Prospective application of Bayesian monitoring and analysis in an ‘open’ randomized clinical trial

A. Vail, J. Hornbuckle, D. J. Spiegelhalter and J. G. Thornton

1 Biostatistics Group, University of Manchester, Manchester, U.K.
2 Department of Paediatrics, Obstetrics & Gynaecology, University of Leeds, Leeds, U.K.
3 MRC Biostatistics Unit, Cambridge, U.K.

SUMMARY

We describe the prospective application of Bayesian monitoring and analysis in an ongoing large multicentre, randomized trial in which interim results are released to investigators. Substantial variability in prior opinion led us to reject the use of elicited clinical priors for monitoring, in favour of archetypal prior distributions representing reasonable scepticism and enthusiasm. Likelihoods for odds ratios for different covariate values are derived from a logistic regression model, which allows us to incorporate information from prognostic factors without resorting to specialized software. Priors, likelihoods and posterior distributions are regularly reported to both an independent Data Monitoring Committee and the trial investigators. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

There has been increasing interest in medical applications of Bayesian methods from statisticians [1–3] and from within medicine [4, 5]. In setting a framework for the conduct of randomized clinical trials, Spiegelhalter et al. [6] sought to familiarize applied statisticians with potential practical advantages of Bayesian methods both in terms of incorporation of external evidence and judgement, and the flexibility in stopping criteria. In particular, they and others [7] have argued that a suitable criterion for stopping and drawing a positive conclusion may be when a ‘sceptic’ has been convinced by the evidence, and conversely a negative conclusion might be drawn when an ‘enthusiast’ has their opinion changed. In each instance such opinions are formally expressed as an archetypal prior distribution, and changes in opinion derived using Bayes theorem in conjunction with a likelihood based on the trial data.

*Correspondence to: Andy Vail, R&D Directorate, Clinical Sciences Building, Hope Hospital, Stott Lane, Salford M6 8HD, U.K.
†E-mail: andy.vail@man.ac.uk
‡Presented at the International Society for Clinical Biostatistics Twenty-first International Meeting, Trento, Italy, September 2000.

Contract/grant sponsor: Medical Research Council; contract/grant number: G9533539
Contract/grant sponsor: European Union, BioMed Concerted Action; contract/grant number: IC20-CT96-0050

Received January 2001
Accepted October 2001
Here we describe the conduct of an ongoing obstetrics trial under the Bayesian paradigm. Our aims during the trial have been to retain simplicity in both analysis and presentation, so that clinicians understand the approach and are able to interpret the data. Our aim in this paper is that applied statisticians who are currently unfamiliar with Bayesian methods will recognize some of their advantages under a framework that requires minimal retraining from frequentist methods.

After a brief outline of the clinical setting, we describe the elicitation of opinion and the eventual choice of prior distributions for monitoring and analysis. The model underlying the likelihood is demonstrated using interim data on a secondary outcome measure, together with a selection of the derived posterior distributions. Finally, some conclusions for potential uses of Bayesian methodology are drawn.

2. CLINICAL SETTING

The Growth Restriction Intervention Trial began in 1996 and seeks to randomize 600 subjects to immediate or deferred delivery when there is evidence of failure to thrive in utero for the pre-term foetus. Obstetricians routinely use both treatment arms in standard clinical practice. At early gestational ages the risks associated with prematurity are relatively high, so that where compromise appears mild, the obstetrician will defer delivery and continue to monitor the pregnancy. At later gestational ages, the risks associated with even mild signs of foetal compromise may outweigh those of prematurity. Immediate delivery is then usual practice. Table I summarizes the dilemma [8]. Foetal compromise, measured by umbilical artery end diastolic frequency (EDF), is less severe in the lower rows.

As we shall show in the next section, there is substantial clinical disagreement as to appropriate action in these difficult circumstances, and this makes it impossible to select rigid enrolment criteria that would be clinically acceptable to a broad range of obstetricians. Recruitment is thus based solely on the ‘uncertainty principle’ [9], in that a patient is eligible if the individual clinician is uncertain as to the preferred course of action. In order to ensure that this judgement of uncertainty is adequately informed, the trial protocol is unusual in stating that results of the trial are not to be blinded and are to be regularly released to investigators. This allows the possibility of ‘drift’ in the type of patient randomized in the study as collaborators update their opinions, and therefore there needs to be careful investigation of possible interactions between treatment effects and important risk factors.

