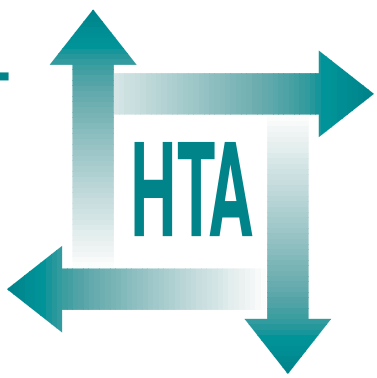
Health Technology Assessment 2000; Vol. 4: No. 38

Methodology

Bayesian methods in health technology assessment: a review

DJ Spiegelhalter
JP Myles
DR Jones
KR Abrams

Health Technology Assessment
NHS R&D HTA Programme



How to obtain copies of this and other HTA Programme reports.
An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (http://www.hta.ac.uk). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents:

– fax (with **credit card** or **official purchase order**)
– post (with **credit card** or **official purchase order** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

Paying by credit card
The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order
You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of *HTA on CD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

Bayesian methods in health technology assessment: a review

DJ Spiegelhalter[1*]
JP Myles[1]
DR Jones[2]
KR Abrams[2]

[1] MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK
[2] Department of Epidemiology and Public Health, University of Leicester, Leicester, UK

[*] Corresponding author

Competing interests: none declared

Published December 2000

This report should be referenced as follows:

Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. *Health Technol Assess* 2000;**4**(38).

Health Technology Assessment is indexed in *Index Medicus*/MEDLINE and *Excerpta Medica*/EMBASE. Copies of the Executive Summaries are available from the NCCHTA website (see opposite).

NHS R&D HTA Programme

The NHS R&D Health Technology Assessment (HTA) Programme was set up in 1993 to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies is produced in the most efficient way for those who use, manage and provide care in the NHS.

Initially, six HTA panels (pharmaceuticals, acute sector, primary and community care, diagnostics and imaging, population screening, methodology) helped to set the research priorities for the HTA Programme. However, during the past few years there have been a number of changes in and around NHS R&D, such as the establishment of the National Institute for Clinical Excellence (NICE) and the creation of three new research programmes: Service Delivery and Organisation (SDO); New and Emerging Applications of Technology (NEAT); and the Methodology Programme.

Although the National Coordinating Centre for Health Technology Assessment (NCCHTA) commissions research on behalf of the Methodology Programme, it is the Methodology Group that now considers and advises the Methodology Programme Director on the best research projects to pursue.

The research reported in this monograph was funded as project number 93/50/05.

The views expressed in this publication are those of the authors and not necessarily those of the Methodology Programme, HTA Programme or the Department of Health. The editors wish to emphasise that funding and publication of this research by the NHS should not be taken as implicit support for any recommendations made by the authors.

Criteria for inclusion in the HTA monograph series
Reports are published in the HTA monograph series if (1) they have resulted from work commissioned for the HTA Programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The editors and publisher have tried to ensure the accuracy of this report but do not accept liability for damages or losses arising from material published in this report. They would like to thank the referees for their constructive comments on the draft document.

ISSN 1366-5278

Published by Core Research, Alton, on behalf of the NCCHTA.
Printed on acid-free paper in the UK by The Basingstoke Press, Basingstoke.

M

Contents

List of abbreviations

List of abbreviations

6-MP	6-mercaptopurine
AMI	acute myocardial infarction
CALGB	Cancer and Leukaemia Group B
CDRH	Center for Devices and Radiological Health
CHART	continuous hyperfractionated accelerated radiotherapy
CMT	conventional medical treatment
CRM	continuous reassessment method
DI	donor insemination
DMC	Data Monitoring Committee
ECMO	extracorporeal membrane oxygenation
ENBS	expected net benefit from sampling
EVPI	expected value of perfect information
EVSI	expected value of sample information
FDA	Food and Drugs Administration
GREAT	Grampian Region Early Anistreplase Trial
GUSTO	Global Utilization of Streptokinase and t-PA for Occluded Coronary Arteries
HFEA	Human Fertility and Embryology Authority
HIB	*Haemophilus influenza*e type b
ISIS-4	Fourth International Study of Infarct Survival
IVF	*in vitro* fertilisation
LIMIT-2	Second Leicester Intravenous Magnesium Intervention Trial
MCMC	Markov chain Monte Carlo
MSUD	maple syrup urine disease
NCI	National Cancer Institute
NSABP	National Surgical Adjuvant Breast and Bowel Project
PORT	Patient Outcomes Research Team
SPPM	Stroke Prevention Policy Model
TRACE	Trandolapril Cardiac Evaluation

Executive summary

Background

Bayesian methods may be defined as the **explicit quantitative use of external evidence in the design, monitoring, analysis, interpretation and reporting of a health technology assessment**. In outline, the methods involve formal combination through the use of Bayes's theorem of:

1. a prior distribution or belief about the value of a quantity of interest (for example, a treatment effect) based on evidence not derived from the study under analysis, with
2. a summary of the information concerning the same quantity available from the data collected in the study (known as the likelihood), to yield
3. an updated or posterior distribution of the quantity of interest.

These methods thus directly address the question of how new evidence should change what we currently believe. They extend naturally into making predictions, synthesising evidence from multiple sources, and designing studies: in addition, if we are willing to quantify the value of different consequences as a 'loss function', Bayesian methods extend into a full decision-theoretic approach to study design, monitoring and eventual policy decision-making. Nonetheless, Bayesian methods are a controversial topic in that they may involve the explicit use of subjective judgements in what is conventionally supposed to be a rigorous scientific exercise.

Objectives

This report is intended to provide:

1. a brief review of the essential ideas of Bayesian analysis
2. a full structured review of applications of Bayesian methods to randomised controlled trials, observational studies, and the synthesis of evidence, in a form which should be reasonably straightforward to update
3. a critical commentary on similarities and differences between Bayesian and conventional approaches
4. criteria for assessing the reporting of a Bayesian analysis
5. a comprehensive list of published 'three-star' examples, in which a proper prior distribution has been used for the quantity of primary interest
6. tutorial case studies of a variety of types
7. recommendations on how Bayesian methods and approaches may be assimilated into health technology assessments in a variety of contexts and by a variety of participants in the research process.

Methods

The BIDS ISI database was searched using the terms 'Bayes' or 'Bayesian'. This yielded almost 4000 papers published in the period 1990–98. All resultant abstracts were reviewed for relevance to health technology assessment; about 250 were so identified, and used as the basis for forward and backward searches. In addition EMBASE and MEDLINE databases were searched, along with websites of prominent authors, and available personal collections of references, finally yielding nearly 500 relevant references. A comprehensive review of all references describing use of 'proper' Bayesian methods in health technology assessment (those which update an informative prior distribution through the use of Bayes's theorem) has been attempted, and around 30 such papers are reported in structured form. There has been very limited use of proper Bayesian methods in practice, and relevant studies appear to be relatively easily identified.

Results

Bayesian methods in the health technology assessment context

1. Different contexts may demand different statistical approaches. Prior opinions are most valuable when the assessment forms part of a series of similar studies. A decision-theoretic approach may be appropriate where the consequences of a study are reasonably predictable.

2. The prior distribution is important and not unique, and so a range of options should be examined in a sensitivity analysis. Bayesian methods are best seen as a transformation from initial to final opinion, rather than providing a single 'correct' inference.
3. The use of a prior is based on judgement, and hence a degree of subjectivity cannot be avoided. However, subjective priors tend to show predictable biasses, and archetypal priors may be useful for identifying a reasonable range of prior opinion. For a prior to be taken seriously, its evidential basis must be explicitly given.
4. The Bayesian approach provides a framework for considering the ethics of randomisation.
5. Monitoring trials with sceptical and other priors may provide a unified approach to assessing whether the results of a trial should be convincing to a wide range of reasonable opinion, and could provide a formal tool for data-monitoring committees.
6. In contrast to earlier phases of development, it is generally unrealistic to formulate a Phase III trial as a decision problem, except in circumstances where future treatments can be accurately predicted.
7. Observational data will generally require more complex analysis: the explicit modelling of potential biasses may be widely applicable but needs some evidence-base in order to be convincing.
8. A unified Bayesian approach is applicable to a wide range of problems concerned with evidence synthesis, for example in pooling studies of differing designs in the assessment medical devices.
9. Priors for the degree of 'similarity' between alternative designs can be empirically informed by studies comparing the results of randomised controlled trials and observational data.
10. Increased attention to pharmaco-economics should lead to further investigation of decision-theoretic models for research planning, although this will not be straightforward.
11. Regulatory agencies are acknowledging Bayesian methods and have not ruled out their use, and the regulation of medical devices is leading the way in establishing the role of evidence synthesis.
12. 'Comprehensive decision modelling' is likely to become increasingly important in policy making.
13. The BayesWatch criteria described in this report may provide a basis for structured reporting of Bayesian analysis.
14. Summaries of fully fledged ('three-star') applications of Bayesian methods in health technology assessment contain few prospective analyses but provide useful guidance.
15. Four case studies show:
    a. Bayesian analyses using a sceptical prior can be useful to the data-monitoring committee of a cancer clinical trial.
    b. Bayesian methods can be used to temper overoptimistic conclusions based on meta-analysis of small trials.
    c. Modern graphical software can easily handle complex assessments previously analysed using the 'confidence profile' method.
    d. Bayesian methods provide a flexible tool for performance estimation and ranking of institutions.

Recommendations and implications for future research and development

Bayesian methods could be of great value within health technology assessment, but for a realistic appraisal of the methodology, it is necessary to distinguish the roles and requirements for five main participant groups in health technology assessment: methodological researchers, sponsors, investigators, reviewers and consumers. Two common themes for all participants can immediately be identified. First, the need for an extended set of case studies showing practical aspects of the Bayesian approach, in particular for prediction and handling multiple substudies, in which mathematical details are minimised but details of implementation are provided. Second, the development of standards for the performance and reporting of Bayesian analyses, possibly derived from the BayesWatch checklist.

Some specific potential areas of research and development include:

1. **Design**. Realistic development of payback models and consideration of 'open' studies.
2. **Priors**. Investigation of evidence-based prior distributions appropriate to the participant group, as well as reasonable default priors in non-standard situations.
3. **Modelling.** Efficient use of all available evidence by appropriate joint modelling of historical controls, related studies, and so on.
4. **Reporting**. Development of criteria along the lines of the BayesWatch checklist, so that future users can reproduce analyses.
5. **Decision-making**. Increased integration with a health-economic and policy perspective, together with flexible tools for implementation.

Chapter 1
Introduction

What are Bayesian methods?

Bayesian statistics began with a posthumous publication in 1763 by Thomas Bayes,[36] a non-conformist minister from Tunbridge Wells.[237] His work was formalised as **Bayes's theorem,** which, when expressed mathematically, is a simple and uncontroversial result in probability theory. However, certain specific uses of the theorem have been the subject of continued controversy for over a century,[128,170] giving rise to a steady stream of polemical arguments in a number of disciplines. In recent years a more balanced and pragmatic perspective has developed.

The basic idea of Bayesian analysis can be illustrated by a simple example and, although we shall try to keep mathematical notation to a minimum in this review, it will be very helpful if we are allowed one Greek letter θ (theta), to denote a currently unknown quantity of primary interest. Suppose our quantity θ is the median life-years gained by using an innovative rather than a standard therapy on a defined group of patients. A clinical trial is carried out, following which conventional statistical analysis of the results would typically produce a *P* value, an estimate and a confidence interval as summaries of what this particular trial tells us about θ. A Bayesian analysis supplements this by focusing on the question 'How should this trial change our opinion about θ?' This perspective forces the analyst to explicitly state

- a reasonable opinion concerning θ **excluding** the evidence from the trial (known as the prior distribution)
- the support for different values of θ based **solely** on data from the trial (known as the likelihood)

and to combine these two sources to produce

- a final opinion about θ (known as the posterior distribution).

The final combination is done using Bayes's theorem.

What, then, is a Bayesian approach to health technology assessment? We have defined it [415] as "**the explicit quantitative use of external evidence in the design, monitoring, analysis, interpretation and reporting of a health technology assessment**".

Reasons for conducting a review

Much of the standard statistical methodology used in health technology assessment revolves around that for the classical randomised controlled trial: these include power calculations at the design stage, methods for controlling type I error within sequential monitoring, calculation of *P* values and confidence intervals at the final analysis, and meta-analytic techniques for pooling the results of multiple studies. Such methods have served the medical research community well.

The increasing sophistication of health technology assessment studies is, however, highlighting the limitations of these traditional methods. For example, when carrying out a clinical trial, the many sources of evidence and judgement available before a trial may be inadequately summarised by a single 'alternative hypothesis', monitoring may be complicated by simultaneous publication of related studies, and multiple subgroups may need to be analysed and reported. Evidence from multiple sources may need to be combined in order to inform a policy decision, such as embarking or continuing on a research programme, regulatory approval of a drug or device, or recommendation of a treatment at an individual or population level. Standard statistical methods are designed for single studies, and have difficulties dealing with this pervading complexity.

A Bayesian perspective leads to an approach to clinical trials and observational studies that is claimed to be more flexible and ethical than traditional methods,[262] and to elegant ways of handling multiple substudies, for example when simultaneously estimating the effects of a treatment on many subgroups.[72] Proponents have also argued that a Bayesian approach enables one to provide conclusions in a suitable form for making decisions: whether for specific patients, for planning research, or for public policy.[299]

The increasing interest in the Bayesian approach is reflected both in the medical and statistical

literature and in the popular scientific press.[318] Pharmaceutical companies are beginning to express an interest, possibly helped by the recent international regulatory authority statistical guidelines[248] explicitly mentioning the possibility of a Bayesian analysis. However, many outstanding questions remain: in particular, to what extent will the scientific community, or the regulatory authorities, allow the explicit introduction of evidence that is not totally derived from observed data, or the formal pooling of data from studies of differing designs?

Objectives

This report is intended to provide:

1. a brief review of the essential ideas of Bayesian analysis
2. a full structured review of applications of Bayesian methods to randomised controlled trials, observational studies and the synthesis of evidence, in a form which should be reasonably straightforward to update
3. a critical commentary on similarities and differences between Bayesian and conventional approaches
4. criteria for assessing the reporting of a Bayesian analysis
5. a comprehensive list of published 'three-star' examples (see the following section for a definition of this term), in which a proper prior distribution has been used for the quantity of primary interest
6. tutorial case studies of a variety of types
7. recommendations on how Bayesian methods and approaches may be assimilated into health technology assessments in a variety of contexts and by a variety of participants in the research process.

Review methodology

What do we mean by a 'systematic' review?

In common with all such methodological reviews, it is essential to define what we mean by 'systematic'. We have identified three levels of review:

Comprehensive. This seeks to identify all relevant references, and has only been attempted for what we have termed 'three-star' Bayesian health technology assessment studies, which we define as those

1. intending to confirm the value of a technology
2. using an informative, carefully considered prior distribution for the primary quantity of interest
3. updating, or planning to update, this prior distribution by Bayes's theorem.

Such three-star studies have been summarised according to a standard pro forma, and are reported in appendix 1. We note that we do not require such studies to be prospective, in that currently most Bayesian examples are re-analyses of previous studies.

The following are therefore not considered as three-star studies:

1. exploratory or Phase II studies
2. those using a 'minimally informative' or reference prior, or only using an informative prior for a nuisance parameter such as between-study heterogeneity
3. decision analyses in which expected utilities are assessed without any updating of beliefs using Bayes's theorem.

