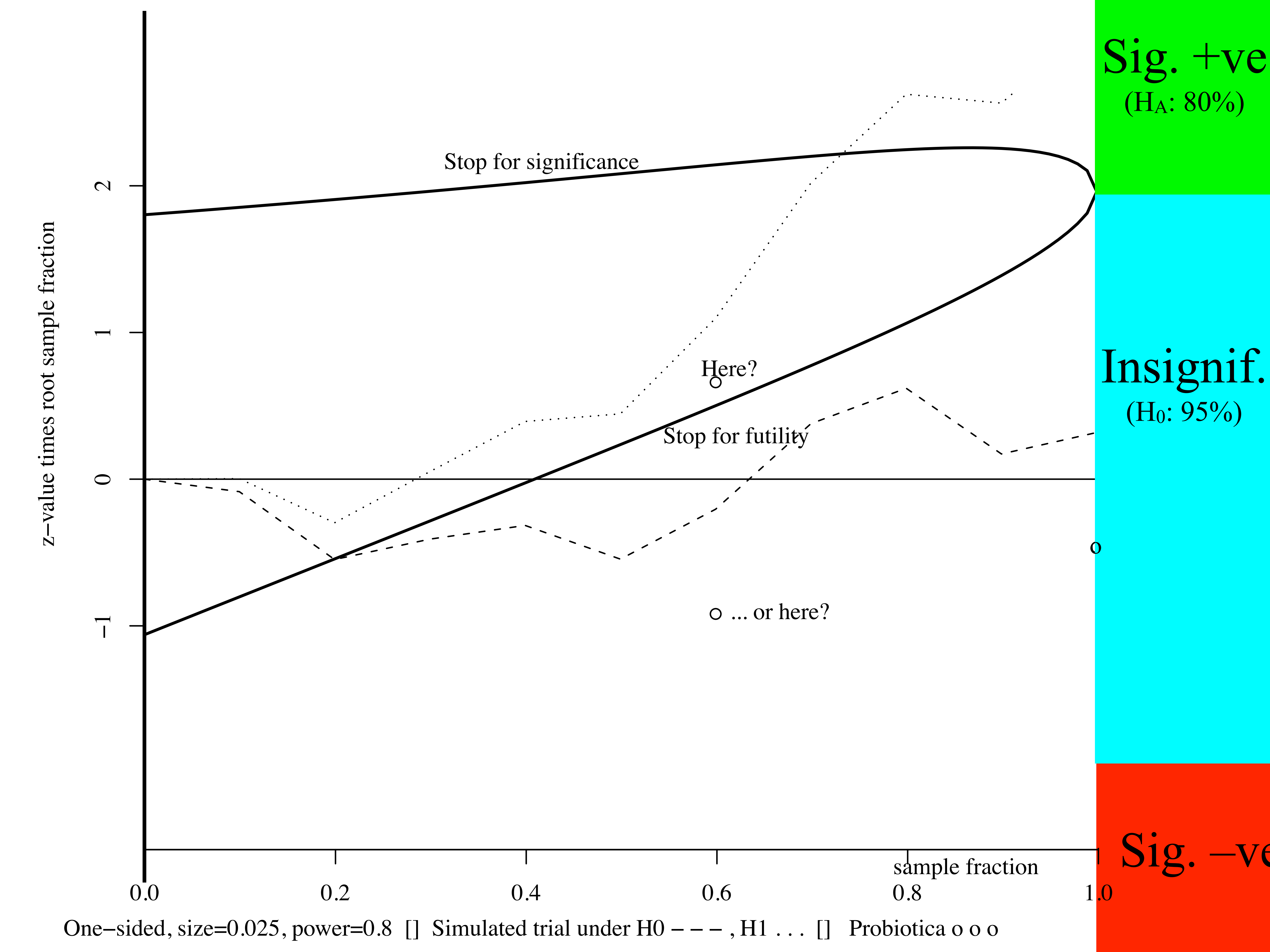
Snapinn: voor een dubbeltje op eerste rang

