How Well-calibrated Should Bayes Procedures Be?

Andy Grieve (Andy.Grieve@UCB.com)
Centre of Excellence for Statistical Innovation
UCB Celltech
Slough, UK.

Long-run behaviour of Bayesian procedures –
Satellite event of CEN-IBS/GMDS 2020 Conference

Inspired by patients.
Driven by science.

Stephen Senn - And thereby hangs a tail
36th Fisher Memorial Lecture, September 2017.

And also, of course, Bayes!

Good

- For 'personal' decision-making
  - Ramsey, De Finetti, Savage, Lindley
    - Involves elicitation problems: O'Hagan
- In pragmatic compromises
  - Good
  - Box (1980)
  - Racine, Grieve, Fluehler, Smith (1986)
- As an aid to thinking
  - The reverse Bayes of Robert Matthews
  - The conditional Bayes approach of Spiegelhalter, Freedman& Parmar JRSSA, 1994 BART

No so Good?

- Bayesian significance tests
- Bayes-factors
- P-values modified to behave like Bayesian tests

(c) Stephen Senn

bayes 22 August 2017

B I also think of it the other way around, searching for objective priors to mirror p-values.

Reply...

Bitter Experience

ASTIN – Acute Stroke Therapy by Inhibition of Neutrophils

A2561002: A double blind, placebo controlled, multi-centre, Bayesian, adaptive design study to assess the dose-response relationship, safety and toleration of UK-279-276 in acute stroke.

G *general*
A *adaptive*
D *dose*
A *allocation*

2 group parallel group design in depression

GADA was run in parallel with a GSD to pilot the dose allocation system.

Bayesian decision rules were chosen to replicate the alpha-spending function.

P(Futility) + P(Efficacy) >1

Berry Consultants
Statistical Innovation

FACTS

Bayesian Research Including Operating Characteristics

Biometrika (1977), **64**, 2, *pp.* 415–8
Printed in Great Britain

A test for normality against symmetric alternatives

BY D. J. SPIEGELHALTER
Department of Statistics and Computer Science, University College London

Biometrika (1980), **67**, 2, *pp.* 493–6
Printed in Great Britain

An omnibus test for normality for small samples

BY D. J. SPIEGELHALTER
Department of Mathematics, University of Nottingham

BIOMETRICS 43, 847–856
December 1987

A Two-Stage Procedure for Bioequivalence Studies

A. Racine-Poon,[1] A. P. Grieve,[1] H. Flühler,[1] and A. F. M. Smith[2]
[1] Mathematical Applications, CIBA-GEIGY AG, CH-4002, Basel, Switzerland

Journal of Biopharmaceutical Statistics, 8(3), 377–390 (1998)

JOINT EQUIVALENCE OF MEANS AND VARIANCES OF TWO POPULATIONS

Andrew P. Grieve

Academic Guidelines for Reporting Bayesian Analyses

ROBUST	BAYESWATCH	BASIS	SAMPL
Prior Distribution Specified Justified Sensitivity analysis Analysis Statistical model Analytical technique Results Central tendency SD or Credible Interval	Intrduction Intervention described Objectives of study Methods Design of Study Statistical model Prior / Loss function? When constructed Prior / Loss descriptions Use of Software MCMC , starting values, run-in, length of runs, convergence, diagnostics Results Interpretation Posterior distribution summarized Sensitivity analysis if alternative priors used	Research Question Statistical model Likelihood, structure, prior & rationale Computation Software - convergence if MCMC, validation, methods for generating posterior summaries Model checks, sensitivity analysis Posterior Distribution Summaries used: i). Mean, std, quintiles ii) posterior shape, (iii) joint posterior for mult comp, (iv) Bayes factors Results of model checks and sensitivity analyses Interpretation of Results Limitation of Analysis	Prior Distribution Specified Justified Sensitivity analysis Analysis Statistical model Analytical technique Software Results Central tendency SD or Credible Interval What's Missing?