Systematic. This is based on a structured search and reporting of the literature, and covers many areas which are not comprehensively reviewed, such as Bayesian analyses using reference priors and Phase I and Phase II studies. An attempt has been made to identify the majority of the relevant literature.

Peripheral. Minimal references are provided on topics such as using Bayes's theorem for prognostic or diagnostic statements, 'empirical Bayes' analysis which uses elements of Bayesian modelling without giving a Bayesian interpretation to the conclusions, preclinical work, pharmacokinetics, decision analysis, and descriptive studies of clinician's personal beliefs.

We emphasise that our definition of 'Bayesian' may be more restrictive than that of other researchers who may, for example, place a much higher emphasis on decision-making using subjective probabilities, without necessarily requiring the use of Bayes's theorem.

The search procedure

A Bayesian approach can be applied to many scientific issues, and a search of the BIDS ISI database using the term 'Bayes' or 'Bayesian' yielded nearly 4000 papers over the period 1990–98. All of the abstracts were handsearched for relevant material, and about 300 of these were relevant to health technology assessment. These were used as a

source for forward and backward searches, and further techniques included searching other databases (EMBASE and MEDLINE), personal collections, handsearching recent journals, and internet searches of prominent authors.

Since it is very difficult to identify appropriate health technology assessment literature from keywords, we would recommend anyone conducting a search for Bayesian methods in health technology assessment to use 'bayes' and 'bayesian' in all searches and then view all abstracts.

We identified about 450 relevant papers, including around 30 reports of studies taking a 'three-star' Bayesian perspective. The published studies are dispersed throughout the literature, apart from one recent collection of papers,[45] and the only general textbook which might be considered as Bayesian health technology assessment is on the confidence profile approach.[150]

It will be clear that the studies are mainly demonstrations of the approach rather than complete assessments, and in spite of numerous articles promoting the use of Bayesian methods the practical take-up seems very low, although increasing. Possible reasons for this will be discussed in our review. There is also a preponderance of articles in the literature on methodology for clinical trials, and an apparent lack of articles on the more complex issues of synthesising data from studies of different designs, in spite of this being an area where Bayesian methods may have much to offer.

Structure of the review

- Chapter 2 briefly reviews the 'classical' statistical approach to health technology assessment, and then outlines the main features of the Bayesian philosophy, including the subjective interpretation of probability, the relation to diagnostic testing, predictions, decision-making and design. There is a brief description of computational methods, complex Bayesian modelling, and the theoretical justification for the approach. Schools of Bayesians are identified, and a commentary attempts to sort out the major ideological issues.

- Chapter 3 deals in detail with the possible sources of prior distributions and their possible criticism in the light of data, and introduces the concept of exchangeability and its relation to hierarchical or multilevel prior distributions.

- Chapter 4 attempts to structure the large literature on Bayesian approaches to all aspects of randomised controlled trials, including sequential analysis, reporting, cost-effectiveness analysis and the stages of drug development.

- Chapter 5 covers observational studies, such as case–control designs, the use of historical controls, and non-randomised comparisons of institutions.

- Chapter 6 considers meta-analysis and its generalisations, in which evidence from multiple studies, possibly of different designs, is pooled using a statistical model.

- Chapter 7 examines how Bayesian analyses for randomised or non-randomised studies may be placed in a concrete decision-making context in order to inform either commercial or public policies, possibly with explicit costs on the consequences of alternative strategies. The view of alternative 'actors' is emphasised.

- Chapter 8 discusses the reporting of Bayesian studies, sets out criteria for assessing the quality of a Bayesian analysis, and provides an example.

- Chapter 9 is a case study in which a sequential cancer clinical trial, the continuous hyperfractioned accelerated radiotherapy (CHART) study, was monitored using a Bayesian procedure.

- Chapter 10 is a case study concerning the much-studied issue of magnesium for acute myocardial infarction (AMI), in which a meta-analysis conflicted with a mega-trial. We show that a reasonably sceptical Bayesian meta-analysis would not have found the initial meta-analysis convincing evidence.

- Chapter 11 shows by four small case studies how modern Bayesian software can deal with the complex modelling problems previously analysed using the confidence profile approach.

- Chapter 12 provides an example of a institutional comparison, in which the success rates of UK *in vitro* fertilisation (IVF) clinics is compared.

- Chapter 13 provides a final summary, general discussion and some suggestions for future research.

- Appendix 1 summarises and lists the 'three-star' applications in a structured format, using the criteria outlined in chapter 9.

- Appendix 2 briefly describes available software and internet sites of interest.

Most of the chapters in the review finish with a critical commentary, in which the arguments against the Bayesian perspective are summarised, and a list of key points for each chapter are repeated in chapter 13. Mathematical and computational methods will be barely mentioned, and details should be sought in the references provided. The review is therefore structured by context rather than methods, although some methodological themes inevitably run throughout; for example, what form of prior distribution is appropriate, and is it reasonable to adopt an explicit loss function?

Finally, we should be quite explicit as to our own subjective biases, which will doubtless be apparent from the organisation and text of this review. We favour the Bayesian philosophy, and would like to see its use extended in health technology assessment, but feel that this should be carried out cautiously, with critical appraisal, and in parallel with the currently accepted methods.

Key points

1. Bayesian methods are defined as the **explicit quantitative use of external evidence in the design, monitoring, analysis, interpretation and reporting of a health technology assessment**.

2. Bayesian methods are a controversial topic in that they may involve the explicit use of subjective judgements in what is conventionally supposed to be a rigorous scientific exercise in health technology assessment.

3. There has been very limited use of proper Bayesian methods in practice, and relevant studies appear to be relatively easily identified.

4. The potential importance of Bayesian methods to a topic is not necessarily reflected in the volume of published literature: in particular, publications on the design and analysis of single clinical trials dominate those on the synthesis of evidence from studies of multiple designs.

Chapter 2

An overview of the Bayesian philosophy in the context of health technology assessment

In this chapter we give an overview of some of the generic features of the Bayesian philosophy that find application in health technology assessment. Limited references to the literature are given at this stage, but relevant sections within randomised trials, observational studies and evidence synthesis are identified. We shall use a simple running example to illustrate the general issues: the Grampian Region Early Anistreplase Trial (GREAT)[204] of early thrombolytic treatment for myocardial infarction, which reported a 49% reduction in mortality (23/148 deaths on control versus 13/163 deaths on active treatment).

The 'classical' statistical approach in health technology assessment

It would be misleading to dichotomise statistical methods as either 'classical' or 'Bayesian', since both terms cover a bewildering range of techniques. It is a little more fair to divide conventional statistics into two broad schools, Fisherian and Neyman–Pearson; different Bayesian approaches will be discussed later in this chapter.

- The **Fisherian** approach to inference on an unknown intervention effect θ is based on the likelihood function mentioned previously, which expresses the relative support given to the different values of θ by the data. This gives rise to an estimate comprising the 'most-likely' value for θ, intervals based on the range of values of θ most supported by the data, and the evidence against specified null hypotheses summarised by *P* values (the chance of getting a result as extreme as that observed were the null hypothesis true).

- The **Neyman–Pearson** approach is focused on the chances of making various types of error so that, for example, clinical trials are designed to have a fixed type I error α (the chance of incorrectly rejecting the null hypothesis), usually taken as 5 or 1%, and fixed power (one minus the type II error β, the chance of not detecting the alternative hypothesis), often 80 or 90%. The situation is made more complex if a sequential design is used, in which the data are periodically analysed and the trial stopped if sufficiently convincing results obtained. Repeated analysis of the data has a strong effect on the type I error, since there are many opportunities to obtain a false-positive result, and thus the *P* value and the confidence interval need adjusting[477] (although this rarely appears to be carried out in the published report of the trial). Such sequential analysis is just one example of a problem of 'multiplicity', in which adjustments need to be made due to multiple analyses being carried out simultaneously. A standard example is the use of Bonferroni adjustments when estimating treatment effects in multiple subsets.

Clinical trials are generally designed from a Neyman–Pearson standpoint, but analysed from a Fisherian perspective.[398] Methods used for observational methods and evidence synthesis tend to be more Fisherian.

Advantages of the traditional framework include its apparent separation of the evidence in the data from subjective factors, the general ease in computation, its wide acceptability and established criteria for 'significance', its relevance to the drug regulatory framework in which quality control of statistical submissions must be ensured, the availability of software, and the existence of robust non- and semi-parametric procedures.

Critique of the classical approach

Hypothesis testing and *P* values

Overemphasis on hypothesis testing has been strongly criticised (although shifting attention to confidence intervals does not avoid all the problems, since these are just the set of hypotheses that cannot be rejected at a certain α level). *P* values are explicitly concerned with the chance of observing the data (or something more extreme) given certain values of the unknown quantity θ, and use an inverse (and frequently misunderstood) argument for deriving statements about θ. Arguments against this procedure include: the null hypothesis may be neither plausible nor of great interest, the arbitrariness of the 0.05 and 0.01 level,

TABLE 1 *Four theoretical studies all with the same* P *value*

Number of patients receiving A and B	Proportion preferring A	*P* value
20	15:5	0.04
200	115:86	0.04
2,000	1,046:954	0.04
2,000,000	1,001,445:998,555	0.04

the focus on statistical rather than clinical significance, the problem over one- or two-sided tests, the fact that even in some simple circumstances, such as a 2 × 2 table, the definition of the *P* value is unclear, and that *P* values tend to create a false dichotomy between 'significant' and 'non-significant' which is inappropriate for consequent policy decisions.[299] See Schervish[392] for a general discussion.

Furthermore, Freeman[184] gives a good example of the limitations of *P* values as expressions of evidence: *Table 1* shows the results of four hypothetical trials in which equal number of patients are given treatments A and B and asked which they prefer, each resulting in an identical 'significant' *P* value of 0.04.

But, as Freeman states, the first trial would be considered too small to permit reliable conclusions, while the last trial (with a preference proportion of 50.07%) would be considered as evidence **for** rather than **against** equivalence. The importance of sample size and plausibility of benefits in interpreting *P* values has often been stressed, and the Fourth International Study of Infarct Survival (ISIS-4) investigators state that "when moderate benefits or negligibly small benefits are both much more plausible than extreme benefits, then a $2p = 0.001$ effect in a large trial or overview would provide much stronger evidence of benefit than the same significance level in a small trial, a small overview, or a small subgroup analysis".[109] Sheiner[400] provides a strong polemic against hypothesis testing and in favour of an approach in which "we gather data to model and quantify nature".

Type I and type II error

Both Bayesians and Fisherians can express strong criticism of Neyman–Pearson theory. Anscombe,[13] quoted by Herson,[232] says "the concept of error probabilities of the first and second kinds … has no direct relevance to experimentation. The formalism of opinions, decisions concerning further experimentation and other required actions, are not dictated in a simple prearranged way by the formal analysis of the experiment, but call for judgement and imagination", while Healy[224] asks "Why the invariable 5% for α? Conditional on this, why the larger 10% or even 20% for β? Is it really more important not to make a fool of yourself than it is to discover something new?". Criticism of fixed type I error has particularly been aimed at sequential analysis (see pages 10 and 29), and relevance of hypothesis testing and decision-making to health technology assessment will be a running theme of this review (see page 14).

Multiple testing

We have already identified the crucial issue that arises in any context in which simultaneous analysis of multiple studies, or multiple analyses of the same study, is required. The traditional approach warns that repeated hypothesis testing is bound to raise the chance of a type I error (wrongly rejecting a true null hypothesis), and so suggests some adjustment, such as Bonferroni, to try to retain a specified overall type I error. This will typically give larger *P* values and wider confidence intervals: it has been shown that such results would be consistent with a rather odd prior distribution in which a constant probability is given to all the null hypotheses being true, regardless of their number.[472]

The need for any such adjustment, which necessarily depends on the number of hypotheses being tested, has been strongly questioned from a non-Bayesian perspective, particularly in epidemiology;[356,382] Cole[108] states that "in every study, every association should be evaluated on its own merits: its prior credibility and its features in the study at hand. The number of other variables is irrelevant", while Cook and Farewell[110] say that it is generally reasonable to report unadjusted conclusions. However, Greenland and Robins[207] are among the many who have argued that some adjustment **is** necessary, but rather than be based on type I errors it should be derived from an explicit model that reflects assumptions about variability. We shall return to this theme on page 12.

Other criticisms of conventional statistical analysis include that it fails to incorporate formally the inevitable background information that is available both at design and analysis and, from a more ideological perspective, that it disobeys certain reasonable axioms of rational behaviour (see page 13). Finally, there is no doubt that classical inferences are often inappropriately interpreted in a Bayesian way, in that *P* values are mistaken for probabilities of null hypotheses being true,

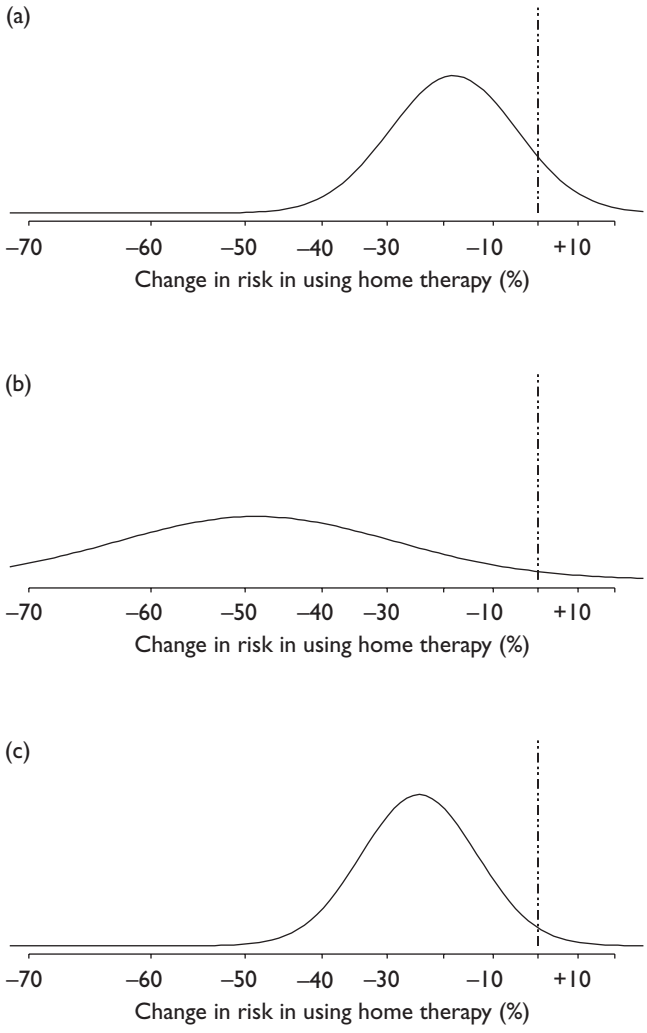


FIGURE 1 *(a) Prior, (b) likelihood (based on 23/148 versus 13/163 deaths) and (c) posterior distributions arising from the GREAT trial of home thrombolysis. (Reproduced by permission of the BMJ from Spiegelhalter* et al.[426]*)*

and 95% confidence intervals as meaning there is a 95% chance of containing the true value.[184]

Bayes's theorem

Suppose θ is some quantity that is currently unknown, for example a specific patient's true diagnosis or the true success rate of a new therapy, and let $p(\theta)$ denote the probability of each possible value of θ (where for the moment we do not concern ourselves with the source of that probability). Suppose we have some observed evidence y whose likelihood of occurrence depends on θ, for example a diagnostic test or the results of a clinical trial. This dependence is formalised by a probability $p(y|\theta)$, which is the (conditional) probability of y for each possible value of θ. We would like to obtain the new probability for different values of θ, taking account of the evidence y; this probability has the conditioning reversed, and is denoted $p(\theta|y)$. Bayes's theorem simply says

$$p(\theta|y) \propto p(y|\theta) \times p(\theta)$$

(The proportionality is made into an equality by making probabilities for all possible values of θ add to 1.) The usual term for $p(\theta)$ is the **prior**, for $p(y|\theta)$ the **likelihood**, and for $p(\theta|y)$ the **posterior**, and hence Bayes's theorem simply says that the posterior distribution is proportional to the product of the prior times the likelihood.