- Stop early in case of –ve effect (safety!)
- Stop early in case of +ve effect (ethics!)
- Stop early if no effect (money!)
- Conserve nominal *significance level* α
- Don't alter evaluation when the trial doesn't stop early
- ***Small power loss*** relative to power $1-\beta$ without interim analysis

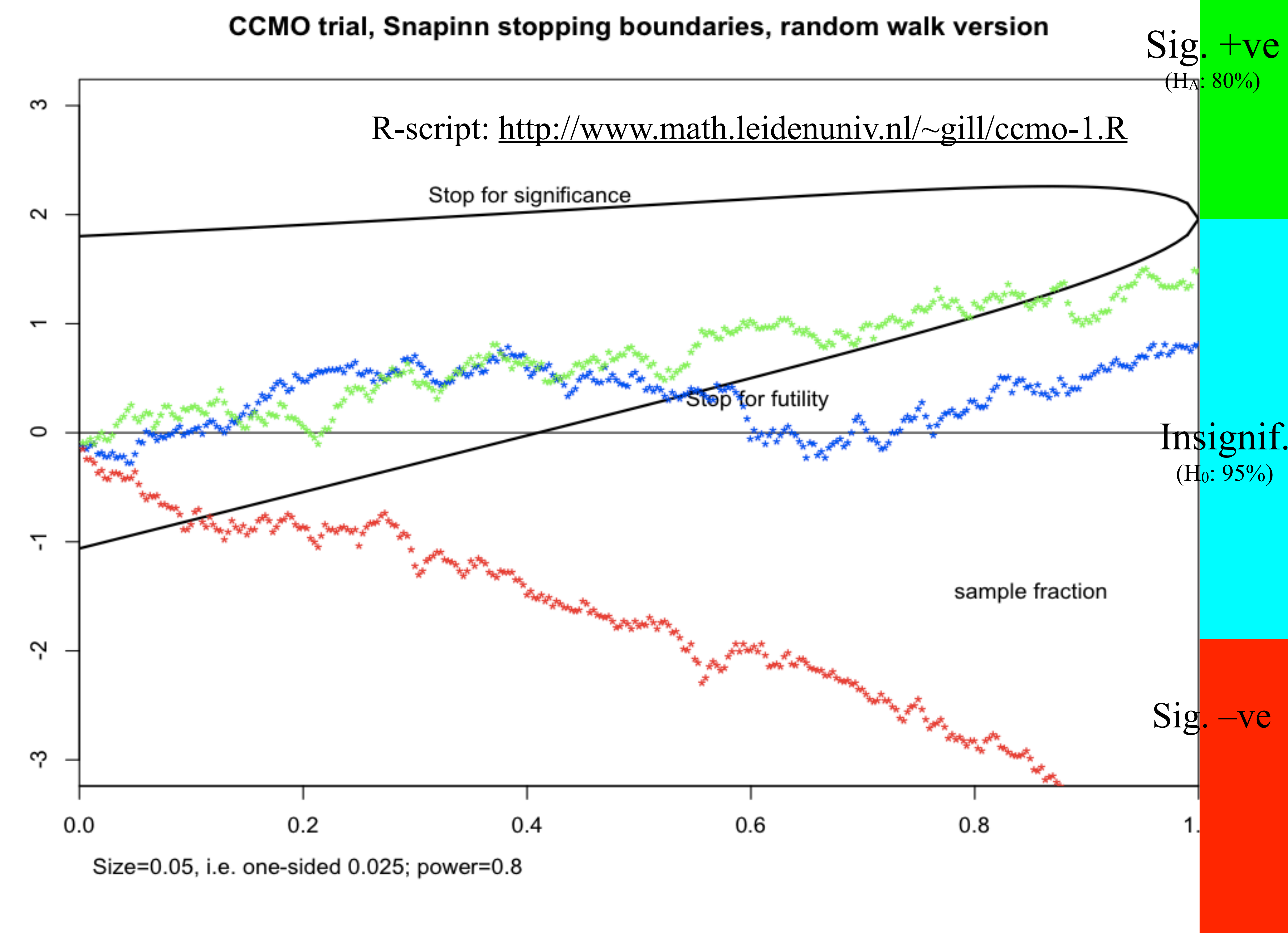
Snapinn does exactly that:
let me explain ...

