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
And also, of course, Bayes!

<table>
<thead>
<tr>
<th>Good</th>
<th>No so Good?</th>
</tr>
</thead>
<tbody>
<tr>
<td>• For ‘personal’ decision-making</td>
<td></td>
</tr>
<tr>
<td>• Ramsey, De Finetti, Savage, Lindley</td>
<td></td>
</tr>
<tr>
<td>• Involves elicitation problems: O’Hagan</td>
<td></td>
</tr>
<tr>
<td>• In pragmatic compromises</td>
<td></td>
</tr>
<tr>
<td>• Good</td>
<td></td>
</tr>
<tr>
<td>• Box (1980)</td>
<td></td>
</tr>
<tr>
<td>• Racine, Grieve, Fluehler, Smith (1986)</td>
<td></td>
</tr>
<tr>
<td>• As an aid to thinking</td>
<td></td>
</tr>
<tr>
<td>• The reverse Bayes of Robert Matthews</td>
<td></td>
</tr>
<tr>
<td>• The conditional Bayes approach of Spiegelhalter, Freedman &amp; Parmar JRSSA, 1994 BART</td>
<td></td>
</tr>
</tbody>
</table>

Bayes 22 August 2017

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

Reply...
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 tolerance 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
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, A. P. Grieve, H. Flühler, and A. F. M. Smith

1 Mathematical Applications, CIBA-GEIGY AG, CH-4002, Basel, Switzerland


JOINT EQUIVALENCE OF MEANS AND VARIANCES OF TWO POPULATIONS

Andrew P. Grieve
# Academic Guidelines for Reporting Bayesian Analyses

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

N=259,250 forecasts

PoP = Probability of Precipitation
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

<table>
<thead>
<tr>
<th>Data</th>
<th>$D \sim N(\delta, \sigma^2/n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior</td>
<td>$\delta \sim N(\delta_0, \sigma^2/(fn))$</td>
</tr>
<tr>
<td>Posterior</td>
<td>$\delta \sim N\left(\frac{nD+fn\delta_0}{n+fn}, \frac{\sigma^2}{n+fn}\right)$</td>
</tr>
<tr>
<td>Decision rule</td>
<td>$\text{Prob}(\delta &gt; 0</td>
</tr>
<tr>
<td>Prob under null</td>
<td>$\Phi\left(\sqrt{1+fnZ}\psi + \frac{fn\delta_0}{\sigma}\right)$</td>
</tr>
<tr>
<td>Control at 2.5%</td>
<td>$Z_{1-\psi} = \frac{Z_{0.975} + \sqrt{fn}Z_0}{\sqrt{1+f}}$ ($Z_0 = \sqrt{nfv\delta_0/\sigma}$)</td>
</tr>
</tbody>
</table>
Contours of Bayesian Decision Rule ($\psi$) to give a One-sided Type I Error of 2.5%

If the prior standardised effect size is large then $\psi$ 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 $Z_{1-\psi} = \frac{Z_{0.975} + \sqrt{f}Z_0}{\sqrt{1 + f}}$

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

to give $D > \frac{\sigma Z_{0.975}}{\sqrt{n}}$
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 a 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 of the study.
Specific Null and Alternative Scenarios

![Graph showing mean response vs dose for different scenarios with control prior and control groups. The graph includes error bars for scenarios A and B with type I error values exceeding 0.95 and 0.047.]
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)

Efficacy (>2 pts)

Futility (< 1 pt)

ED95
POC Study in Neuropathic Pain

Probability of futility and dose-response curve. Change from baseline in mean pain score

Probability of futility (<=1.5 improvement over PBO)

Horizontal reference line at P(Futility)=0.8

NDLM estimate of dose-response curve

NDLM estimate
80% CI limits

Change from baseline in mean pain score

Probability of futility and dose-response curve. Change from baseline in mean pain score
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”
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”