Table I. Schematic presentation of the clinical question. At early gestational age most clinicians would defer delivery (D) unless foetal compromise were extreme. At late gestational age most would recommend immediate delivery (I) for all but very mild signs of compromise. Uncertainty (?) exists on the leading diagonal.

<table>
<thead>
<tr>
<th>Umbilical artery end diastolic frequency (EDF)</th>
<th>Gestational age (weeks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reversed</td>
<td>24 25 26 27 28 29 30 31 32 33 34 35 36</td>
</tr>
<tr>
<td>Moderately reduced</td>
<td>D D D D D D D D D D ?</td>
</tr>
</tbody>
</table>
Table I shows 52 combinations of gestational age (13 weeks) and EDF (four categories). It is clear that the relative advantage of immediate or deferred delivery is expected to change dramatically over these circumstances. Formally, we might expect a strong interaction between treatment effect and these two prognostic factors. In the following sections we describe how this potential interaction influences both the use of prior opinion and the structure of the likelihood.

3. INVESTIGATORS’ PRIOR OPINIONS

Various elicitation exercises preceding this trial [8, 10] demonstrated major disagreement amongst potential collaborators regarding their beliefs concerning both the absolute and relative risks of immediate and deferred delivery. At subsequent meetings collaborators expressed opinions regarding likely outcomes for one of four preprepared clinical scenarios representing case histories of actual patients falling on the leading diagonal of Table I. Formal elicitation of expert opinion in this context provided a basis for discussion and avoided problems of dominant opinion that may occur in unstructured debate.

Figure 1 shows the variety of opinions expressed by international collaborators for the risk of death in one such clinical scenario: 28 to 29 weeks gestation with absent end diastolic frequency. The method of elicitation was adapted from Gore [11]. We asked collaborators to write their ‘best guess’ as to the risk of death with immediate delivery and with deferred delivery. We then asked them to place 20 crosses in lanes to form a histogram expressing their distribution of uncertainty concerning risk with deferred delivery. Estimates of the risk of death with immediate delivery ranged from 10 per cent to 65 per cent. Three collaborators (1, 2 and 8) failed to provide interpretable data regarding their range of uncertainty for deferred delivery, two forgot to give a scale and one drew a curve rather than marking crosses. Two
(4 and 17) were confident that deferral could be no worse than immediate delivery. Conversely, two (11 and 14) were essentially certain that deferral must be worse than immediate delivery.

This lack of consistent expert opinion may be a more common feature of clinical trials than is generally acknowledged. It has been debated whether the appropriate prerequisite for an ethical clinical trial is that collaborating clinicians have conflicting views (lack of clinical equipoise), or that individual clinicians have no preference for either treatment (the uncertainty principle) [12]. It appears that both criteria may be generally fulfilled in this context, in that for each clinician one can imagine situations in which personal uncertainty may exist, but these situations will not be the same for all clinicians. Such variability in opinions concerning both absolute and relative rates justifies a protocol without rigid entry criteria, but perhaps casts serious doubt on the value of eliciting and using ‘expert’ priors for formal Bayesian analysis.

### 4. MONITORING AND FEEDBACK USING ARCHETYPAL PRIORS

A simple Bayesian procedure would be to consider each clinical scenario for which prior opinions were elicited, derive a likelihood function on a log-odds scale from a $2 \times 2$ table based solely on the data for that scenario, and combine this likelihood with each individual clinical prior to produce the relevant posterior distribution. Normal approximations were reasonable for both the prior and likelihood of the log-odds ratio. We have presented such interim analyses to collaborators as well as to a formal Data Monitoring and Ethics Committee. Graphs such as Figure 2, based on a real opinion and early interim data for the risk of death at 28 or 29 weeks gestation with absent EDF, have been well received by clinicians with no previous experience of Bayesian methods.

This presentation has the advantage of simplicity in explaining the prior-to-posterior principle of Bayesian interpretation. However, there are three immediate drawbacks. First, it would not have been practical to elicit opinions for all 52 combinations of gestational age by EDF category. Second, there is a strongly subjective element in the selection of which elicited priors to present. Third, based on the description above, it is not clear that individual clinical priors are expressing realistic, generalizable information in terms of parameter values for the analysis, and in particular it may be inappropriate to produce a ‘composite’ prior in the light of such diverse opinion.