What's Missing? - Operating Characteristics

Type I Error, "Power" etc

Guidelines written by Bayesians

Frequentist properties of Bayesian Procedures

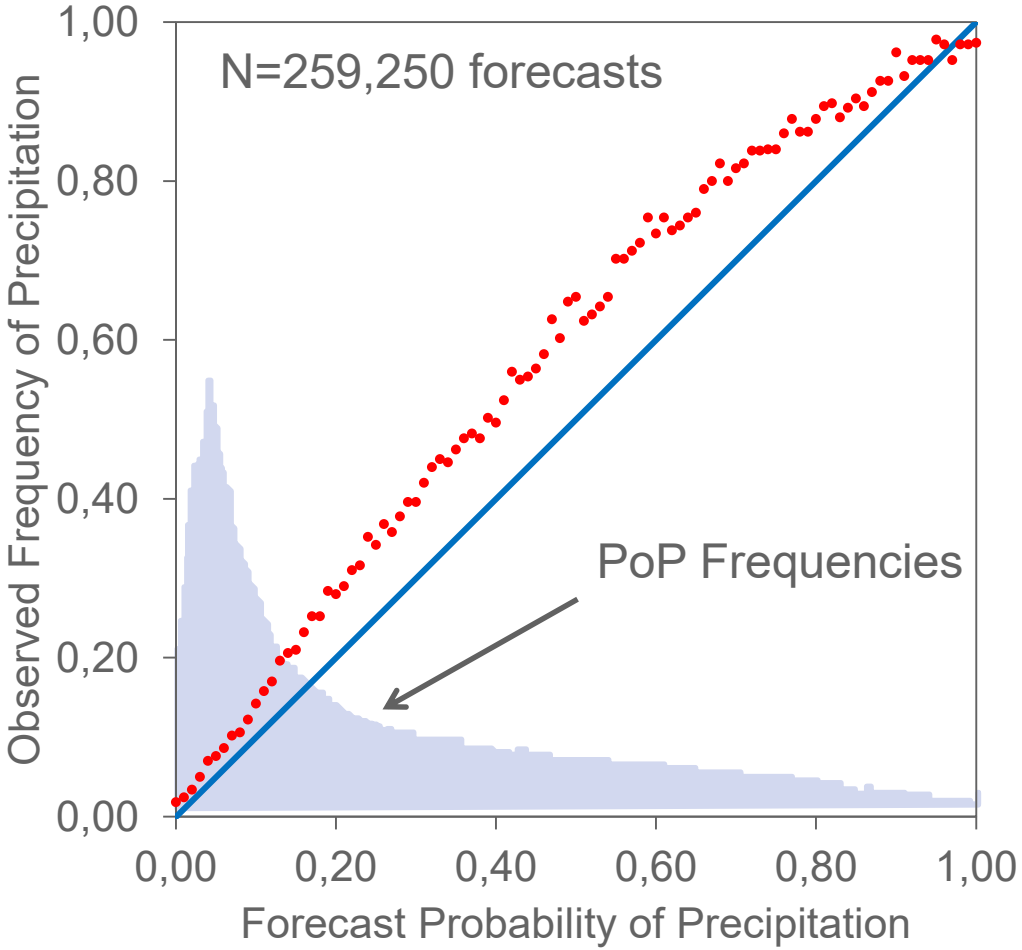
- "Bayesianly Justifiable And Relevant Frequency Calculations For The Applied Statistician" – Don Rubin (1979)

Objective Bayes – Berger & Bernardo (Uniformative)

Calibrated Bayes – Rubin, Lewis & Berry, Spiegelhalter

- Important for pharmaceutical statisticians?