- Some technicalities for connoisseurs
  [if time allows]

- Monte Carlo experiment:
  100000 Probiotica trials under 3 scenarios,
  w. & w.out Snapinn)
  [if computer+beamer works]

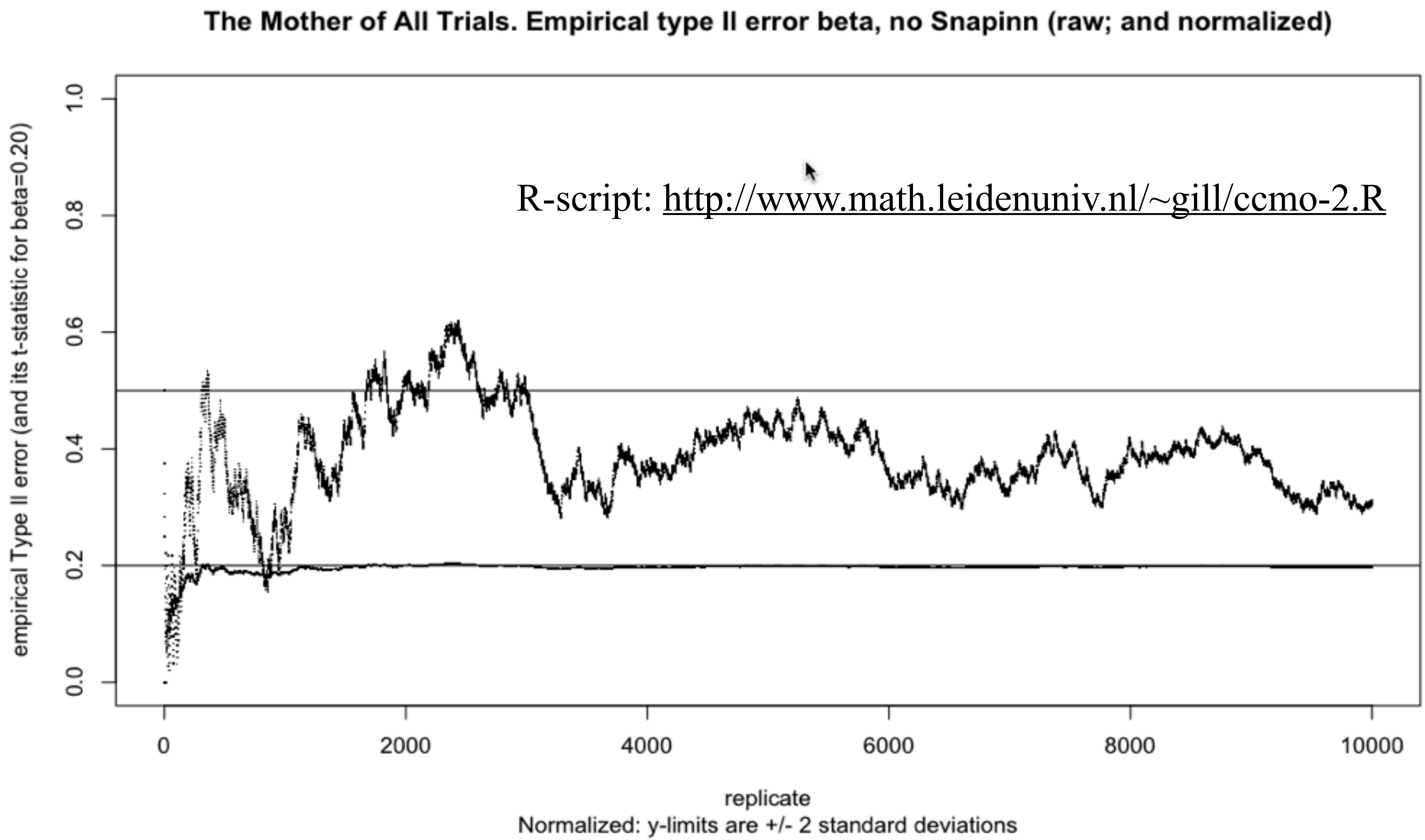


Sig. +ve
(H_A: 80%)
Stop for significance
2
1
Here?
Insignif.
(H_0: 95%)
Stop for futility
0
o
z−value times root sample fraction
−1
o ... or here?
Sig. −ve
sample fraction
0.0
0.2
0.4
0.6
0.8
1.0
One−sided, size=0.025, power=0.8 [] Simulated trial under H0 − − − , H1 . . . [] Probiotica o o o


CCMO trial, Snapinn stopping boundaries, random walk version
R-script: http://www.math.leidenuniv.nl/~gill/ccmo-1.R
Stop for significance
Stop for futility
sample fraction
Size=0.05, i.e. one-sided 0.025; power=0.8
Sig. +ve
(HA: 80%)
Insignif.
(H0: 95%)
Sig. –ve
3
2
1
0
-1
-2
-3
0.0
0.2
0.4
0.6
0.8
1.

The Mother of all Trials

10000 Probiotica trials, three scenarios, with and without Snapinn early stopping rules



The Mother of all Trials

10000 Probiotica trials, three scenarios, with and without Snapinn early stopping rules

- Null hypothesis true: one-sided Type 1 error = 2.48%, stopped for futilty = 66.82%, stopped for significance = 0.48%.

- Alternative hypothesis true: Type 2 error = 21.25%, stopped for futilty = 6.45%, stopped for significance = 23.59%.

- Disaster hypothesis true: Disaster = 0%, stopped for futilty = 99.32 %, stopped for significance = 0 %.