Example

Pocock and Spiegelhalter[363] discuss the GREAT trial, in which the unknown quantity θ is the true percentage change in risk of mortality from using home thrombolytic therapy. They obtained a prior distribution for θ expressing belief that "a 15–20% reduction in mortality is highly plausible, while the extremes of no benefit and a 40% reduction are both unlikely". This prior is shown in *Figure 1a*, while *Figure 1b* shows the likelihood expressing the support by the data (23/148 deaths on control versus 13/163 deaths on active treatment) for various values of θ. In contrast to the prior distribution, *Figure 1b* displays strong support for values of θ representing a 40–60% risk reduction.

Figure 1c shows the posterior distribution, obtained by multiplying the prior and likelihood together and then making the total area under the curve be equal to one (i.e. 'certainty'). The evidence in the likelihood has been pulled back towards the prior distribution – a formal representation of the belief that the results were 'too good to be true'.

Reporting probability statements

Having obtained a posterior distribution, to produce probabilities of exceeding certain thresholds, or lying in certain intervals, is only a computational task. In *Figure 1*, the posterior distribution provides an easily interpretable summary of the evidence, and probabilities for hypotheses of interest can then be read off the graph by calculating the relevant areas under the curve. For example, the most likely benefit is around a 24% risk reduction (half that observed in the trial), the posterior probability that the reduction is at least 50% is only 5%, and a 95% interval runs from a 43% to 0% risk reduction.

Such an interval is generally termed a 'credible interval' – unlike a confidence interval, it can be directly interpreted as saying that, given the prior assumptions, the model and the data, there is a 95% chance that the true reduction lies between 0 and 43%.

In many standard situations a traditional confidence interval is essentially equivalent to a credible interval based on the likelihood alone, and hence equivalent to using a 'flat' prior. Burton[82] claims that "it is already common practice in medical statistics to interpret a frequentist confidence interval as if it did represent a Bayesian posterior probability arising from a calculation invoking a prior density that is uniform on the fundamental scale of analysis".

The subjective interpretation of probability

The standard use of probability describes long-run frequency properties of repeated random events. This is known as the **frequency** interpretation of probability, and so both Fisherian and Neyman–Pearson schools are often referred to as 'frequentist'. We have allowed probability to refer to generic uncertainty about any unknown quantity, and this is an application of the **subjectivist** interpretation of probability.

This subjective view of probability is not new, and used to be standard. Fienberg[170] points out that Jakob Bernoulli in 1713 introduced "the subjective notion that the probability is personal and varies with an individual's knowledge", and that Laplace and Gauss both worked with posterior distributions, which became known as 'the inverse method'. However from the mid-nineteenth century the frequency approach started to dominate, and controversy has sporadically continued. Dempster[128] quotes Edgeworth in 1884 as saying the critics who "heaped ridicule upon Bayes's theorem and the inverse method" were trying to elicit "knowledge out of ignorance, something out of nothing". Polemical opinions are still expressed: in defence of the explicit introduction of subjective judgement into scientific research, Matthews[319] states that "it simply makes no sense to take seriously every apparent falsification of a plausible theory, any more than it makes sense to take seriously every new scientific idea".

The Bayesian perspective thus extends the remit of standard statistical analysis, in that there is explicit concern for what it is reasonable for an observer to believe in the light of data. Thus the perspective of the **consumer** of the analysis is explicitly taken into account; for example, in a trial on a new drug being carried out by a pharmaceutical company, the viewpoints of the company, the regulatory authorities and the medical profession may be substantially different. The **subjective** nature of the analysis is therefore unapologetically emphasised. Berger and Berry[38] state that "Bayesian statistics treats subjectivity with respect by placing it in the open and under the control of the consumer of data".

The prior distribution shown in *Figure 1a* was based on the subjective judgement of a senior cardiologist, informed by empirical evidence derived from one unpublished and two published trials. Of course, conclusions strongly based on beliefs that

TABLE 2 *The expected results when carrying out 200 clinical trials with α = 5% and β = 20% if only 10% of treatments are truly effective*

Trial conclusion	Treatment truly ineffective	Treatment truly effective	Total
Not significant	171	4	175
Significant	9	16	25
Total	180	20	200

cannot be supported by concrete evidence are unlikely to be widely regarded as convincing, and so it is important to attempt to find consensus on reasonable sources of external evidence. The assessment and use of prior beliefs is discussed further below and in chapter 3.

The relation to the use of Bayes's theorem in diagnostic tests

If θ is something that is potentially observable, and $p(\theta)$ can be derived from known data, then the use of Bayes's theorem is uncontroversial. For example, it has long been established that sensitivity and specificity are insufficient characteristics to judge the result of a diagnostic test for an individual – the disease prevalence is also needed.

From the health technology assessment perspective, the more interesting and controversial context is that of an unknown θ, which is a quantity that is not potentially directly observable, such as the mean benefit of a new therapy in a defined group of patients. There have been many arguments[80,132,361,429] for the connection between the use of Bayes's theorem in diagnostic testing and in general clinical research, pointing out that just as the prevalence is required for the assessment of a diagnostic test, so the prior distribution on θ is required to supplement the usual information (*P* values and confidence intervals) which summarises the likelihood. We need only think of the huge number of clinical trials that are carried out, with few clear successes, to realise that the 'prevalence' of truly effective treatments is low. We should thus be cautious about accepting extreme results, such as observed in the GREAT trial, at face value: indeed, Grieve[213] suggests a Bayesian approach provides "a yardstick against which a surprising finding may be measured".

Brophy and Joseph[76] defend a Bayesian approach to clinical studies by analogy to the differing levels of certainty which might be demanded by a diagnostic test under different circumstances, and Simon[405] provides the following example shown in *Table 2*. Suppose 200 trials are performed, but only 10% are of truly effective treatments. Suppose each trial is carried out with a type I error of 5% (the chance of claiming an ineffective treatment is effective) and a type II error of 20% (the chance of claiming an effective treatment is ineffective). Then *Table 2* shows that 9/25 = 36% of trials with significant results are in fact of totally ineffective treatments: in diagnostic testing terms, the 'predictive value positive' is only 64%.

Simon refers to this as the 'epidemiology of clinical trials', and suggests this should imbue a spirit of scepticism about unexpected significant trial results, which is naturally handled within a Bayesian perspective.

The prior distribution

Chapter 3 provides a full discussion of the source and use of prior distributions, including elicitation from experts, the use of 'default' priors to represent archetypal positions of ignorance, scepticism and enthusiasm and, when multiple related studies are being simultaneously analysed, the assumption of a common prior that may be 'estimated'.

Four important points should be emphasised immediately:

1. Despite the name 'prior' suggesting a temporal relationship, it is quite feasible that a prior distribution is decided on **after** seeing the results of a study, since it is simply intended to summarise reasonable uncertainty given evidence external to the study in question. Cox[117] states that "I was surprised to read that priors must be chosen before the data have been seen. Nothing in the formalism demands this. Prior does not refer to time, but to a situation, hypothetical when we have data, where we assess what our evidence would have been if we had had no data. This assessment may rationally be affected by having seen the data, although there are considerable dangers in this, rather similar to those in frequentist theory". Naturally when making predictions or decisions one's prior distribution needs to be unambiguously

TABLE 3 *Classical* P *values (adjusted for sequential examination of the data) and posterior probabilities of treatment superiority (no adjustment necessary because of repeated looks at the data)*

Patient pair	Preferred	$n_A - n_B$	Two-sided *P*	*P*(B > A)
1	A	1	1.0	0.25
2	B	0	1.0	0.50
3	A	1	1.0	0.31
4	A	2	0.63	0.19
5	A	3	0.38	0.11
6	B	2	0.69	0.23
7	A	3	0.45	0.14
8	A	4	0.29	0.090
9	A	5	0.18	0.055
10	A	6	0.11	0.033
11	A	7	0.065	0.019
12	A	8	0.039	0.011
13	A	9	0.022	0.0065
14	B	8	0.057	0.018
15	A	9	0.035	0.011
16	A	10	0.021	0.0064
17	A	11	0.013	0.0038
18	A	12	0.0075	0.0022
19	A	13	0.0044	0.0013
20	A	14	0.0026	0.0008
21	A	15	0.0015	0.0005

specified, although even then it is reasonable to carry out sensitivity analysis to alternative choices.

2. There is no such thing as the 'correct' prior. Instead, researchers have suggested using a 'community' of prior distributions expressing a range of reasonable opinions. Thus, a Bayesian analysis of evidence is best seen as providing a mapping from specified prior beliefs to appropriate posterior beliefs.

3. When **multiple** related studies are being simultaneously analysed, it may be possible to 'estimate' the prior for each study – see page 12.

4. As the amount of data increases, the prior will, unless it is of a pathological nature, be overwhelmed by the likelihood and will exert negligible influence on the conclusions.

Predictions

Suppose we wish to predict some future observations z, which depend on the unknown quantity θ through a distribution $p(z|\theta)$: for example, z may be the outcomes to be observed on some future patients. Since our current uncertainty concerning θ is expressed by the posterior distribution $p(\theta|y)$, then we can average over the current beliefs regarding the unknown θ to obtain the predictive distribution $p(z|y)$ for the future observations z.

Such predictive distributions are useful in many contexts: Berry and Stangl[45] describe their use in design and power calculations, model checking, and in deciding whether to conduct a future trial, while Grieve[208] provides examples in bioequivalence, trial monitoring and toxicology. Applications of predictions include power calculations (see page 28), sequential analysis (see page 31), payback from research (see page 52) and health policy making (see page 54).

Sequential analysis

Sequential data fall naturally within the Bayesian framework, as the posterior distribution following each observation becomes the prior for the next. This is discussed with regard to clinical trials in chapter 4, but for illustration we consider the example of Berry,[53] concerning the 6-mercaptopurine (6-MP) trial for acute

leukaemia.[186] In this study patients were allocated in pairs to 6-MP (A) or placebo (B), and the treatment assigned to whichever stayed in remission longer was considered the preference for that pair. Of the first 18 pairs, 15 preferred A, and the trial was stopped with $P < 0.01$; three subsequent pairs also preferred A. The data are shown in *Table 3*, with the frequentist two-sided P values and the posterior probability of B being the preferred treatment, assuming an initial uniform prior on the probability of B being preferable to A.

We note that the posterior probabilities can be calculated without regard to any stated stopping procedure, that the posterior probability that $P(B > A)$ is simply the prior probability that the next pair will prefer B, and that the posterior probabilities suggest the trial could have ended considerably sooner.

The idea that the data influence the posterior only through the likelihood is known as the likelihood principle[37,49] (see page 13), and underlies the strong Bayesian criticism of sequential analysis.

Decision-making

Suppose we wish to make one of a set of decisions, and that we are willing to assess some value $u(d, \theta)$, known as a utility, of taking the decision d when θ is the true unknown 'state of nature'. The theory of optimal decision-making says we should choose the decision that maximises our expected utility, where the expectation is taken with respect to our current probability distribution for θ.[307]

The use of Bayesian ideas in decision-making is a huge area of research and application, in which attention is focused on the utility of consequences rather than the use of Bayesian methods to revise beliefs.[471] This activity blends naturally into cost-effectiveness analysis (see page 51), but nevertheless the subjective interpretation of probability is essential, since the expressions of uncertainty required for a decision analysis can rarely be based purely on empirical data. There is a long history of attempts to apply this theory to medicine, and in particular there is a large literature on decision analysis, whether applied to the individual patient or for policy decisions.[301] The journal *Medical Decision Making* contains an extensive collection of policy analyses based on maximising expected utility, some of which particularly stress the importance of Bayesian considerations.

Therefore there has been a long debate on the use of loss functions (just the negative of utility), in parallel to that concerning prior distributions. Berry[53] and others have long argued that the design, monitoring and analysis of a study must explicitly take into account the consequences of eventual decisions. It is important to note that there is also a frequentist theory of decision-making that uses loss functions, but does not average with respect to prior or posterior distributions: the decision-making strategy is generally 'minimax', where the loss is minimised whatever the true value of the parameter might be. This can be thought of as assuming the most pessimistic prior distribution: see, for example, Bather[30] and Palmer.[344] Thus all combinations of the use of prior distributions and/or loss functions are possible: this is further discussed in the commentaries to this chapter and chapter 4.

The explicit use of utility functions within the design and monitoring of clinical trials is controversial but has been explored in a number of contexts: for example, Berry and Stangl[45] discuss whether to stop a Phase II trial based on estimating the number of women in the trial and in the future who will respond; whether to continue a vaccine trial by estimating the number of children who will contract the disease; and the use of adaptive allocation in a Phase III trial such that at each point the treatment which maximises the expected number of responders is chosen. See pages 32 and 37 for further discussion. A decision-theoretic approach also leads to formal techniques for assessing the payback from research (see page 52) and policy making (see page 54).

Hypothesis testing

Just as the Neyman–Pearson approach focuses on two competing hypotheses, it is possible to take a Bayesian hypothesis-comparison approach. This requires a prior distribution that places a 'lump' of probability on each competing hypotheses, which is updated to produce a posterior probability in, say, the null hypothesis that two drugs have equal effect (note that this is **not** the same as a P value, although such a mistaken interpretation of a P value is often made). Priors that explicitly consider the 'truth' of a null hypothesis are discussed on page 19. These were particularly promoted by Cornfield, who emphasised the 'relative betting odds' between two hypotheses, defined as the ratio of the posterior to the prior odds. This measure of the relative likelihood of two hypotheses is also known as the 'Bayes factor', and has been the subject of much research.[266]

Design

Bayesian design of experiments can be considered as a natural combination of prediction and decision-making, in that the investigator is seeking to choose a design which will achieve the desired goals. Chaloner and Verdinelli[99] provide a general review, while Simes[401] emphasises that trials that are designed to help treatment decisions require specification of utilities and detailed subgroup information in order to individualise treatment decisions. A simple design problem, choosing the size of a clinical trial, is considered on page 28.

Sequential designs present a particular problem known as 'backwards induction',[124] in which one must work backwards from the end of the study, examine all the possible decision points that one might face, and optimise the decision, allowing for all the possible circumstances in which one might find oneself. This is computationally very demanding since one must consider what to do in all possible future eventualities, but limited examples will be described on pages 32 and 37. Early phases of clinical trials have attracted this approach: Brunier and Whitehead[81] consider the balancing of costs of experimentation and errors in treatment allocation (see page 38).