1-Day Ahead Forecasts - Custom Weather



PoP =

Probability
of
Precipitation

Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials – FDA/CDRH 2010

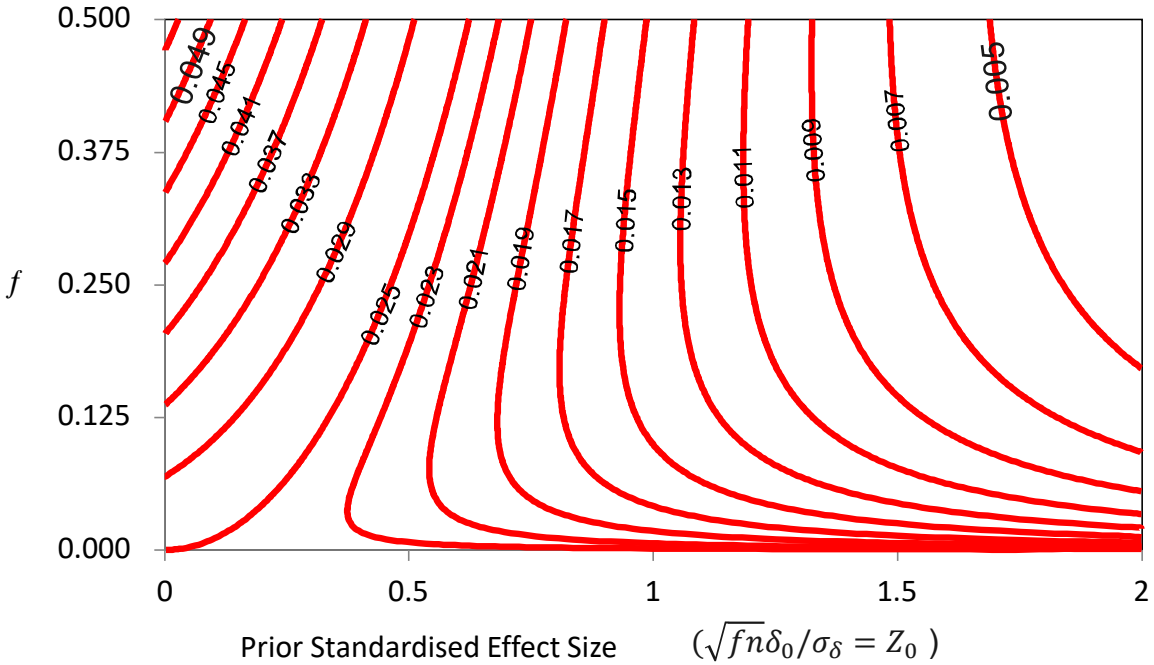
> "Because of the inherent flexibility in the design of a Bayesian clinical trial, a thorough evaluation of the operating characteristics should be part of the trial design. This includes evaluation of:

- probability of erroneously approving an ineffective or unsafe device (type I error)
- probability of erroneously disapproving a safe and effective device (type II error)
- power (the converse of type II error: the probability of appropriately approving a safe and effective device)
- sample size distribution (and expected sample size)
- prior probability of claims for the device
- if applicable, probability of stopping at each interim look. "

Bayesian Analysis of Clinical Trial with Real Prior Evidence

Data	$D \sim N(\delta, \sigma^2/n)$
Prior	$\delta \sim N(\delta_0, \sigma^2/(fn))$
Posterior	$\delta \sim N\left(\frac{nD+fn\delta_0}{n+fn}, \frac{\sigma^2}{n+fn}\right)$
Decision rule	$Prob(\delta > 0 \mid D) > 1-\psi = D > -\frac{\sqrt{1+f}Z_\psi\sigma}{\sqrt{n}} - f\delta_0$
Prob under null	$\Phi\left(\sqrt{1+f}Z_\psi + \frac{f\sqrt{n}\delta_0}{\sigma}\right)$
Control at 2.5%	$Z_{1-\psi} = \frac{Z_{0.975}+\sqrt{f}Z_0}{\sqrt{1+f}}$ $\left(Z_0 = \sqrt{nf}\delta_0/\sigma\right)$

Contours of Bayesian Decision Rule (ψ) to give a One-sided Type I Error of 2.5%



- If the prior standardised effect size is large then ψ must be considerably reduced to control the type I error.
- In contrast, for small Z_0 and large *f*, the nominal level may be relaxed.
- This is intuitively correct because the prior distribution is providing a significant penalty towards zero.
- Substitute $Z_{1-\psi} = \dfrac{Z_{0.975} + \sqrt{f}Z_0}{\sqrt{1+f}}$

  into decision rule $D > -\dfrac{\sqrt{1+f}Z_\psi\sigma}{\sqrt{n}} - f\delta_0$

  to give $D > \dfrac{\sigma Z_{0.975}}{\sqrt{n}}$

Implications

- “requiring strict control of the type-I error results in 100% discounting of the prior information.” (Grieve, Pharm Stats, 2016)
- If we require absolute control of the type I error - “perfectly-calibrated” - then throw away any prior information.
- FDA’s Bayesian guidance for devices - “it may be appropriate to control the type I error at a less stringent level than when no prior information is used”.
- The FDA’s remark is a recognition of the phenomenon and an endorsement of a less strict control of type I error - “well-calibrated”.