- No Snapinn: one-sided Type 1 error = 2.4%, Type II error = 20.02%, Type 1 error in disaster scenario = 0%.

R-script: http://www.math.leidenuniv.nl/~gill/ccmo-2.R

References

- Besselink et al. (2004) *BMC Surgery*
- Besselink et al. (2008) *Lancet*
- Gill (2008) *Statistica Neerlandica*

  http://www.math.leidenuniv.nl/~gill/CCMO.pdf
- Snapinn (1992) *Statistics in Medicine*
- Schouten (1995) *Klinische Statistiek*

STATISTICS IN MEDICINE, VOL. 11, 659–672 (1992)

MONITORING CLINICAL TRIALS WITH A CONDITIONAL PROBABILITY STOPPING RULE

STEVEN M. SNAPINN

Merck Sharp & Dohme Research Laboratories, BL3-2, West Point, Pennsylvania 19486, U.S.A.

SUMMARY

Conditional probability procedures offer a flexible means of performing sequential analysis of clinical trials. Since these procedures are not based on repeated significance tests, the number and schedule of the interim analyses is less important than with group sequential procedures. Their main disadvantage is that the magnitude of their effect on the significance level is difficult to assess. This paper describes a conditional probability procedure which attempts to maintain the overall significance level by balancing the probabilities of false early rejection and false early acceptance. Monte Carlo sampling results suggest that this procedure can achieve a large reduction in expected sample size without greatly affecting either the significance level or power of the trial.