We emphasise, however, that the likelihood principle states that Bayesian inference is carried out independently of design, so that the **reasons** for obtaining particular data points is irrelevant: the only aspect of interest is the relative likelihood of observing those particular data given possible values of the parameters of interest.

Multiplicity

The context of clinical trials may present issues of "multiple analyses of accumulating data, analyses of multiple end-points, multiple subsets of patients, multiple treatment group contrasts and interpreting the results of multiple clinical trials".[404] Observational data may feature multiple institutions, and meta-analysis involves synthesis of multiple studies. The general Bayesian approach to multiplicity involves specifying a common prior distribution for the substudies that expresses a belief in the expected 'similarity' of all the individual unknown quantities being estimated. This produces a degree of pooling, in which an individual study's results tend to be 'shrunk' towards the average result by an amount depending on the variability between studies and the precision of the individual study: relevant contexts include subset analysis, between-centre variability, and random-effects meta-analysis. This is essentially a random-effect approach, often labelled as 'empirical Bayes' or 'multilevel' modelling; we shall later distinguish these from a 'fully Bayes' approach.

This is later discussed with respect to hierarchical models (see page 21), subset analysis (see page 35), between-centre variability (see page 36), and institutional comparisons (see page 44 and chapter 12).

Complex modelling

Health technology assessments will generally involve some synthesis of evidence from a variety of sources, and a single clinical trial will rarely provide a definitive conclusion as to a policy recommendation. Standard statistical methods are designed for summarising the evidence from single studies, and although there has been a huge growth in methods and application of meta-analysis, these have generally been concerned with pooling evidence from studies with very comparable designs, outcome measures and so on.

A Bayesian approach allows evidence from diverse sources to be pooled through assuming that their underlying probability models (their likelihoods) have parameters of interest in common. For example, the 'true' effect of an intervention will feature in models for both randomised trials and observational data, even though there may be additional adjustments for potential biases, different populations, cross-overs between treatments and so on. One context in which this has been emphasised is in the regulation of medical devices (see page 53). This ability to deal with very complex models is discussed in chapter 6, but in general this great flexibility brings with it mathematical and computational challenges.

Computational issues

The Bayesian approach applies probability theory to a model derived from substantive knowledge and can, in theory, deal with realistically complex situations – the approach can also be termed 'full probability modelling'. It has to be acknowledged, however, that the computations may be difficult, with the specific problem being to carry out the integrations necessary to obtain the posterior distributions of quantities of interest in situations where non-standard prior distributions are used, or where there are additional 'nuisance parameters' in the model. These problems in integration

for many years restricted Bayesian applications to rather simple examples. However, there has recently been enormous progress in methods for Bayesian computation, generally exploiting modern computer power to carry out simulations known as Markov chain Monte Carlo (MCMC) methods.[189,193]

Although we will not be particularly concerned with computational issues, we refer to Carlin *et al.*[90] and Etzioni and Kadane[162] who discuss a range of methods which may be used (normal approximations, Laplace approximations and numerical methods including MCMC), Gelman and Rubin's[189] review of MCMC methods in biostatistics, and Vanhouwelingen's[464] commentary on the importance of computational methods in the future of biostatistics.

The theoretical justification for the Bayesian approach

The theoretical foundations for the optimality of Bayesian inference have been discussed at length by Cornfield,[112] Degroot,[124] Lindley,[306] Bernardo and Smith[43] and others. The crucial idea is that if one accepts certain intuitively plausible behavioural axioms, such as not placing bets that are guaranteed to lose money, then one should act according to decision theory based on subjective probabilities.

Apart from this somewhat ideological argument, the most important theoretical construct (aside from probability theory itself) is the likelihood principle,[37] which states that all the information that the data provides about the parameter is contained in the likelihood. This entails, for example, that frequentist stopping rules that retain type I error (with their consequence that the inferences are influenced by what one would have done had one observed something different) are entirely irrelevant.

Schools of Bayesians

It is important to emphasise that there is no such thing as a single Bayesian approach, and that many ideological differences exist between the researchers whose work is discussed in this chapter. Four levels of Bayesian approach, of increasing 'purity', may be identified:

1. The **empirical** Bayes approach, in which a prior distribution is estimated from multiple experiments. Analyses and reporting are in frequentist terms.

2. The **reference** Bayes approach, in which a Bayesian interpretation is given to conclusions expressed as posterior distributions, but an attempt is made to use 'objective' or 'reference' prior distributions.

3. The **proper** Bayes approach, in which informative prior distributions are based on available evidence, but conclusions are summarised by posterior distributions without explicit incorporation of utility functions.

4. The **decision-theoretic** or 'full' Bayes approach, in which explicit utility functions are used to make decisions based on maximising expected utility.

Our focus in this review is primarily on the third, **proper**, school of Bayesianism.

Making the health technology assessment context explicit

Bayesian methods explicitly allow for the possibility that the conclusions of an analysis may depend on who is conducting it and their available evidence and opinion, and therefore the context of the study is vital. Apart from methodological researchers, at least four different viewpoints can be identified for any health technology assessment:

- **sponsors,** for example the pharmaceutical industry, medical charities or granting agencies
- **investigators**, that is, those responsible for the conduct of a study, whether industry or publicly funded
- **reviewers,** for example regulatory bodies or journal editors
- **consumers**, for example agencies setting health policy, clinicians or patients.

An analysis which might be carried out solely for the investigators, for example, may not be appropriate for presentation to reviewers or consumers, and Racine *et al.*[370] point out that "experimentalists tend to draw a sharp distinction between providing their opinions and assessments for the purposes of experimental design and in-house discussion, and having them incorporated into any form of externally disseminated report".

TABLE 4 *A taxonomy of six possible 'philosophical' approaches to health technology assessment, depending on their objective and their quantitative use of prior information*

	Objective		
	Inference (estimation)	**Hypothesis testing**	**Decision (loss function)**
No prior	Fisherian	Neyman–Pearson	Classical decision theory
Prior	Proper Bayesian	'Bayes's factors'	Full decision-theoretic Bayesian

A characteristic of health technology assessment is that the investigators who plan and conduct a study are generally not the same body as those that make decisions on the basis of the evidence provided in part by that study: such decision makers may be regulatory authorities or healthcare providers. This division is acknowledged both in the structure of this report (in which chapter 4 considers design and monitoring of trials, and chapter 7 examines decision-making) and in the lower profile given to decision-making compared with inferences.

Further reading

A wide-ranging introductory chapter by Berry and Stangl[45] in their textbook[61] covers a whole range of modelling issues, including elicitation, model choice, computation, prediction and decision-making. Non-technical tutorial articles include those by Lewis and Wears,[295] Bland and Altman[68] and Lilford and Braunholtz.[299] Other authors emphasise different merits of Bayesian approaches in health technology assessment: Eddy *et al.*[148] concentrate on the ability to deal with varieties of outcomes, designs and sources of bias, Breslow[72] stresses the flexibility with which multiple similar studies can be handled, Etzioni and Kadane[162] discuss general applications in the health sciences with an emphasis on decision-making, while Freedman[177] and Lilford and Braunholtz[299] concentrate on the ability to combine 'objective' evidence with clinical judgement. Stangl and Berry[426] provide a recent review of biomedical applications.

There is a huge methodological statistical literature on general Bayesian methods, much of it quite mathematical. Dempster[128] gives a historical background, while Cornfield[112] provides a theoretical justification of the Bayesian approaches, in terms of ideas such as **coherence**. A rather old article[156] is still one of the best technical introductions to the Bayesian philosophy.

Good tutorial introductions are provided by Lindley[307] and Barnett,[29] while more recent books, in order of increasing technical difficulty, include those by Berry,[56] Lee,[287] O'Hagan,[338] Gelman *et al.*,[188] Carlin and Louis,[92] Berger[36] and Bernardo and Smith.[43]

There is a large literature on using Bayesian methods for specific modelling issues, and only a few references are provided here. Specific problems that have attracted a Bayesian analysis in health technology assessment include survival analysis,[2,7,218,427] longitudinal models,[257,282] missing data and dropouts,[116,270,285,354,385] and model criticism and selection.[205]

Commentary

As mentioned on page 11, we need to carefully distinguish the debate as to whether prior distributions should be used for inferences, from the question of whether our objective should be estimation, hypothesis testing or a decision requiring a loss function of some kind. *Table 4* considers all possible combinations of these elements.

All six approaches have been investigated in theory and, to some extent, in practice, and the resulting arguments become complex. Here we shall only give a very brief overview; see the commentaries in each chapter for further detailed objections.

Prior or no prior?

There have been generic warnings about a Bayesian approach, and Feinstein[167] claims that "a statistical consultant who proposes a Bayesian analysis should therefore be expected to obtain a suitably informed consent from the clinical client whose data are to be subjected to the experiment". However, some seem to misunderstand the Bayesian explicit acceptance of subjectivity: Lane[280] states that "We are told that elimination of subjectivity by use of Bayesian inference paves the way to truly objective, evidence-based practice. Yet who but a statistically minded minority can begin to interpret Bayesian analysis?".

In general, the objections to the use of a prior distribution have been pragmatic, and strict ideological standpoints have been played down: Cox and Farewell[118] say it would be "a great pity if differences of technical approach were exaggerated into differences about qualitative issues", while Armitage[19] maintains it is not appropriate to polarise the argument as a choice between two extremes. Practical difficulties in obtaining and using prior opinions will be described in the commentary to succeeding chapters: Fisher[171] and O'Rourke[343] provide lists of general objections, while Tukey[454] suggests that the Bayesian approach to multiplicity has little to contribute due to it being rarely appropriate to assume exchangeability.

Following the discussion on page 13, many authors have argued that the appropriateness of the Bayesian approach depends crucially on the circumstances and the precise question being asked: for example, both Koch[272] and Whitehead[476] claim that a proper Bayesian approach may be reasonable at early stages of a drug's development but is not acceptable in Phase III trials.

Estimation, hypothesis testing or decision-making?

Whether or not to incorporate an explicit loss function would appear to depend on the question being asked, and the extent to which a health technology assessment should lead to an inference about a treatment effect or a decision as to future policy has been strongly debated.[15,72,223,249,402] Important objections that have been raised to a decision-theoretic approach include the lack of a coherent theory for decision-making on behalf of multiple audiences with different utility functions, the difficulty of obtaining agreed utility values, and the fact that future treatments would be recommended on the basis of even marginal expected gains, without any concern with the confidence with which such a recommendation is made (see also page 39). We also note Kass and Greenhouse's[267] argument against the Bayes factor approach, claiming that "in most RCTs, estimation would be more appropriate than testing".

Key points

1. Claims of advantages and disadvantages of Bayesian methods are now largely based on pragmatic reasons rather than blanket ideological positions.

2. A Bayesian approach can lead to flexible modelling of evidence from diverse sources.

3. Bayesian methods are best seen as a transformation from initial to final opinion, rather than providing a single 'correct' inference.

4. Different contexts may demand different statistical approaches, both regarding the role of prior opinion and the role of an explicit loss function. It is vital to establish contexts in which Bayesian approaches are appropriate.

5. A decision-theoretic approach may be appropriate where the consequences of a study are predictable, such as when dealing with rare diseases treated according to a protocol, within a pharmaceutical company, or in public health policy.

Chapter 3

Where does the prior distribution come from?

Introduction

There is no denying that quantifiable prior beliefs exist in medicine. For example, in the context of clinical trials, Peto and Baigent[358] state that "it is generally unrealistic to hope for large treatment effects" but that "it might be reasonable to hope that a new treatment for acute stroke or AMI could reduce recurrent stroke or death rates in hospital from 10% to 9% or 8%, but not to hope that it could halve in-hospital mortality". However, turning informally expressed opinions into a mathematical prior distribution is perhaps the most difficult aspect of Bayesian analysis. Four broad approaches are outlined below, ranging from the elicitation of subjective opinion to the use of statistical models to estimate the prior from data. Sensitivity analysis to alternative assumptions is considered vital, whatever the method used to construct the prior distribution.

From a mathematical and computational perspective, it is extremely convenient if the prior distribution is a member of a family of distributions that is **conjugate** to the form of the likelihood, in the sense that they 'fit together' to produce a posterior distribution that is in the same family as the prior distribution. For example, many likelihoods for treatment effects can be assumed to have an approximately normal shape,[421] and thus in these circumstances it will be convenient to use a normally shaped prior (the conjugate family), provided it approximately summarises the appropriate external evidence. Similarly when the observed data are to be proportions (implying a binomial likelihood), the conjugate prior is a member of the 'beta' family of distributions which provide a flexible way of expressing beliefs about the magnitude of a true unknown frequency. Modern computing power is, however, reducing the need for conjugacy, and in this section we shall concentrate on the source of the prior rather than its precise mathematical form.

We should repeat the statements made on page 9 regarding the fact that there is no 'correct' prior: Bayesian analysis can be seen as a means of transforming prior into posterior opinions, rather than producing **the** posterior distribution. It is therefore vital to take into account the context and audience for the assessment (see page 13).

Elicitation of opinion

A true subjectivist Bayesian approach requires only a prior distribution that expresses the personal opinions of an individual but, if the health technology assessment is to be generally accepted by a wider community, it would appear to be essential that the prior distributions have some evidential or at least consensus support. In some circumstances there may, however, be little 'objective' evidence available and summaries of expert opinion may be indispensable.

There is a very large literature on eliciting subjective probability distributions from experts, with some good early references on statistical[391] and psychological aspects,[456] as well as on methods for pooling distributions obtained from multiple experts.[190] The fact that people are generally not good probability assessors is well known, and the variety of biases they suffer are discussed by Kadane and Wolfson.[265] Nevertheless it has been shown that training can improve experts' ability to provide judgements that are 'well calibrated', in the sense that if a series of events are given a probability, say 0.6, then around 60% of these events will occur. Murphy and Winkler[331] discuss this issue with regard to weather forecasting.

Chaloner[97] provides a thorough review of methods for prior elicitation in clinical trials, including interviews with clinicians, postal questionnaires, and the use of an interactive computer program to draw a prior distribution. She concludes that fairly simple methods are adequate, using interactive feedback with a scripted interview, providing experts with a systematic literature review, basing elicitation on 2.5 and 97.5% percentiles, and using as many experts as possible.

Berry and Stangl[45] discuss methods for eliciting conjugate priors and checking whether the assessments are consistent with each other. Kadane and Wolfson[260] distinguish between **experiments to learn** for which only the prior of the experimenter needs to be considered, and **experiments to prove**, in which the priors (and utilities) of an audience to be persuaded need to be considered. They discuss general methods of elicitation, and give an example applied to an observational study in

healthcare. They emphasise the method by which experts may be asked for predictions about future observations; from these assessments an implicit prior distribution can be derived.

The methods used in case studies can be divided into four main categories using increasingly formal methods:

1. **Informal discussion.** Examples include Rosner and Berry,[381] who consider a trial of paclitaxel (Taxol®) in metastatic breast cancer, in which a beta prior for the overall success rate was assessed using a mean of 25% and 50% belief that the true success rate lay between 15 and 35%. Similarly, Lilford and Braunholtz[299] describe how priors were obtained from two doctors for the relative risk of venous thrombosis from oral contraceptives. In the context of meta-analysis, Smith *et al.*[412] thought that it was unlikely that the between-study odds would vary by more than an order of magnitude, and hence derived a prior distribution on the heterogeneity parameter.