Bayesian Adaptive Design with Historical Control Data

- Phase II, randomized, double-blind, active-controlled, adaptive, parallel design.
- 6 treatment arms
  - 5 single doses of Drug X
  - Control: single doses of an active comparator (Historical and Contemporary)
- Acute Treatment duration: minimum of 24 hours or discharge – continuous measure
- Dose Selection: All doses with an mean effect compared to active of > 0.8 units with a given posterior probability
- Prior distribution: based on ~3600 historic controls – discounted to 40
- Interim Analysis
  - Allows testing of assumptions
    - Prior distribution
    - Effect sizes
  - Early stopping for futility
- Randomization
  - Stage 1: 1:1:1:1:1:1 randomisation
  - Stage 2: unequal depending on shape of dose-response curve

Regulatory Agencies Review

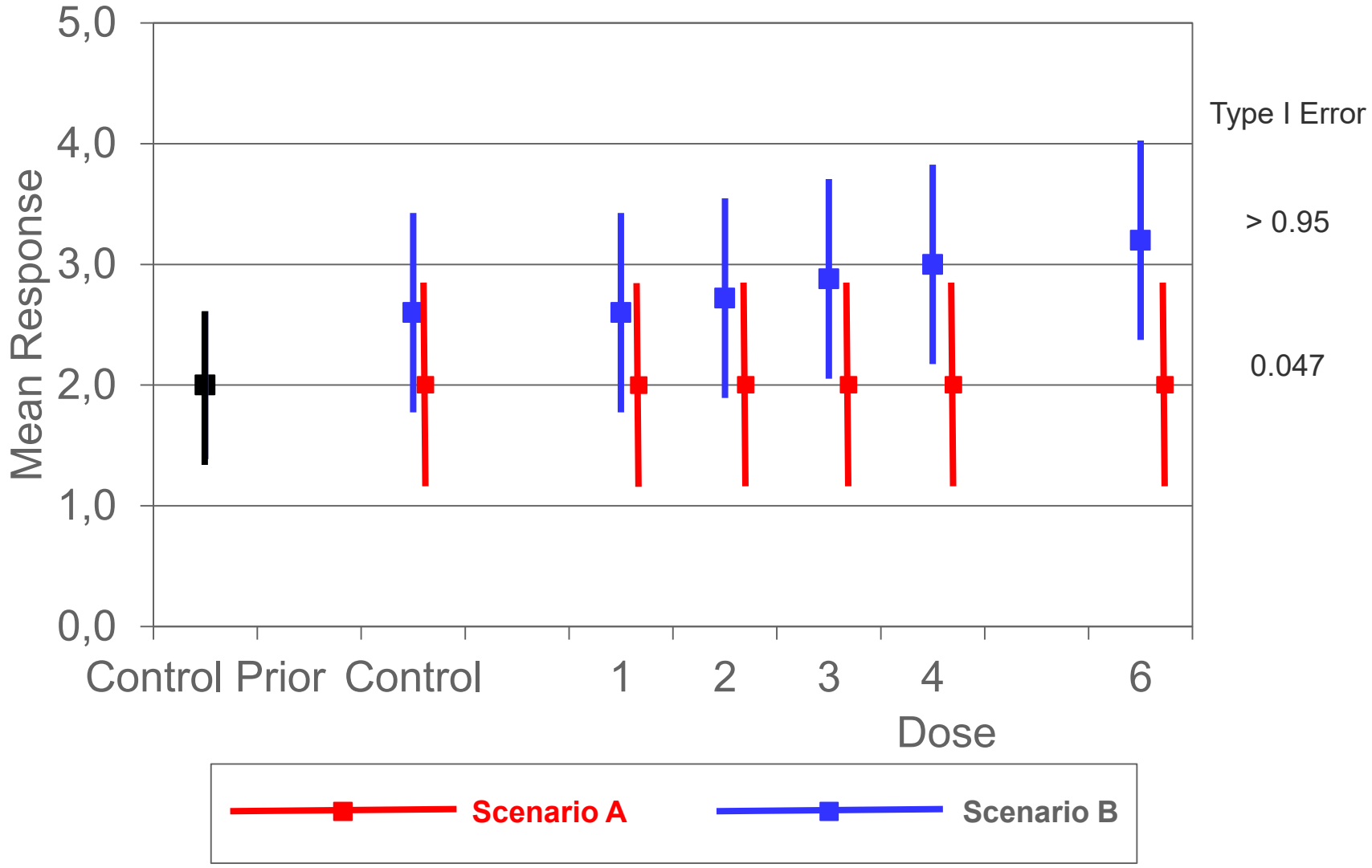
Regulatory Agencies consulted

- FDA,UK,Germany,Poland,Russia,Ukraine.