1. INTRODUCTION

Sequential analysis of clinical trials is usually performed for two reasons: to reduce the expected sample size and thus spare study resources, and for ethical reasons. Typically, some type of group sequential procedure[1–3] is used. Using one of these procedures, a prespecified number of interim

The critical p-values are

$$\text{Rejection boundary} = 1 - \Phi\left(\frac{z_{1-\alpha} + \sqrt{(1-f)}\,z_{p_{\text{rej}}}}{\sqrt{f(2-f)}}\right)$$

and

$$\text{Acceptance boundary} = 1 - \Phi\left(\frac{f(2-f)z_{1-\alpha} + \sqrt{(1-f)}\,z_{p_{\text{acc}}} - (1-f)^2 z_{1-\beta}}{\sqrt{f(2-f)}}\right).$$

Table II gives an example of these boundaries as a function of f, using $\alpha = 0{\cdot}025$, $\beta = 0{\cdot}05$, $p_{\text{rej}} = 0{\cdot}95$, and $p_{\text{acc}} = 0{\cdot}10$. So, for example, if an analysis without any adjustment for multiple testing is done after 60 per cent of the patients have completed the trial, then an attained p-value of 0·0028 or less is required to reject the null hypothesis, and an attained p-value of 0·298 or more is required to accept the null hypothesis. If the attained p-value is between these limits, then the trial continues.

Table II. Critical P-values required for early rejection and early acceptance, based on $\alpha = 0{\cdot}025$, $\beta = 0{\cdot}05$, $p_{\text{rej}} = 0{\cdot}95$ and $p_{\text{acc}} = 0{\cdot}10$

	Critical value for early	
f	Rejection	Acceptance
0·10	< 0·0001	> 0·999
0·20	< 0·0001	0·968
0·30	0·0002	0·826
0·40	0·0007	0·629
0·50	0·0016	0·444
0·60	0·0028	0·298
0·70	0·0043	0·195
0·80	0·0060	0·123
0·90	0·0087	0·072
1·00	0·0250	0·025

PHARMACEUTICAL STATISTICS
Pharmaceut. Statist. 2006; **5**: 273–281
Published online 5 May 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/pst.216

Assessment of futility in clinical trials

MAIN PAPER

Steven Snapinn*,†, Mon-Gy Chen, Qi Jiang and Tony Koutsoukos
Amgen Inc., One Amgen Center Drive, 24-2-C, Thousand Oaks, CA 91320, USA

The term 'futility' is used to refer to the inability of a clinical trial to achieve its objectives. In particular, stopping a clinical trial when the interim results suggest that it is unlikely to achieve statistical significance can save resources that could be used on more promising research. There are various approaches that have been proposed to assess futility, including stochastic curtailment, predictive power, predictive probability, and group sequential methods. In this paper, we describe and contrast these approaches, and discuss several issues associated with futility analyses, such as ethical considerations, whether or not type I error can or should be reclaimed, one-sided vs two-sided futility rules, and the impact of futility analyses on power.

Keywords: *conditional power; predictive power; predictive probability; sequential analysis; stochastic curtailment*

Klinische statistiek / druk 2

een praktische inleiding in methodologie en analyse
Auteur: H.J.A. Schouten
Paperback
273 pagina's | Bohn Stafleu van Loghum | mei 1999

In tabel 4 s moment, waarbij de significantie-grens vo bekende 0,05. Deze grenzen werden door dan door O'Brien en Fleming (1979) zelf. I ls in het onderzoeks-protocol staat dat in wordt het betreffende experiment gestaakt 0,015 bij de tweede analyse. Als P = 0,04 ficant verschil als we corrigeren voor het g g de P-waarde zelf te vermelden en niet alle significant is.

Tabel 4.
Stopregels van O'Br 0,05.