2. **Structured interviewing and formal pooling of opinion**. Freedman and Spiegelhalter[180] describe an interviewing technique in which a set of experts were individually interviewed and hand-drawn plots of their prior distributions elicited. Deliberate efforts were made to prevent the opinions being over-confident (too 'tight'). The distributions were converted to histograms, and averaged to produce a composite prior. This was carried out twice for trials of thiotepa in superficial bladder cancer and these priors used later in discussing initial and interim power of the study,[416,418] (see pages 28 and 29). A similar exercise was carried out before a trial in osteosarcoma[414] and used in a discussion of the power of the trial.[420] Gore[198] introduced the concept of 'trial roulette', in which 20 gaming chips, each representing 5% belief, could be distributed amongst the bins of a histogram: in a trial of artificial surfactant in premature babies, 12 collaborators were interviewed using this technique to obtain their opinion on the possible benefits of the treatment.[442] Using an electronic tool so that individuals in a group could respond without attribution, Lilford *et al.*[297] presented collaborators in a trial with a series of imaginary patients in order to elicit their opinions on the benefit of early delivery.

3. **Structured questionnaires**. The 'trial roulette' scheme described above was administered by post by Hughes[242,243] for a trial in treatment of oesophageal varices, and by Abrams *et al.*[1,421] for a trial of neutron therapy.

   Parmar *et al.*[348] elicited prior distributions for the effect of a new radiotherapy regime (CHART), in which the possible treatment effect was discretised into 5% bands and the form was sent by post to each of nine clinicians. Each provided a distribution over these bands and an arithmetic mean was then taken. Parmar *et al.* provide a copy of the form in their paper, and results are provided for both lung[348] and head and neck cancer.[421]

4. **Computer-based elicitation**. Chaloner *et al.*[98] provide a detailed case study of the use of a rather complex computer program that interactively elicited distributions from five clinicians for a trial of prophylactic therapy in AIDS (see also Carlin *et al.*[90]). Kadane[258] reports the results of an hour-long telephone interview with each of five clinicians, using software to estimate prior parameters from the results of a series of questions eliciting predictive probability distributions for responses of various patient types. When a second round of elicitation became necessary, the proposal was met by "little enthusiasm". Kadane and Wolfson[260] provide an edited transcript of a computerised elicitation session in a non-trial context. Dumouchel[141] describes a computer program for eliciting prior distributions in hierarchical models, in which beliefs about contrasts give rise to informative prior distributions on between-unit variability.

Summary of evidence

If the results of previous similar studies are available it is clear that may be used as the basis for a prior distribution. Three main approaches have been used:

1. **Use of a single previous experimental result as a prior distribution**. Brown *et al.*[78] and Spiegelhalter *et al.*[421] both provide examples of prior distributions proportional to the likelihoods arising from a single pilot trial.

   Korn *et al.*[276] could not identify a beta prior matching the characteristics of their past empirical experience, but Bring[75] pointed out that had they fitted the mode rather than the mean of the prior to their past data then the problem would have disappeared.

2. **Direct use of a meta-analysis of many previous studies**. Lau *et al.*[284] point out that cumulative meta-analysis can be given a Bayesian interpretation in which the prior for each trial is obtained from the meta-analysis of preceding studies, while Dersimonian[129] derives priors for a trial of the effectiveness of calcium supplementation in the prevention of pre-eclampsia in pregnant women by a meta-analysis of previous trials using both random effects and fixed effects models. Smith *et al.*[410] and Higgins and Whitehead[233] both use a review of past meta-analyses as a basis for a prior distribution for the degree of heterogeneity in a current meta-analysis. Gilbert *et al.*[192] report an empirical distribution of past trial results in surgery which could be used as a prior.

3. **Use of previous data in some discounted form**. Previous studies may not be directly related to the one in question and we may wish to discount its influence. Kass and Greenhouse[267] state that "we wish to use this information, but we do not wish to use it as if the historical controls were simply a previous sample from the same population as the experimental controls", and they investigate a range of such discounting methods (see appendix 1). Berry and Stangl[45] discuss the use of historical information with trial data, giving an example of a Bayesian logistic regression analysis with historical data weighted by various amounts: see page 5.4 for discussion of historical controls in randomised trials. Brophy and Joseph[76] pool two previous studies to form a prior but investigate the effect of downweighting their influence to 50 and 10%, Cronin *et al.*[121] report a mixture of past evidence and 'prior belief of what is a reasonable range of values', while Greenhouse and Wasserman[206] downweight a previous trial with 176 subjects to have weight equivalent to only 10 subjects.

Default priors

It would clearly be attractive to have prior distributions that could be taken 'off-the-shelf', rather than having to consider all available evidence external to the study in their construction. Four main suggestions can be identified.

'Non-informative' or 'reference' priors

There has been substantial research in the Bayesian literature into so-called **non-informative** or **reference** priors, which usually take the form of uniform distributions over the range of interest, possibly on a suitably transformed scale of the parameter.[70] Formally, a uniform distribution means the posterior distribution has the same shape as the likelihood function, which in turn means that the resulting Bayesian intervals and estimates will essentially match the traditional results, so that posterior tail areas will match one-sided *P* values (ignoring any adjustment for multiplicity). Hughes[243] points out that a careful choice of scale is necessary and that a uniform prior on, say, a log(relative risk) might entail an inappropriate prior on a risk difference, while Kass and Wasserman[268] review the current status of such reference priors.

Results with reference priors are generally quoted as one part of a Bayesian analysis. In particular, Burton[82] suggests that most doctors interpret frequentist confidence intervals as posterior distributions, and also that information prior to a study tends to be vague, and that therefore results from a study should be presented by performing a Bayesian analysis with a non-informative prior and quoting posterior probabilities for the parameter of interest being in various regions. He gives examples from a hypothetical logistic regression setting and from the evaluation of the effectiveness of a *Haemophilus influenzae* type b (HIB) vaccine in Aboriginal children.

Other applications of reference priors include Achcar *et al.*,[7] Briggs[73] on net benefit in cost-effectiveness studies, and almost all the confidence profile examples in Eddy *et al.*[150] However, Fisher[171] points out that "there is no such thing as a "noninformative" prior. Even improper priors give information: all possible values are equally likely".

'Sceptical' priors

Informative priors that express scepticism about large treatment effects have been put forward both as a reasonable expression of doubt, and as a way of controlling early stopping of trials on the basis of fortuitously positive results. Kass and Greenhouse[267] suggest that a "cautious reasonable sceptic will recommend action only on the basis of fairly firm knowledge", but that these sceptical "beliefs we specify need not be our own, nor need they be the beliefs of any actual person we happen to know, nor derived in some way from any group of "experts" ".

Mathematically speaking, a sceptical prior about a treatment effect will have a mean of zero and some spread which determines the degree of scepticism. Fletcher *et al.*[173] use such a prior, while Spiegelhalter *et al.*[421] argue that a reasonable

degree of scepticism may be equivalent to feeling that the alternative hypothesis is optimistic, formalised by a prior with only a small probability (say 5%) that the treatment effect is as large as the alternative hypothesis (see *Figure 2*).

This approach has been used in a number of case studies,[183,348] and has been suggested as a basis for monitoring trials[166] (see page 29) and when considering whether or not a confirmatory study is justified (see page 33). Other users of sceptical priors include Dersimonian[129] and Heitjan[227] in the context of Phase II studies, while a senior US Food and Drugs Administration (FDA) biostatistician[339] has stated that he "would like to see [sceptical priors] applied in more routine fashion to provide insight into our decision making".

'Enthusiastic' priors

As a counterbalance to the pessimism expressed by the sceptical prior, Spiegelhalter *et al.*[421] suggest an 'enthusiastic' prior centred on the alternative hypothesis and with a low chance (say 5%) that the true treatment benefit is negative. Use of such a prior has been reported in case studies[183,227] and as a basis for conservatism in the face of early negative results[166] (see page 31). Digman *et al.*[133] provide an example of such a prior, but call it 'optimistic'.

Priors with a point mass at the null hypothesis ('lump-and-smear' priors)

The traditional statistical approach expresses a qualitative distinction between the role of a null hypothesis, generally of no treatment effect, and alternative hypotheses. A prior distribution that retains this distinction would place a 'lump' of probability on the null hypothesis, and 'smear' the remaining probability over the whole range of alternatives: Cornfield[112] uses a normal distribution centred on the null hypothesis, while Hughes[243] uses a uniform prior over a suitably restricted range. The resulting posterior distribution retains this structure, giving rise to a posterior probability of the truth of the null hypothesis: this is apparently analogous to a *P* value but is neither numerically nor conceptually equivalent (see page 11).

Cornfield repeatedly argued for this approach, which naturally gives rise to the 'relative betting odds' as a sequential monitoring tool, defined as the ratio of the likelihood of the data under the null hypothesis to the average likelihood (with respect to the prior) under the alternative.[113] The relative betting odds is independent of the 'lump' of prior probability placed on the null (although it does depend on the shape of the 'smear' over the alternatives), and does not suffer from the problem of 'sampling to a foregone conclusion' (see page 32). He suggests a 'default' prior under the alternative as a normal distribution centred on the null hypothesis and with expectation (conditional on the effect being positive), equal to the alternative hypothesis δ, that is, with prior standard deviation $\sqrt{(\pi/2\delta)}$. This is essentially equivalent to the formulation of a sceptical prior described above, but with probability of exceeding the alternative hypothesis of $\gamma = 0.21$ – this is larger than the value of 5% often used for sceptical priors, but the lump of probability on the null hypothesis is already expressing considerable scepticism. Values for these prior distributions for 11 outcome measures are reported for the Urokinase Pulmonary Embolism Trial.[389] This method was used in a number of major studies alongside more standard approaches,[114,389,458] although relative betting odds were dropped from the final report of the Coronary Drug Project.[115]

A mass of probability on the null hypothesis has also been used in a cancer trial[182] and for sensitivity analysis in trial reporting[243] (see page 33).

Although such an analysis provides an explicit probability that the null hypothesis is true, and so appears to answer the question of interest, the prior might be somewhat more realistic were the lump to be placed on a small range of values representing the more plausible null hypothesis of 'no clinically effective difference' (although Cornfield[112] points out that the 'lump' is just a mathematical approximation to such a prior). Lachin[277] has extended the approach to this situation where the null hypothesis forms an interval.

'Robust' priors

To save all the effort of eliciting a prior distribution, it seems reasonable when reporting or monitoring a study to identify how the prior affects the conclusions. This essentially means identifying a class of prior distributions, and then seeing how the posterior conclusions vary with priors within that class. Greenhouse and Wasserman[206] carry out two such case studies, while Carlin and Sargent[93,388] use 'prior partitioning' to identify the set of priors that would lead, say, to stopping a trial according to specified tail-area posterior probabilities. Those in authority then simply have to assess whether their prior lies in the appropriate partition. See pages 31 and 33 for further discussion of this approach in clinical trials.

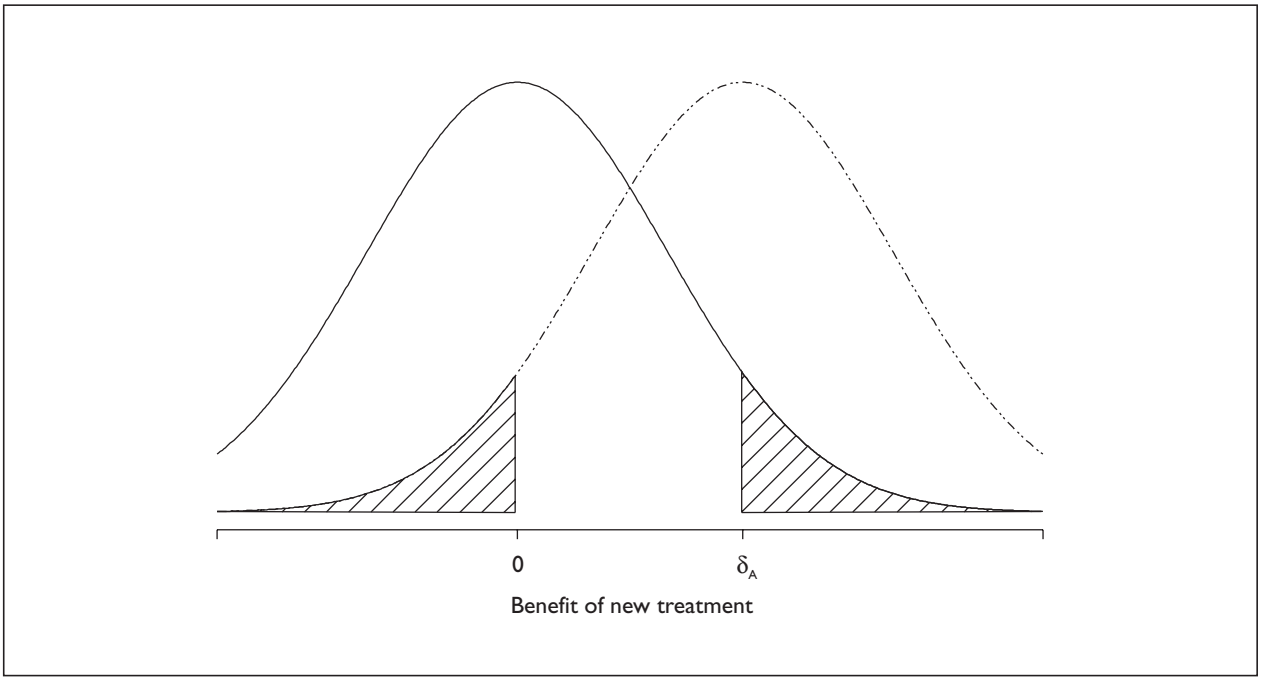


FIGURE 2 *Sceptical (solid line) and enthusiastic (broken line) priors for a trial with alternative hypothesis δ_A. The sceptics' probability that the true difference is greater than δ_A is γ (shown shaded). This value has also been chosen for the enthusiasts' probability that the true difference is less than 0*

Exchangeability, hierarchical models and multiplicity

Suppose we are interested in making inferences on many parameters $\theta_1,\ldots, \theta_k$ measured on k 'units' which may, for example, be true treatment effects in subsets of patients, multiple clinics, or for each of a series of trials. We can identify three different assumptions:

1. All the θs are identical, in which case all the data can be pooled and the individual units forgotten.

2. All the θs are entirely unrelated, in which case the results from each unit are analysed independently (for example using a fully specified prior distribution).

3. The θs are assumed to be 'similar' in the following sense. Suppose we were blinded as to which unit was which, and all we had was a label for each, say A, B, C and so on. Then our prior opinion about any particular set of θs would not be affected by only knowing the labels, in that we have no reason to think specific units are systematically different. Such a set of θs are called 'exchangeable',[43] and this assumption can be shown under broad conditions to be mathematically equivalent to assuming they are drawn at random from some population distribution, just as in a traditional random effects model. Note that there does not need to be any actual sampling – perhaps these k units are the only ones that exist – since the probability structure is a consequence of the belief rather than a physical randomisation mechanism.

We emphasise that an assumption of exchangeability is a judgement based on our knowledge of the context. If there are known reasons to suspect specific units are systematically different, then those reasons need to be modelled. Dixon and Simon[134] discuss the reasonableness of exchangeability assumptions.