European agencies raised questions mainly about CMC,QP related and labeling

FDA raised some questions about the prior distribution and its impact. They were not concerned with the adaptive nature od the study.

Specific Null and Alternative Scenarios

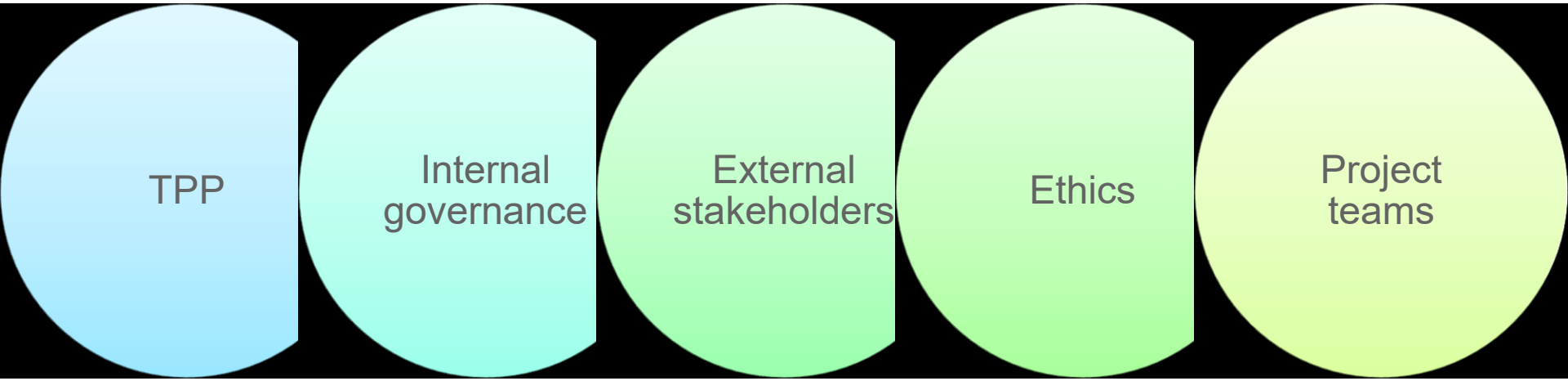


Determining Decision Criteria

Appropriate approach:

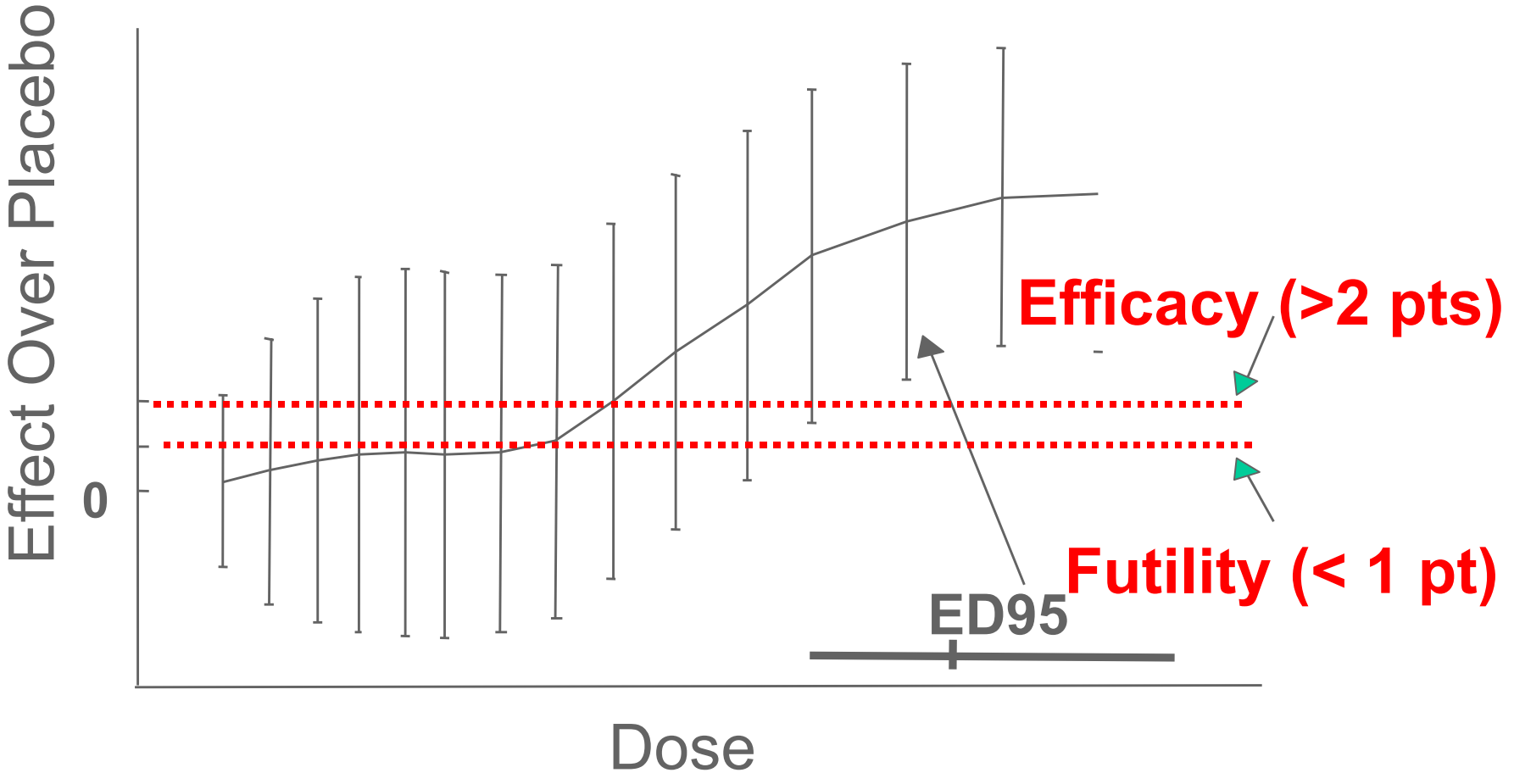
- Choose decision rule based on clinical or commercial criteria.