		5
Opeen-	1	0,001
volgende	4	0,001
signifi-	9	0,008
cantie	3	0,023
grenzen		0,041

12.4 Een éénzijdige stopregel volgens Snapinn

Om praktische redenen ben ik zeer enthousiast over de stopregels volgens het systeem van Snapinn (1992). De Snapinn stopregels maken het niet alleen mogelijk te stoppen bij een significant verschil, maar bieden ook de mogelijkheid te stoppen wanneer een significant verschil niet langer te verwachten is. Een ander belangrijk voordeel is dat de uiteindelijke statistische analyse niet gecorrigeerd hoeft te worden voor de gebruikte stopregel, hetgeen de interpretatie aanzienlijk kan vereenvoudigen. Een plezierig gevolg hiervan is dat het aantal patiënten gewoon op de conventionele manier kan worden berekend; het gaat dan om het aantal patiënten dat nodig is als het experiment niet tussentijds wordt afgebroken.

Een schijnbaar nadeel van de stopregels van Snapinn is dat vrijwel nooit in de eerste helft van het experiment zal worden gestopt. Maar bij het opstellen van de alternatieve hypothese worden vaak wonderen verwacht, omdat anders het berekende aantal patiënten onhaalbaar groot wordt. Dit betekent dat de gevoeligheid van de statistische toets in de eerste helft van het experiment te gering is om een enigszins realistisch verschil te kunnen aantonen. Bovendien is de bewijskracht van een significant verschil tamelijk klein bij een gering aantal patiënten, zoals in paragraaf 11.2 werd uitgelegd.

Het aantal tussentijdse analyses hoeft niet bij voorbaat vast te staan, ofschoon dat om organisatorische redenen wenselijk kan zijn. De stopgrenzen voor de P-waarde hangen af van de fractie f volledig geëvalueerde patiënten; bij de analyse van overlevingsduren is f de fractie reeds opgetreden sterfgevallen (van het totale aantal sterfgevallen dat te verwachten is als de alternatieve hypothese waar is). In de tweede kolom van tabel 5 staan de stopgrenzen waarbij

Statistica Neerlandica (2009) Vol. 63, nr. 1, pp. 1–12
doi:10.1111/j.1467-9574.2008.00411.x

Statistics, ethics and probiotica

Richard D. Gill*

Mathematical Institute, Leiden University, P. O. Box 9512, 2300 RA Leiden, The Netherlands

Ethical issues involved in the design of the 'PROPATRIA' probiotica trial are discussed. This randomized clinical trial appeared to be well conducted according to accepted good practices. The finding that the treatment was actually rather harmful, and that despite this, and despite a built-in interim analysis, the trial was not stopped earlier, led to strong criticism in the media. I argue that 'accepted good practices' need to be reconsidered in the light of this experience. First, a much stronger distinction needs to be recognized between the immediate interests of the patients being treated in the trial and the interests of future patients of future doctors elsewhere. Secondly, it is in the interests of future patients that well-conducted clinical trials are accepted by society. As it is unavoidable that an occasional trial will result in an unpredicted severely negative outcome, ethical screening committees must ensure that those performing a trial can never be accused of putting the interest of 'science' above the interest of their own patients when

BMC Surgery

Study protocol

Open Access

Probiotic prophylaxis in patients with predicted severe acute pancreatitis (PROPATRIA): design and rationale of a double-blind, placebo-controlled randomised multicenter trial [ISRCTN38327949]

Marc GH Besselink*[1], Harro M Timmerman[1], Erik Buskens[2], Vincent B Nieuwenhuijs[1], Louis MA Akkermans[1], Hein G Gooszen[1] and the members of the Dutch Acute Pancreatitis Study Group

Address: [1]Department of Surgery, University Medical Center Utrecht PO Box 85500, HP G04.228, 3508 GA Utrecht, The Netherlands and [2]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht PO Box 85060, 3500 AB Utrecht, The Netherlands

Email: Marc GH Besselink* - m.besselink@chir.azu.nl; Harro M Timmerman - h.timmerman@chir.azu.nl; Erik Buskens - e.buskens@umcutrecht.nl; Vincent B Nieuwenhuijs - v.b.nieuwenhuijs@hetnet.nl; Louis MA Akkermans - l.m.a.akkermans@chir.azu.nl; Hein G Gooszen - h.gooszen@chir.azu.nl; the members of the Dutch Acute Pancreatitis Study Group - info@pancreatitis.nl