These drawbacks are substantially addressed by use of archetypal ‘sceptical’ and ‘enthusiastic’ priors [6]. A sceptical prior is centred on an odds ratio of 1, corresponding to no difference in treatments, with standard deviation determined by ensuring that the central 95 per cent of the density lies between two odds ratios representing clinically important and plausible differences. Choice of this standard deviation remains a somewhat arbitrary element of the analysis. Here we have taken odds ratios of 0.5 and 2.0 as plausible and important alternative hypotheses, so our sceptical prior is expressing severe doubt that the true difference could be as extreme as a doubling or halving of the odds. The standard deviation encapsulates the information regarding the strength of scepticism that the trial data will need to overcome. For monitoring purposes, we recommend that this standard deviation is explicit before examination of the trial data. For analysis we recommend use of the sceptical prior when data appear to favour an intervention. Sensitivity analysis may be used to calculate for any combination of patient characteristics the degree of scepticism required (minimum prior standard deviation) in order not to be persuaded by the data. Sceptical priors have also been recommended in
Figure 2. Example ‘prior to posterior’ analysis for the odds ratio of death at 28 or 29 weeks gestation with absent EDF. In this example, the clinician’s prior preference for immediate delivery (— · —) is tempered by equivocal data (– – ). The resulting posterior (——) may give this clinician encouragement for randomization of such future cases.

Figure 3. Archetypal sceptical (—–) and enthusiastic (- - -) priors for the odds ratio.

cumulative meta-analysis [13], sequential phase II studies [14] and in deciding whether to carry out a confirmatory trial [15].

An enthusiastic prior represents an archetypal opinion of belief in the efficacy of one treatment over another. In the context of trialling two established treatments, there are two enthusiastic priors – one enthusiast for immediate delivery and one enthusiast for deferral. These have standard deviations equal to that of the sceptical prior, and are centred at odds ratios of 0.5 and 2.0 respectively (Figure 3), and therefore also express severe doubt (2.5 per cent) that the favoured treatment could actually be inferior.

The sceptical and enthusiastic priors side-step debate on the value of ‘expert’ prior opinion, can be applied to any of the clinical scenarios, and are less subjective than the choice of which clinical priors to present. We apply them uniformly in each of the 52 cells of Table I. By presenting the likelihood at interim analyses, we are essentially also illustrating the use of a ‘reference’ prior which is uniform on the log-odds scale.
5. GENERATING AN APPROPRIATE LIKELIHOOD

The problem remains that the numbers of allocated patients and events of interest are likely to be low for any of the 52 particular clinical scenarios displayed in Table I, and this would lead to the specified priors dominating the data in determination of the posterior distributions. However, if a representative sample of all pregnancies were randomized, we might anticipate trends in the log-odds ratio across both the rows and columns of Table I. That is, the log-odds ratio may be positive (in favour of deferral) for low gestational age and decrease with increasing gestational age, whilst also decreasing with increasing signs of foetal compromise. This suggests smoothing the likelihood over the situations described in the gestational age (GA) by EDF table by fitting a statistical model allowing for interaction between the treatment group and the covariates GA and EDF.

A fully Bayesian model could have been fitted, in which random treatment effects were estimated for each of the 52 combinations of covariates. However, this would have negated our aim of conducting the analysis using standard statistical software, and in any case a much simpler model was found to be adequate.

Using the whole data set, we fit a standard (frequentist) logistic regression including linear main effect terms for compromise (EDF coded 1=reversed to 4=moderately reduced) and gestational age (GA in weeks −24), and the interactions of each with the binary treatment allocation term (TREAT coded 0=defer, 1=immediate). Thus the probability \( p \) of an event is determined by

\[
\logit(p) = b_0 + b_1 \text{TREAT} + b_2 \text{EDF} + b_3 \text{GA} + b_4 \text{TREAT.EDF} + b_5 \text{TREAT.GA}
\]

This provides smoothed estimates of the log-odds ratio \((b_1 + b_4 \times \text{EDF} + b_5 \times \text{GA})\) with associated standard error for all gestational age by EDF combinations.