Who decides what the decision criteria should be?

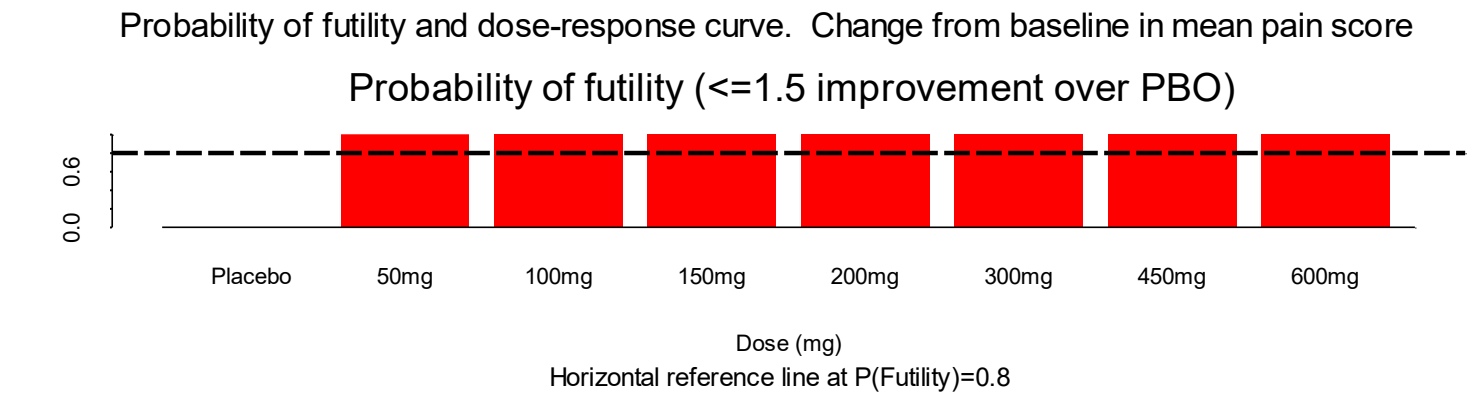


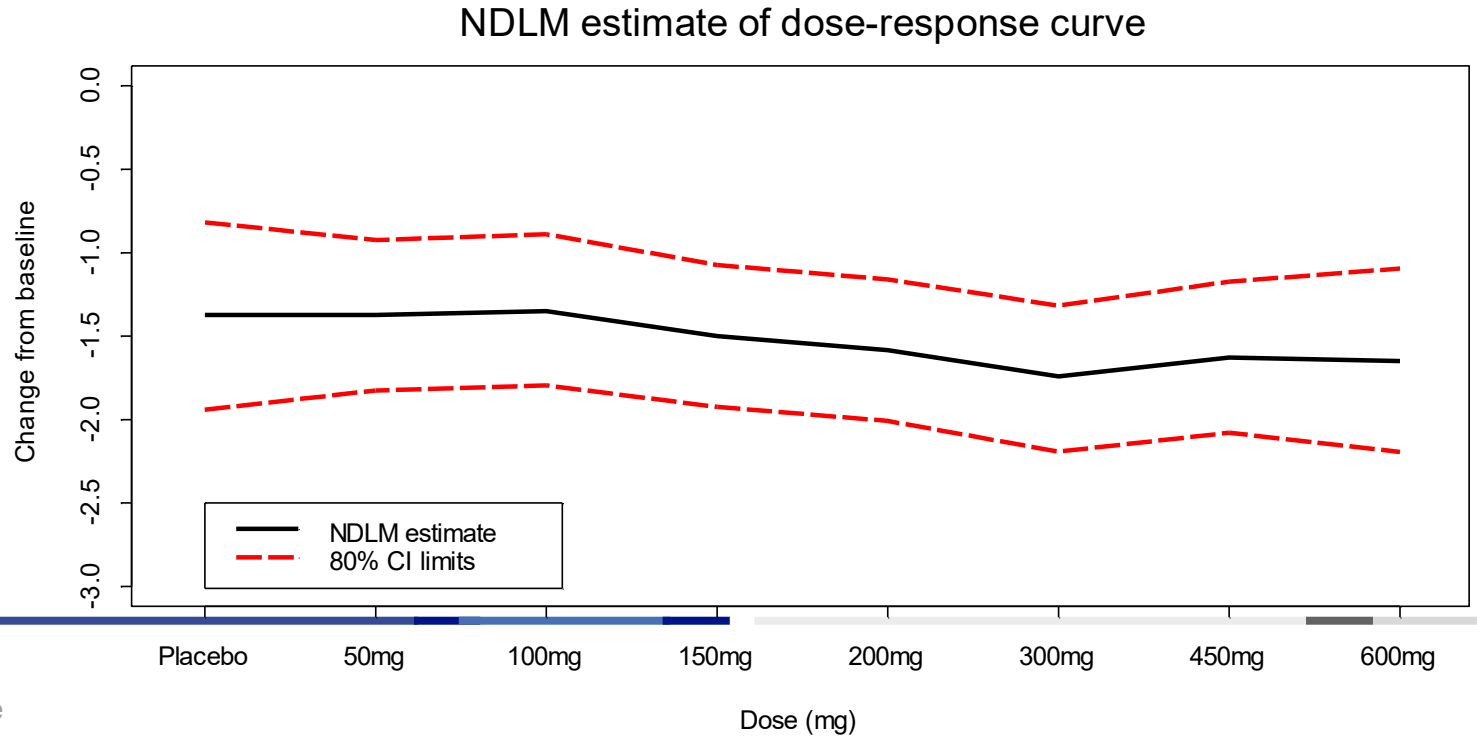
Consult,BUT don’t leave it to the statistician alone!

ASTIN Trial – Acute Stroke: Dose Effect Curve (Grieve and Krams,Clinical Trials,2005)



POC Study in Neuropathic Pain
Smith et al (Pharmaceutical Statistics,2006)





Conclusions: Determining Decision Criteria

Appropriate approach:

- Choose decision rule based on clinical or commercial criteria.
- Investigate operating characteristics
- If they are unacceptable e.g. type I error > 20% then look to change them – “well-calibrated”

Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials – FDA/CDRH 2010

Requires simulations to assess Bayesian approaches.

If type I error too large

- change success criterion (posterior probability)
- reduce number of interim analyses
- discount prior information
- increase sample size
- altering calculation of type I error

"the degree to which we might relax the type I error control is a case-by-case decision that depends …. Primarily on the confidence we have in prior information"

Conclusions: Determining Decision Criteria

Appropriate approach:

- Choose decision rule based on clinical or commercial criteria.
- Investigate operating characteristics
- If they are unacceptable e.g. type I error > 20% then look to change them – “well-calibrated”
- BUT don’t strive to get exact control – “perfectly-calibrated”