* Corresponding author

Published: 29 September 2004

BMC Surgery 2004, **4**:12 doi:10.1186/1471-2482-4-12

Received: 21 July 2004
Accepted: 29 September 2004

This article is available from: http://www.biomedcentral.com/1471-2482/4/12

Probiotic prophylaxis in predicted severe acute pancreatitis: a randomised, double-blind, placebo-controlled trial

Marc G H Besselink, Hjalmar C van Santvoort, Erik Buskens, Marja A Boermeester, Harry van Goor, Harro M Timmerman, Vincent B Nieuwenhuijs, Thomas L Bollen, Bert van Ramshorst, Ben J M Witteman, Camiel Rosman, Rutger J Ploeg, Menno A Brink, Alexander F M Schaapherder, Cornelis H C Dejong, Peter J Wahab, Cees J H M van Laarhoven, Erwin van der Harst, Casper H J van Eijck, Miguel A Cuesta, Louis M A Akkermans, Hein G Gooszen, for the Dutch Acute Pancreatitis Study Group

Summary

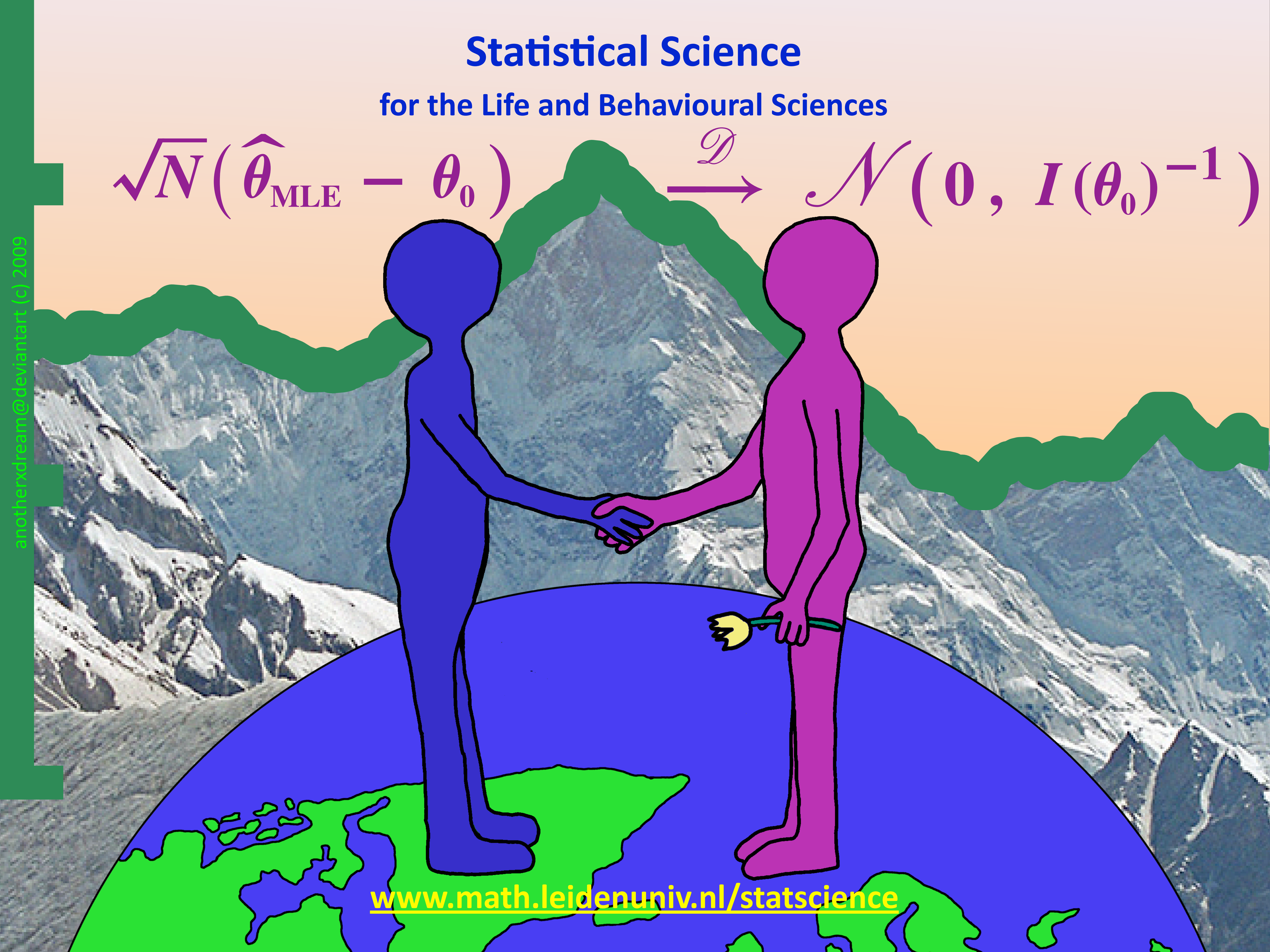
Background Infectious complications and associated mortality are a major concern in acute pancreatitis. Enteral administration of probiotics could prevent infectious complications, but convincing evidence is scarce. Our aim was to assess the effects of probiotic prophylaxis in patients with predicted severe acute pancreatitis.