6. RESULTS FOR A SECONDARY OUTCOME

We present some interim results for a secondary outcome, the requirement for ventilation or intubation (Table II). Publication of primary outcome data could jeopardize eventual clinical publication and is not necessary to demonstrate our method. There is a clear effect of gestational age \((b_3)\) but not of EDF \((b_2)\) on the overall risk – treating EDF as a factor gave a negligible improvement in fit. There is evidence for a deleterious effect of immediate intervention \((b_1)\), but some suggestion that this effect decreases with increasing gestational age.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant ((b_0))</td>
<td>2.68 (0.65)</td>
</tr>
<tr>
<td>\text{TREAT} ((b_1))</td>
<td>1.76 (0.96)</td>
</tr>
<tr>
<td>\text{EDF} ((b_2))</td>
<td>−0.01 (0.20)</td>
</tr>
<tr>
<td>\text{GA} ((b_3))</td>
<td>−0.52 (0.08)</td>
</tr>
<tr>
<td>\text{TREAT.EDF} ((b_4))</td>
<td>0.02 (0.28)</td>
</tr>
<tr>
<td>\text{TREAT.GA} ((b_5))</td>
<td>−0.14 (0.12)</td>
</tr>
</tbody>
</table>
age ($b_3$). Relatively fewer babies in the ‘Defer’ arm of the trial require ventilation or intubation at earlier gestational ages, but there is no evidence that this comparison with the ‘Immediate’ arm is affected by EDF status at randomization. Note that the modelled log-odds ratio exists by extrapolation even where an odds ratio has not been observed (Figure 4). Since the results suggest the superiority of one of the treatments, their influence on the sceptical prior should be considered to see if the evidence is sufficiently strong to overcome the expressed degree of scepticism. The modelled log-odds ratio replaces the observed value

![Figure 4](image-url)

Figure 4. Observed OR with 95 per cent confidence interval (vertical bars) for ventilation or intubation requirement: OR $\geq 1$ corresponds to increased risk with immediate delivery and hence favours deferral. Estimated OR with 95 per cent confidence bands from logistic regression are shown for: (a) reversed EDF; (b) absent EDF; (c) severely reduced EDF; (d) moderately reduced EDF.
in the simpler analysis based on treating each cell of Table I individually. Pointwise application of the sceptical prior produces a smoothed posterior distribution (Figure 5) with ready interpretation for trial monitoring. For cases with reversed EDF (a) or only moderately reduced EDF (d) there is insufficient evidence to persuade a sceptic of an effect despite the modelled likelihoods presented in Figure 4. Where most data are available, corresponding to the central categories of Table I, there may be sufficient evidence to persuade a sceptic that deferral is preferable for categories (b) and (c) (absent or severely reduced EDF). This is only according to this secondary outcome measure. If such a pattern were present for primary outcomes it may be reasonable for investigators to stop randomizing such patients.
Although conventionally seen as a Bayesian process, elicitation of expert prior opinion can serve useful purposes for all trials. It may lead to a consensus view of specific eligibility criteria and outcome measures, and clinicians’ estimates of effect size are helpful in guiding power calculations in determining the target sample size. We have found elicitation exercises to be informative in generating constructive discussion at collaborators’ meetings. However, in the face of substantial clinical disagreement, the value of elicited priors for formal monitoring and reporting within the Bayesian framework may be limited. Giving priority to interpretation

![Graph](https://via.placeholder.com/150)

**Figure 5.** Sceptical posterior mean odds ratio and 95 per cent credibility bands for ventilation or intubation requirement calculated by pointwise updating of the sceptical prior with Figure 4 modelled data for: (a) reversed EDF; (b) absent EDF; (c) severely reduced EDF; (d) moderately reduced EDF.
using archetypal priors allays concerns regarding both subjectivity and the extent to which opinion is genuinely representative.

We adopted a Bayesian approach in this trial for practical rather than philosophical benefits. In a trial where slow recruitment was anticipated to be a major difficulty, the ability to report interim findings has maintained the interest and enthusiasm of collaborators. Such an ‘open’ approach could be adopted under the frequentist paradigm using sequential methods, but its impact on early stopping is unpredictable and hence could influence type I error. This is not a concern within the Bayesian paradigm. We have been able to restrict ourselves to simple methods of analysis using standard software, and the approach has proved popular with our target audience. We feel that, when carried out with care and in parallel with standard presentation of results, a Bayesian approach based on archetypal priors provides valuable insight as to the extent to which clinical trial results should be convincing to a wide variety of clinical opinion.
We have not made any attempt to elicit the utilities of the participants or their demands regarding the benefits of either intervention. While this is formally part of a full Bayesian decision-theoretic approach, we currently feel that it is more appropriate to focus only on inferences regarding the magnitude of benefit. We are therefore avoiding any attempt at answering the crucial open question: to what extent can a clinical trial be treated as a decision problem?

The type of analysis described in this paper may be useful both in publicly funded trials and in the deliberations of regulatory authorities concerning commercial submissions. In each case a full decision-theoretic analysis is currently very difficult. Instead, we recommend a consideration of the persuasiveness of current evidence to a broad range of clinical opinion.

REFERENCES