The Bayesian approach to multiplicity is thus to integrate all the units into a single model, in which it is assumed that $\theta_1,\ldots, \theta_k$ are drawn from some common prior distribution whose parameters are unknown: this is known as a hierarchical model. These unknown parameters may then be estimated directly from the data using the technique of 'marginal maximum likelihood' (known as the 'empirical Bayes' approach), or given a (usually minimally informative) prior distribution (known as the 'full Bayes' approach). The results from either an empirical or full Bayes analysis will generally be similar, and lead to the inferences for each unit having narrower intervals than if they are assumed independent, but **biased** towards mean response. Thus there is a similar conservatism to

TABLE 5 *A comparison of some elicited subjective prior distributions and the consequent results of the clinical trials. In each case a pooled prior was provided, assumed normal on a log(hazard ratio) scale – Box's P value is calculated on this scale. This is transformed to a hazard ratio scale where a hazard ratio > 1 corresponds to benefit of the new treatment: median and 95% intervals are given*

Study	Prior			Likelihood			Z	P
	Hazard ratio	95% interval	Reference	Hazard ratio	95% interval	Reference		
CHART (lung)	1.37	0.87–2.14	348	1.32	1.09–1.59	390	0.14	0.89
Thiotepa XI	1.65	0.99–2.74	418	0.90	0.63–1.29	375	1.90	0.06
Osteo	1.11	0.66–1.83	420	0.94	0.69–1.27	413	0.54	0.59
Neutron (clinical prior)	1.19	0.67–2.08	421	0.66	0.40–1.10	421	1.51	0.14

the traditional adjustment, but for a radically different reason.

Cornfield[112,113] was an early proponent of the Bayesian approach to multiplicity (see page 35). Breslow[72] gives many examples of problems of multiplicity and reviews the use of empirical Bayes methods for longitudinal data, small area mapping, estimation of a large number of relative risks in a case–control study, and multiple tumour sites in a toxicology experiments, while Louis[309] reviews the area and provides a detailed case study. Other examples in clinical trials are described on page 35.

Empirical criticism of priors

The ability of subjective prior distributions to predict the true benefits of interventions is clearly of great interest, and Box[69] suggested a methodology for comparing priors with subsequent data. The prior is used to create a predictive distribution of the future summary statistic, which provides the prior predictive probability of observing data as extreme as that actually seen. This probability may be termed Box's *P* value. Spiegelhalter[416] showed how this could be applied to a clinical trial, and applications have been reported for discrete data,[369,422] while Spiegelhalter *et al.*[421] display predictive distributions and Box's *P* values in all their analyses (see chapter 9 for an application in the CHART trial.) For normal prior and likelihood, Box's *P* value is simply derived from the standardised difference (denoted *Z*) between the two distributions (difference in means divided by the square root of the sum of their variances).[421]

There have been a number of prospective elicitation exercises for clinical trials, and many of these trials have now reported their results. *Table 5* shows a selection of results, including the intervals for the prior distributions for treatment effects, the evidence from the likelihood, and Box's *P* value summarising the conflict between the prior and the likelihood.

Table 5 shows the generally poor experience obtained from prior elicitation. The clinicians are universally optimistic about the new treatments (median of prior hazard ratios > 1), whereas only one of the trials eventually showed any evidence of benefit from the new treatment (likelihood hazard ratio > 1). This also reflects the experience of Carlin *et al.*[90] in their elicitation exercise.

Commentary

Here we shall only consider comments on how priors may be best obtained: whether they should be used or not is discussed elsewhere (see pages 14 and 39).

There have been many criticisms of the process of eliciting subjective prior distributions in the health technology assessment context. Claims include:

1. **Subjects are biased in their opinions**. Gilbert *et al.*[192] state that "innovations brought to the stage of randomised trials are usually expected by the innovators to be sure winners", and Hughes[242] points out that the very fact that clinicians are participating in a trial is likely to suggest they expect the new therapy to be of benefit; this appears to be borne out in the results shown in *Table 5*. Altman[12] warns that investigators may even begin to exaggerate their prior beliefs in order to make their prospective trial appear more attractive (although this appears to already happen both in public and industry-funded studies). Fisher[171] believes the effort put into elicitation is misplaced, since the measured beliefs are likely to be based more on emotion than scientific evidence.

2. **The choice of subject biases results**. The biases discussed above mean that the choice of subject for elicitation is likely to influence the results. If we wish to know the distribution of opinions among well-informed clinicians, then trial investigators are not a random sample and will give biased conclusions. Lewis[292] says statisticians reviewing the literature should provide much better prior distributions than clinicians, while Chalmers[95] suggests even lay people are biased towards believing new therapies will be advances, and therefore we need empirical evidence on which to base the prior probability of superiority.

   Fisher[171] claims that uncertainty as to whose prior to use militates against any use of Bayesian methods. However, it can be argued that taking context into account (see page 13) means that it is quite reasonable to allow for differing perspectives.

3. **Timing of elicitation has an influence**. Senn[397] objects to any retrospective elicitation of priors as "present remembrance of priors past is not the same as a true prior", while Hughes[242] points out that opinions are likely to be biased by what evidence has recently been represented and by whom.

4. **Subjective judgement of exchangeability may be inappropriate**. Tukey[454] says that "to treat the true improvements for the classes concerned as a sample from a nicely behaved population … does not seem to me to be near enough the real world to be a satisfactory and trustworthy basis for the careful assessment of strength of evidence".

These concerns have led to a call for the evidential basis for priors to be made explicit, and for effort to go into identifying reasons for disagreement and attempting to resolve these.[171] Even advocates of Bayesian methods have suggested that the biases in clinical priors suggest more attention to empirical evidence from past trials, possibly represented as sceptical priors: Fayers[164] asks, given the long experience of negative trials, "should we not be using priors strongly centred around 0, irrespective of initial opinions, beliefs and hopes of clinicians?", while Spiegelhalter *et al.*[421] say that elicited priors from investigators show predictable positive bias and may possibly be replaced by archetypal 'enthusiastic' priors.

Key points

1. The use of a prior is based on judgement and hence a degree of subjectivity cannot be avoided.

2. The prior is important and not unique, and so a range of options should be examined in a sensitivity analysis.

3. The intended audience for the analysis needs to be explicitly specified.

4. The quality of subjective priors (as assessed by predictions) show predictable biases in terms of enthusiasm.

5. For a prior to be taken seriously, its evidential basis must be explicitly given, as well as any assumptions made (e.g. downweighting of past data). Care must, however, be taken of bias in published results.

6. Archetypal priors may be useful for identifying a reasonable range of prior opinion.

7. Great care is required in using default priors intended to be minimally informative.

8. Exchangeability assumptions should not be made casually.

Chapter 4

A guide to the Bayesian health technology assessment literature: conduct of randomised controlled trials

A Bayesian: one who asks you what you think before a clinical trial in order to tell you what you think afterwards.

Stephen Senn[398]

General arguments

Randomised controlled trials have provided fertile territory for arguments between alternative statistical philosophies. There are many specific issues in which a distinct Bayesian approach is identifiable, such as the ethics of randomisation, power calculations, monitoring, subset analysis, alternative designs and so on, and these are dealt with in separate sections below.

There are also a number of general discussion papers on the relevance of Bayesian methods to trials. These include tutorial introductions at a non-technical[295] and slightly more technical level,[1] while Brophy and Joseph[76] explain the Bayesian approach by demonstrating its use in re-analysing data from the Global Utilization of Streptokinose and t-PA for Occluded Coronary Arteries (GUSTO) trial of different thrombolytic regimes, using prior information from other studies. More detailed but non-mathematical discussions are given by Cornfield[113] and Kadane,[262] who particularly emphasises the internal consistency of the Bayesian approach, and welcomes the need for explicit prior distributions and loss function as producing scientific openness and honesty. Pocock and Hughes[365] again provide a non-mathematical discussion concentrating on estimation issues in trials, while Armitage[22] attempts a balanced view of the competing methodologies. Particular emphasis has been placed on the ability of Bayesian methods to take full advantage of the accumulating evidence provided by small trials.[302,317] A special issue of *Statistics in Medicine* on 'methodological and ethical issues in clinical trials' contains papers both for[53,420,459] and against[476] the Bayesian perspective, and features incisive discussion by Armitage, Cox and others. Palmer and Rosenberger[346] review non-standard trial designs and suggest circumstances where they may be appropriate.

Somewhat more technical reviews are given by Spiegelhalter and colleagues.[420,421] Berry[52,53,55] has long argued for a Bayesian decision-theoretic basis for clinical trial design, and has described in detail methods for elicitation, monitoring, decision-making and using historical controls.

Earlier (see page 14), we identified the fundamental dual issues of prior distributions and loss functions, and we have followed this division by focusing on **inferences** from a clinical trial in this chapter, and delaying discussion of subsequent decisions, whether based on randomised or non-randomised evidence, to chapter 7. We realise this is a somewhat artificial separation, as the potential role for explicit statement of a loss function is a running theme throughout discussions on design, sample size, sequential analysis, adaptive allocation and payback from research programmes, and many would argue that the eventual decision is inseparable from design and analysis of a study. However, the health technology assessment context often means that the investigators who design and carry out a study are generally not the same body who make decisions on the basis of the evidence (see page 13), and so if we take a pragmatic rather than ideological perspective, then the attempt at separation appears reasonable.

Ethics and randomisation

Is randomisation necessary?

Randomisation has two traditional justifications: it ensures treatment groups are directly comparable (up to the play of chance), and it provides a fundamental basis for the probability distributions underlying conventional statistical procedures.

Since Bayesian probability models are derived from subjective beliefs and do not require any underlying random mechanism, the latter requirement is irrelevant, and this has led some to

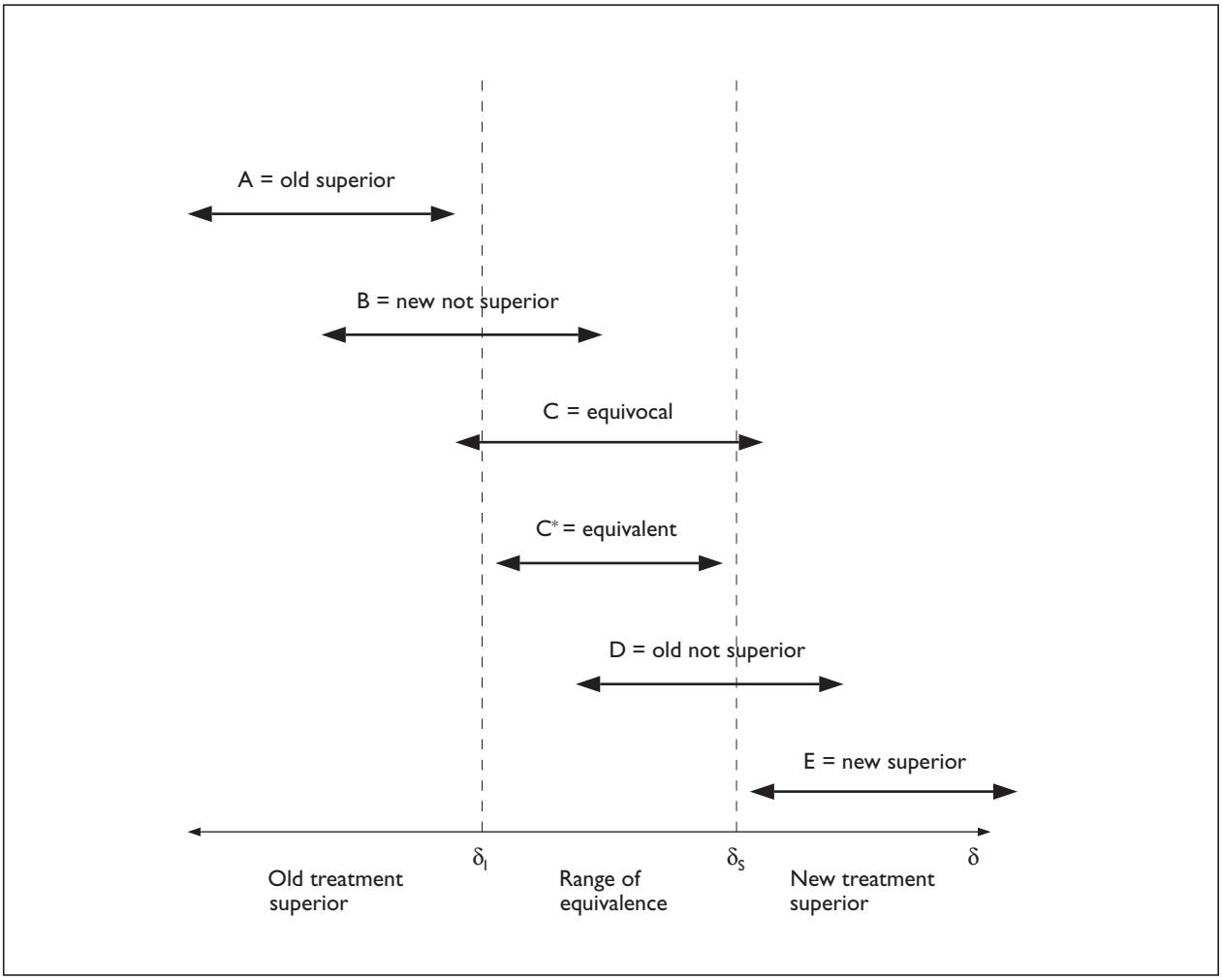


FIGURE 3 *Possible situations at any point in the progress of a trial, derived from superimposing an interval estimate (say 95%, denoted ↔) on the range of equivalence*

question the need for randomisation at all, provided alternative methods of balancing groups can be established. For example, Urbach[459] argues that a "Bayesian analysis of clinical trials affords a valid, intuitively plausible rationale for selective controls, and marks out a more limited role for randomisation than it is generally accorded", while Kadane[261] suggests updating clinician's priors and only assigning treatments that at least one clinician considers optimal. Berry goes further in claiming[46] "Randomised trials are inherently unethical". Papineau[347] refutes Urbach's position and claims that despite it not being essential for statistical inference, experimental randomisation forms a vital role in drawing causal conclusions (see also Rubin[383]). The relationship between randomisation and causal inferences is beyond the scope of this chapter, but in general the need for sound experimental design appears to dominate philosophical statistical issues.[246] In fact Berry and Kadane[65] suggest that if there are several parties who make different decisions and observe different data, randomisation may be a strictly optimal procedure since it enables each observer to draw his or her own appropriate conclusions. Kadane and Seidenfeld[263] make a useful distinction between experiments to learn and those to prove, which we will find useful when it comes to discussing confirmatory trials.

That careful analysis of databases can to some extent replace randomised trials has been argued by Howson and Urbach[240] and Hlatky.[235] Byar[84] puts an opposing view.

When is it ethical to randomise?

If we agree that randomisation is in theory useful, then the issue arises of when it is ethical to randomise. This is closely associated with the process of deciding when to stop a trial (discussed further on page 29) and is often represented as a balance between individual and collective ethics.[346,362] The ethics of randomisation and clinical trials have been covered in Edwards *et al.*[155]

Freedman[176] introduced the idea of professional equipoise, in which disagreement among the medical profession makes randomisation ethical. The trial design of Kadane[258] (see appendix 1) is an expression of this principle, in that only a treatment that at least one clinician thought optimal could be given to a patient (although it unfortunately turned out that a computer bug meant that some patients were allocated to treatments that **all** clinicians felt were suboptimal). Perhaps a more appealing approach is the 'uncertainty principle', which is often argued as a basis for ethical randomisation[85]: this may be thought of as 'personal equipoise'[154] in which the clinician is uncertain as to the best treatment for the patient in front of him or her. However, a quantified degree of uncertainty has not been specified.