Methods In this multicentre randomised, double-blind, placebo-controlled trial, 298 patients with predicted severe acute pancreatitis (Acute Physiology and Chronic Health Evaluation [APACHE II] score ≥8, Imrie score ≥3, or C-reactive protein >150 mg/L) were randomly assigned within 72 h of onset of symptoms to receive a multispecies probiotic preparation (n=153) or placebo (n=145), administered enterally twice daily for 28 days. The primary endpoint was the composite of infectious complications—ie, infected pancreatic necrosis, bacteraemia, pneumonia, urosepsis, or infected ascites—during admission and 90-day follow-up. Analyses were by intention to treat. This study is registered, number ISRCTN38327949.

Findings One person in each group was excluded from analyses because of incorrect diagnoses of pancreatitis; thus, 152 individuals in the probiotics group and 144 in the placebo group were analysed. Groups were much the same at baseline in terms of patients' characteristics and disease severity. Infectious complications occurred in 46 (30%) patients in the probiotics group and 41 (28%) of those in the placebo group (relative risk 1·06, 95% CI 0·75–1·51). 24 (16%) patients in the probiotics group died, compared with nine (6%) in the placebo group (relative risk 2·53, 95% CI 1·22–5·25). Nine patients in the probiotics group developed bowel ischaemia (eight with fatal outcome), compared

Published Online
February 14, 2008
DOI:10.1016/S0140-6736(08)60207-X

Department of Surgery (M G H Besselink MD, H C van Santvoort MD, H M Timmerman PhD, Prof L M A Akkermans PhD, Prof H G Gooszen MD) **and Julius Center for Health Sciences and Primary Care** (E Buskens MD), **University Medical Center Utrecht, Utrecht, Netherlands; Department of Epidemiology** (E Buskens) **and Department of Surgery** (V B Nieuwenhuijs MD, Prof R J Ploeg MD), **University Medical Center Groningen, Groningen, Netherlands; Department of Surgery, Academic Medical Center**


Statistical Science
for the Life and Behavioural Sciences
$\sqrt{N}\left(\hat{\theta}_{\mathrm{MLE}} - \theta_0\right) \xrightarrow{\mathscr{D}} \mathcal{N}\left(0,\ I(\theta_0)^{-1}\right)$
www.math.leidenuniv.nl/statscience
anotherxdream@deviantart (c) 2009