The Bayesian approach can be seen as formalising the uncertainty principle by explicitly representing, in theory, the belief of an individual clinician that a treatment may be beneficial to a specific patient – this could be provided by superimposing the clinician's posterior distribution on the range of equivalence (see page 27) relevant to a particular patient.[421] It has been argued that a Bayesian model naturally formalises the individual ethical position,[300,345] in that it explicitly confronts the personal belief in the clinical superiority of one treatment. Berry,[53] however, has suggested that if patients were honestly presented with numerical values for their clinician's belief in the superiority of a treatment, then few might agree to be randomised. One option might be to randomise but with a varying probability that is dynamically weighted towards the currently favoured treatment: such adaptive allocation designs are discussed on page 37.

Kass and Greenhouse[267] argue that "the purpose of a trial is to collect data that bring to conclusive consensus at termination opinions that had been diverse and indecisive at the outset", and go on to state that "randomisation is ethically justifiable when a cautious reasonable sceptic would be unwilling to state a preference in favour of either the treatment or the control". This approach leads naturally to the development of sceptical prior distributions (see page 19) and their use in monitoring sequential trials (see page 30).

Specification of null hypotheses

Attention in a trial usually focuses on the null hypothesis of treatment equivalence expressed by $\theta = 0$, but realistically this is often not the only hypothesis of interest. Increased costs, toxicity and so on may mean that a certain improvement would be necessary before the new treatment could be considered clinically superior, and we shall denote this value θ_S. Similarly, the new treatment might not actually be considered clinically inferior unless the true benefit were less than some threshold denoted θ_I. The interval between θ_I and θ_S has been termed the 'range of equivalence'[179] (see *Figure 3*): often θ_I is taken to be 0.

This is not a specifically Bayesian idea,[22] and there are several published examples of the elicitation and use of such ranges of equivalence.[172,180]

Using historical controls

A Bayesian basis for the use of historical controls in clinical trials, generally in addition to some contemporaneous controls, is based on the idea that it is wasteful and inefficient to ignore all past information on control groups when making a new comparison. This was first suggested by Pocock,[360] and has since been particularly developed within the field of carcinogenicity studies.[386] The crucial issue is the extent to which the historical information can be considered equivalent to contemporaneous data: essentially the inclusion of historical controls is indistinguishable from using such data as the basis for a prior opinion (see page 18).

Five broad approaches have been taken:

1. Assume the historical control **individuals** are exchangeable with those in the current control group, which leads to a complete pooling of historical with experimental controls.

2. Assume the historical control **groups** are exchangeable with the current control group, and hence build or assume a hierarchical model for the response within each group.[127,440] This leads to a degree of pooling between the control groups, depending on their observed or assumed heterogeneity. Gould[199] suggests using past trials to augment current control group information, assuming exchangeable control groups. Rather than directly producing a posterior distribution on the contrast of interest, he uses this historical information to derive predictive probabilities of obtaining a significant result were a full trial to have taken place (see page 28) (this paper carefully avoids the term 'Bayesian', and we might suspect this was a deliberate policy).

3. Assume that the parameter being estimated in historical data is some function of the parameter that is of interest, thus explicitly modelling potential biases, as in the confidence profile method[150] (see page 49).

   Atkinson[25] suggests this technique in a context where the historical control data is available on a general population, but experimental therapy is only given to an identifiable subset of new patients. A posterior distribution for the benefit of treatment can be obtained by constructing a model for what would have been the control response on that subset.

4. If there is only one historical group, then assume a parameter representing the probability that any past individual is exchangeable with current individuals, so discounting the contribution of past data to the likelihood, as used by Berry[44,45] in reanalysis of the extracorporeal membrane oxygenation (ECMO) study (see page 37).

5. Assume a certain prior probability that the historical control group exactly matches the contemporaneous controls and hence can be pooled. Racine *et al.*[370] provide a Bayes's factor formulation within the context of bioequivalence studies.

Design: sample size of non-sequential trials

Here we only deal with trials of fixed sample size: see page 29 for sequential designs.

We described a taxonomy of six different broad approaches to health technology assessment studies earlier (see page 14). Here we focus on how the four main concepts (ignoring the Bayesian hypothesis testing and the classical decision-theory approach) deal with selecting the size of an experiment.

1. **Fisherian.** In principle there is no need for preplanned sample sizes, but a choice may be made by selecting a particular precision of measurement and informally trading that off against the cost of experimentation.

2. **Neyman–Pearson.** Trials have traditionally been designed to have reasonable power, defined as the chance of correctly detecting an alternative hypothesis, best defined as being both 'realistic and important'. Power is generally set to 80 or 90%.

3. **Proper Bayesian.** As in the Fisherian approach, there is in principle no need for preplanned sample sizes, as pointed out by Lilford *et al.*[153,300,302] Alternatively, it is natural to focus on the eventual precision of the posterior distribution of the treatment effect. There is an extensive literature on non-power-based Bayesian sample size calculations which may be very relevant for trial design.[9,247,255,256]

   Considerable attention has been paid to a hybrid between Neyman–Pearson and Bayesian approaches, in which the prior might be used in the design but not in reporting the analysis. Thus the requirement for a trial being large enough to detect a plausible difference naturally leads to the use of prior distributions: either the prior mean could be taken as the alternative hypothesis or the power for each value of θ could be averaged with respect to the prior distribution to obtain an 'expected' or 'average' power, which should be a more realistic assessment of the chance that the trial will yield a positive conclusion.

   Prior distributions can thus be used for sample size calculations for trials that will be analysed in a traditional frequentist or Bayesian manner.[421] The prior distributions might be from any of the sources described in chapter 3, and examples include sets of subjective assessments,[150,180,414,418,442] a single previous study,[78] or a meta-analysis of previous results.[96,129]

   Brown[78] predicts the chance of correctly detecting a positive improvement, rather than the overall chance of getting a positive result, and this might be considered a more reasonable target. Predictive power can also be applied to null hypotheses defined as ranges of equivalence.[328] Gould considers nuisance parameters, for example overall response rate, and uses historical data[200] or interim data[201] for power calculations.

   It is natural to express a cautionary note on projecting from previous studies,[273] and possible techniques for discounting past studies are very relevant (see page 27).

4. **Decision-theoretic Bayesian.** If we are willing to express a utility function for the cost of experimentation and the potential benefit of the treatment, then sample sizes can chosen to

maximise the expected utility. This was considered by Canner,[88] while Thompson[452] estimated that a trial of electronic foetal monitoring with 180,000 cases per arm would cost US$22 million but would be expected to provide a benefit of US$118 million. Detsky[131] conducted an early attempt to model the impact of a trial in terms of future lives saved, which required modelling beliefs about the future number to be treated and the true benefit of the treatment, while Tan and Smith[439] fit the ideas in with ranges of equivalence. In other examples,[107,238,239] sample sizes are explicitly determined by a trade-off between the cost of the trial and the expected future benefit: see page 39 for some comments on the consequences. This approach also attempts to answer the question 'what is the expected net benefit from carrying out the trial?', which is discussed further on page 52.

Instead of attempting to model the future benefit of a trial, Lindley[305] considers a utility function based only on the final interval.

Design and monitoring of sequential trials

Introduction

Whether or not to stop a trial early is a complex ethical, financial, organisational and scientific issue, in which statistical analysis plays a considerable role. Furthermore the monitoring of sequential clinical trials can be considered the 'front line' between Bayesian and frequentist approaches, and Etzioni and Kadane state that the reasons for their divergence "reach to the very foundations of the two paradigms".[162] Pocock,[362] O'Brien[336] and Whitehead[478] provide good reviews.

Four main statistical approaches can be identified, again corresponding to the four main entries in *Table 4*.

1. **Fisherian.** This is perhaps best exemplified in trials influenced by Richard Peto's group, in which protocols state[109] that the data-monitoring committee should only alert the steering committee to stop the trial on efficacy grounds if there is "**both** (a) 'proof beyond reasonable doubt' that for all, or for some, types of patient one particular treatment is clearly indicated…, **and** (b) evidence that might reasonably be expected to influence the patient management of many clinicians who are already aware of the results of other main studies".

2. **Neyman–Pearson.** This classical method attempts to retain a fixed type I error through pre-specified stopping boundaries. These may be considered as stopping guidelines: Demets[126] states that "while they are not stopping rules, such methods can be useful in the decision-making process", although regulatory authorities require good reasons for not adhering to such boundaries.[248] Whitehead[476,477] is a major proponent, and Jennison and Turnbull[250] provide a detailed review.

3. **Proper Bayesian.** Probabilities derived from a posterior distribution may be used for monitoring, without formally prespecifying a stopping criterion – see page 30. There is no real need in this framework even to prespecify a sample size.[53] As for fixed sample size trials, prior distributions have been used at the design stage but assuming a Neyman–Pearson analysis.[180,321]

4. **Decision-theoretic Bayesian.** This assumes we are willing to explicitly assess the losses associated with consequences of stopping or continuing the study (see page 32), and therefore requires a full specification of the 'patient horizon', the allocation rule and so on. This approach also quantifies the expected benefit of the trial and therefore helps decide whether to conduct the trial at all.

Here we briefly describe some of the huge literature on this subject.

A brief critique of Neyman–Pearson methods in sequential trials

As introduced on page 10, the need to adjust conclusions because the data have been looked at during the study has been roundly criticised. Anscombe[13] baldly states that "Sequential analysis is a hoax", and Meier[323] considers that "provided the investigator has faithfully presented his methods and all of his results, it seems hard indeed to accept the notion that **I** should be influenced in my judgement by how frequently **he** peeked at the data while he was collecting it". The crucial technical point is that Neyman–Pearson theory disobeys the likelihood principle (see page 13), and hence there is no need for Bayesians or Fisherians to take any account of what would have happened had something other been observed.[50]

If we were to assign weights to the relative importance of the two types of error that could be made, any resulting design would seek to minimise a linear combination of the type I error rate α and type II error rate β. The fact that such a design

would obey the likelihood principle led Cornfield[111] to point out that "the entire basis for sequential analysis depends upon nothing more profound than a preference for minimising β for given α rather than minimizing their linear combination. Rarely has so mighty a structure, and one so surprising to scientific common sense, rested on so frail a distinction and so delicate a preference".

Freedman[177] points out that there is no agreed method of estimation following a sequential trial, and Hughes and Pocock[244,364] argue that frequentist sequential rules are "prone to exaggerate magnitude of treatment effect" since they would tend to stop when on a random high. However, Whitehead[478] says one can adjust to get rid of such biases, and Hughes and Pocock are really looking at inconsistency with clinical opinion.

Armitage[17] agrees that adjusted *P* values are "too tenuous to be quoted in an authoritative analysis of the data", but still considers frequency properties of stopping rules may be useful guides for "mental adjustment". Heitjan *et al.*[228] says the loss function implicit in a sequential analysis reflects a focus on inference rather than decisions.

Pocock and Hughes[364] say that "control of the overall type I error is a vital aid to restricting the flood of false positives in the medical literature", but this appears to introduce the extraneous issue of selective reporting.

From a practical perspective, the responsibility for recommending the early termination of a trial is increasingly vested in an independent data and safety monitoring committee, which will need to take into account multiple sources of evidence when making their judgements. Classical sequential analysis may be a useful warning to them against overinterpretation of naive *P* values.

Monitoring using the posterior distribution

Following the 'proper Bayesian' approach, it is natural to consider terminating a trial when one is confident that one treatment is better than the other, and this may be formalised by assessing the posterior probability that the treatment benefit θ lies above or below some boundary, such as the ends of the range of equivalence. It is generally recommended that a range of priors are considered, and applications have been reported in a variety of trials.[38,51,77,90,129,182,183,191,348,381] Explicit comparison with boundaries obtained by frequentist procedures have been displayed,[129,181,183] and the similar conservativism noted. Armitage[17] and Simon[405] discuss the choice of an appropriate boundary.

The informal stopping procedure described above explicitly takes into account the impact of the results on a range of clinical opinion, and so follow Kass and Greenhouse[267] in claiming that a successful trial should contain sufficient evidence to bring both a sceptic and an enthusiast to broadly the same conclusions (see page 26). This may be formalised by using the concept of sceptical and enthusiastic priors (see page 19), in which stopping with a positive result might be considered if a posterior based on a sceptical prior suggested a high probability of treatment benefit, whereas stopping without a negative result may be based on whether the results were sufficiently disappointing to make a posterior based on an enthusiastic prior rule out a treatment benefit: in other words we should stop if we have convinced a reasonable adversary that he or she is wrong. Fayers *et al.*[166] provide a tutorial on such an approach, and Freedman *et al.*[178] consider it as part of an exercise for a data-monitoring committee. Digman *et al.*[133] give an example in which the data have overwhelmed an optimistic prior centred on a 40% risk reduction, and hence justifies assuming a negative result and early stopping.

Greenhouse and Wasserman[206] and Carlin and Sargent[93,388] consider stopping rules based on 'robust priors': Carlin[89] argues that this enables statements of the form "given the data so far, the prior would have to place a mass of at least *p* on the range where the new treatment is considered superior, in order to avoid stopping now and rejecting this treatment as inferior". Emerson[159] retrospectively examines what class of priors would have been needed to replicate stopping used in a study (see page 20). Posterior probabilities of two responses can be monitored jointly, and stopping considered when an event of interest, such as either outcome occurring,[161] exceeds a certain threshold.

A specific problem occurs when additional information becomes available as a trial is continuing, such as the publication of similar studies. Brophy and Joseph[77] argue that this is a good opportunity for Bayesian methods, generating considerable discussion[271] and letters.

This monitoring scheme has also been proposed for single-arm studies, and is discussed within the context of Phase I and II trials (see page 38). Criticisms of this procedure include its sampling

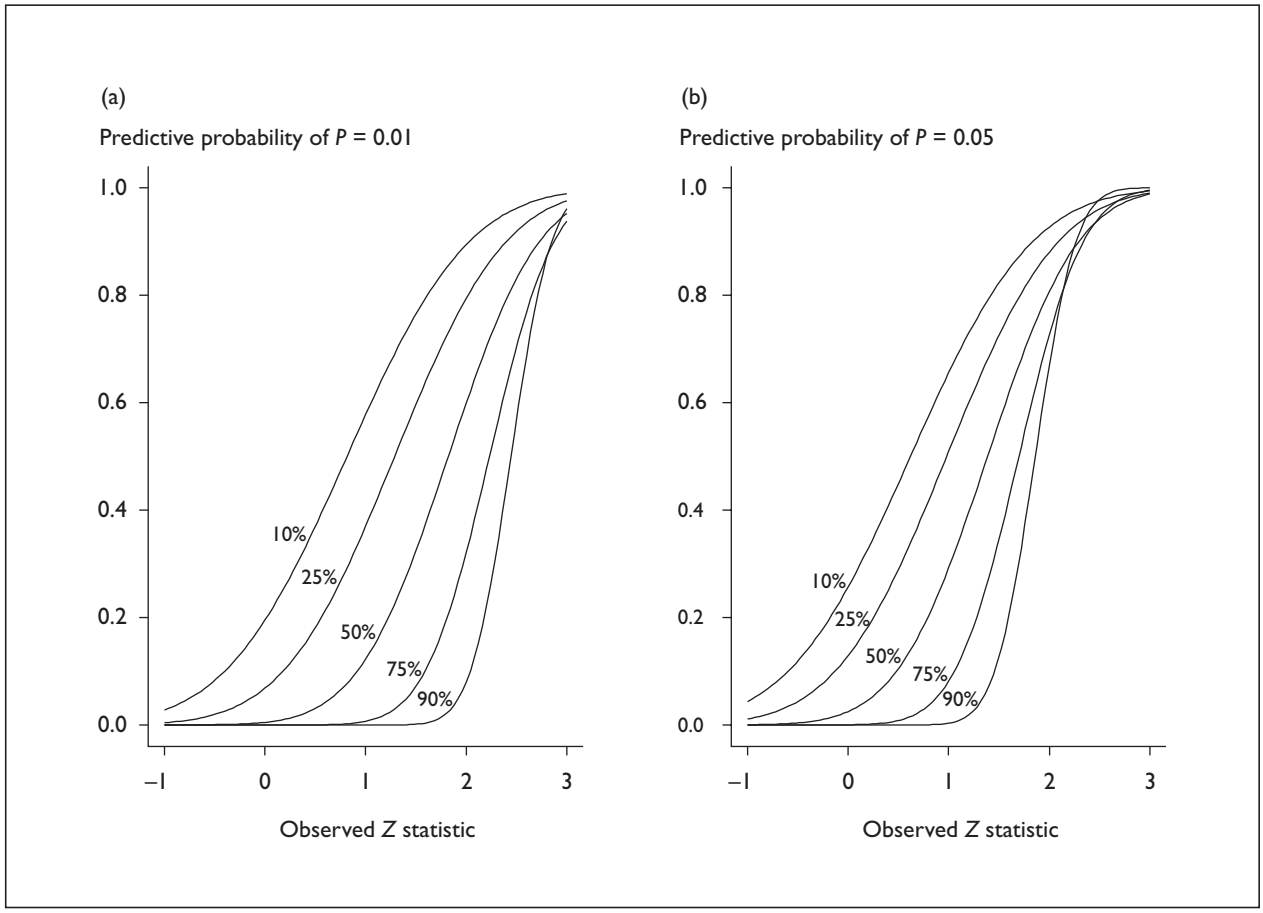


FIGURE 4 *Predictive probability of obtaining a significant (two-sided* P *= 0.01 or 0.05) result, given the fraction* f *of study completed (*f *= 10, 25, 50, 75 and 90%) and the current standardised test statistic* Z*. For example, if halfway through a study (*f *= 50%), the treatment effect is currently one standard error away from 0 (*Z *= 1), then* ***based on this information alone*** *there is only 29% chance that the trial will eventually show a significant (*P *< 0.05) benefit of treatment*

properties (see page 32), lack of explicit loss function (see page 32), and dependence on prior (see page 39).

Monitoring using predictions

Investigators and funders are often concerned with the question 'Given the data so far, what is the chance of getting a 'significant' result?' The traditional approach to this question is 'stochastic curtailment',[221] which calculates the conditional power of the study, given the data so far, for a range of alternative hypotheses.[250]

Digman *et al.*[133] point out that it is not, however, reasonable to condition on a hypothesis that is no longer tenable. From a Bayesian perspective it is natural to average such conditional powers with respect to the current posterior distribution, just as the pretrial power was averaged with respect to the prior to produce the average or expected power (see page 28). The methodology has been illustrated in a variety of contexts[103,104,286,368,419] in which initial trial results are used to predict the probability of eventually gaining 'significance'. Since data are available at an interim stage a minimally informative pretrial prior may be used, and hence the method does not strictly speaking require a **proper** Bayesian justification. Armitage[22] points out circumstances in which the predictions can be based on a pivotal quantity that does not depend on the parameter, and Lan and Wittes's[278] '*B* value' enables calculation of predictive probability of significance. Chang and Shuster[100] use non-parametric predictions of survival times without a prior, while Frei *et al.*[185] and Hilsenbeck[234] provide practical examples of stopping studies due to the futility of continuing.

The fact that predicted probabilities of success are often surprisingly low has been emphasised and is shown in *Figure 4*.[420] The chance that the result will even change sign may also be reported.[48] The technique has been used with results that currently show approximate equivalence to justify the 'futility' of continuing,[470] and may be particularly useful for data-monitoring committees and funders[275]

when accrual or event rates are lower than expected.[5]

Nevertheless, Armitage[17,18,22] warns against using this predictive procedure as any kind of formal stopping rule as it gives an undue weight to 'significance', and makes strong assumptions about the direct comparability of future data with those already observed – for example if future data involve extended follow-up there may be undue reliance on an assumption of proportional hazards.[421]

Monitoring using a formal loss function

The full Bayesian decision-theoretic approach requires the specification of losses associated with all combinations of possible true underlying states and all possible actions. The decision whether to terminate a trial is then, in theory, based on whether termination has a lower expected loss than continuing, where the expectation is with respect to the current posterior distribution, and the consequences of continuing have to consider all possible future actions – this 'backwards induction' requires the computationally intensive technique of 'dynamic programming'.

Reasonably straightforward solutions can be found in some circumstances. For example, Anscombe[13] considers n pairs of patients randomised equally to two groups, a total patient horizon of N, a uniform prior on true treatment benefit, and loss function proportional to the number of patients given the inferior treatment times the size of the inferiority: he concludes it is approximately optimal to stop and give the 'best to the rest' when the standard one-sided P value is less than n/N – half the proportion of patients already randomised. Berry and Pearson[60] and others[59,120,228] have extended such theory to allow for unequal stages and so on. Backwards induction is extremely computationally demanding, but Carlin *et al.*[91] do a retrospective analysis on a trial,[90] and claim it is computationally feasible using MCMC methods, in which forward sampling is used as an approximation to the optimal strategy. There is also an extensive theoretical literature on trials designed from a non-Bayesian decision-theoretic perspective.[30]

As a worked (but retrospective) example, Berry *et al.*[64] consider a trial of influenza vaccine for Navajo children. They construct a model consisting of priors for the effectiveness of the vaccine and the placebo treatment, the probability of obtaining regulatory approval and the length of time taken to obtain it, and the probability of a superior vaccine appearing in the next 20 years and the length of time taken for it to appear. After each month the expected number of cases of the strain amongst Navajo children in the next 20 years is calculated in the case of stopping the trial, and continuing the trial (the latter being calculated by dynamic programming). The trial is stopped when the former exceeds the latter.

The level of detail required for such an analysis has been criticised as being unrealistic,[72,421] but it has been argued that trade-offs between benefits for patients within and outside the trial should be explicitly confronted.[162] See page 39 for further discussion.

A recent development is reported by Kadane *et al.*,[264] who are to be allowed to elicit prior distributions and utilities from members of the data-monitoring committee for a large collaborative cancer trials group National Surgical Adjuvant Breast and Bowel Project (NSABP). They intend to do individual elicitations using predictive methods (see page 17) and use the forward sampling approach to solve the dynamic programming problem.[91] Their success at this ambitious venture remains to be seen.

Frequentist properties of sequential Bayesian methods

Although the long-run sampling behaviour of sequential Bayesian procedures is irrelevant from a strict Bayesian perspective, a number of investigations have taken place which generally show good sampling properties.[236,293,294,381] In particular, Grossman *et al.*[217] show that a sceptical prior (see page 19), centred on zero and with precision equivalent to that arising from an 'imaginary' trial of around 26% of the planned sample size, gives rise to boundaries that have type I error around 5% for five interim analyses, with good power and expected sample size. Thus an 'off-the-shelf' Bayesian procedure has been shown to have good frequentist properties: essentially the conservative behaviour of a Neyman–Pearson approach is mirrored by that obtained from assuming a sceptical prior. The sampling properties of Bayesian designs has been particularly investigated in the context of Phase II trials (see page 28).

One contentious issue is 'sampling to a foregone conclusion'.[21] This mathematical result proves that repeated calculation of posterior tail areas will, **even if the null hypothesis is true**, eventually lead a Bayesian procedure to reject that null hypothesis. This does not, at first, seem an attractive frequentist property of a Bayesian procedure.

Nevertheless, Cornfield[111] argued that "if one is seriously concerned about the probability that a stopping rule will certainly result in the rejection of a null hypothesis, it must be because some possibility of the truth of the hypothesis is being entertained", and if this is the case then they should be placing a lump of probability on it, as discussed on page 19, and so fit within the Bayesian hypothesis testing framework (see page 11). He shows that if such a lump, however small, is assumed then the problem disappears in the sense that the probability of rejecting a true null hypothesis does not tend to one. Breslow[72] agrees with this solution, but Armitage is not persuaded,[16] claiming that even with a continuous prior distribution with no lump at the null hypothesis, one might still be interested in type I error rates at the null as giving a bound to those at non-null values.

The role of 'scepticism' in confirmatory studies

After a clinical trial has given a positive result for a new therapy, there remains the problem of whether a confirmatory study is needed. Fletcher *et al.*[173] argue that the results of the first trial might be treated with scepticism, and Berry[57] points out that using a sceptical prior is a means of dealing with 'regression to the mean', in which early extreme results tend to return to the average over time.

Example

Parmar *et al.*[349] consider two trials and show that, when using a reasonable sceptical prior, doubt can remain after both the first trial and the confirmatory trial about whether the new treatment provides a clinically important benefit.

Figure 5 shows an example with a sceptical prior distribution with a median of 0 months benefit, which is equivalent to an 'imaginary' trial in which 33 patients died on each treatment. The dashed vertical lines display the null hypothesis of no improvement and the minimum clinically worthwhile improvement of 4 months: between these lie what can be termed the 'range of equivalence', and the figure shows that the sceptical prior expresses a probability of 41% that the true benefit lies in the range of equivalence, and only 9% that the new treatment is clinically superior.

The likelihood plot shows the inferences to be made from the data alone, assuming a 'uniform' prior on the range of possible improvements: Parmar *et al.* call this an 'enthusiastic' prior. The probability that the new treatment is actually inferior is 0.004 (equivalent to the one-sided *P* value of 0.008/2.) The probability of clinical superiority is 80%, which might be considered sufficient to change treatment policy.

The posterior plot shows the impact of the sceptical prior, in that the chance of clinical superiority is reduced to 44% – hardly sufficient to change practice. In fact, Parmar *et al.* report that the National Cancer Institute (NCI) Intergroup Trial investigators were unconvinced by the Cancer and Leukemia Group B (CALGB) trial due to their previous negative experience, and so carried out a further confirmatory study. They found a significant median improvement but of only 2.4 months, suggesting the sceptical approach might have given a more reasonable estimate.

Reporting, sensitivity analysis, and robustness

The only 'guidelines' available for reporting Bayesian analyses appear to be those of Lang and Secic:[281]

1. Report the pretrial probabilities and specify how they were determined.
2. Report the post-trial probabilities and their probability intervals.
3. Interpret the post-trial probabilities.

These seem very limited: see chapter 8 for an attempt to set more stringent standards for reporting.

An integral part of any good statistical report is a sensitivity analysis of assumptions concerning the form of the model (the likelihood). Bayesian approaches have the additional concern of sensitivity to the prior distribution, both in view of its controversial nature and because it is by definition a subjective assumption that is open to valid disagreement. As part of the general discussion of priors in chapter 3, the need to consider the impact of a 'community of priors'[267] was stressed, and three main types of 'community' may be identified:

1. **Discrete set.** Many case studies carry out sensitivity analysis to a limited list of possible priors, possibly embodying scepticism, enthusiasm, clinical opinion and 'ignorance': the studies described in appendix 1 provide many examples.

2. **Parametric family.** O'Rourke[343] emphasises that posterior probabilities "should be clearly and primarily stressed as being a 'function' of the prior probabilities and not **the** probability of treatment effects". If the community of priors can be described by one varying parameter, then it is possible to graphically display the dependence of the main conclusion to that parameter. Hughes[242] suggested examining sensitivity of conclusions to priors based on previous trial results and that reflecting investigators' opinions, and later[243] gives an example which features a point-mass prior on zero, and an explicit plot of the posterior probability against the prior probability of this null hypothesis. Hughes's approach of plotting prior and against posterior summaries is an example of the 'robust' Bayesian approach, in which an attempt is made to characterise the class of priors leading to a given decision (see appendix 1).

3. **Non-parametric family.** The 'robust' Bayesian approach has been further explored by allowing the community of priors to be a non-parametric family in the neighbourhood of an initial prior (see page 20). For example, Gustafson[219]

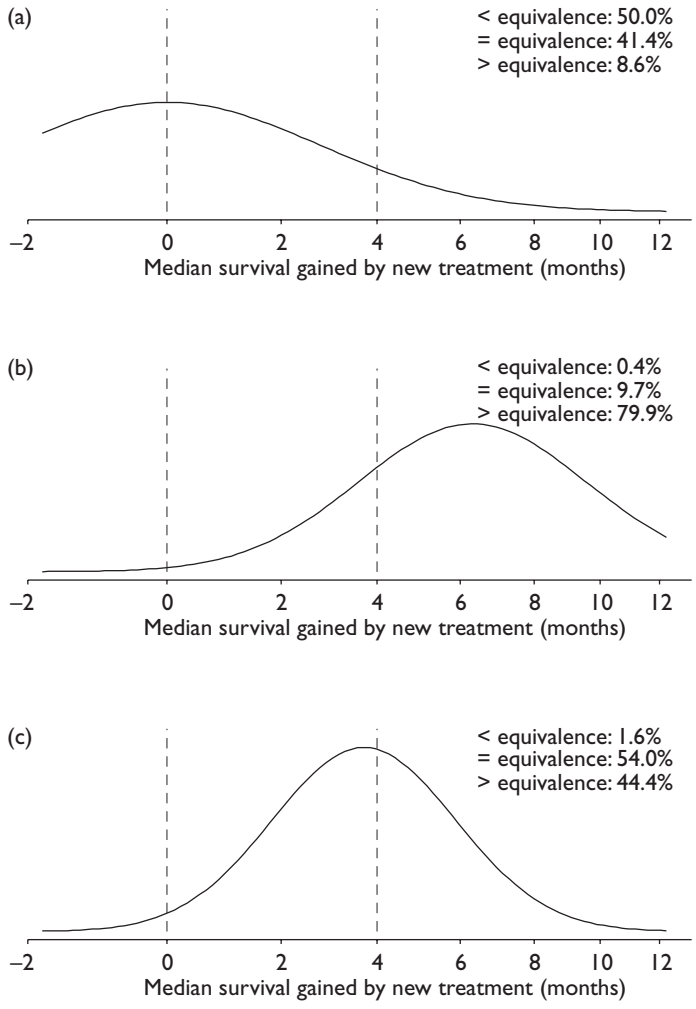


FIGURE 5 *(a) Sceptical prior (equivalent to 33 deaths in each group), (b) likelihood (observed hazard ratio = 1.63 after 120 deaths) and (c) posterior distributions arising from the CALGB trial of standard radiotherapy versus additional chemotherapy in advanced lung cancer. The dashed lines give the boundaries of the range of clinical equivalence, taken to be 0–4 months median improvement in survival. Percentages by each graph show the probabilities of lying below, within and above the range of equivalence. (Reproduced by permission of the BMJ from Spiegelhalter* et al.[426]*)*

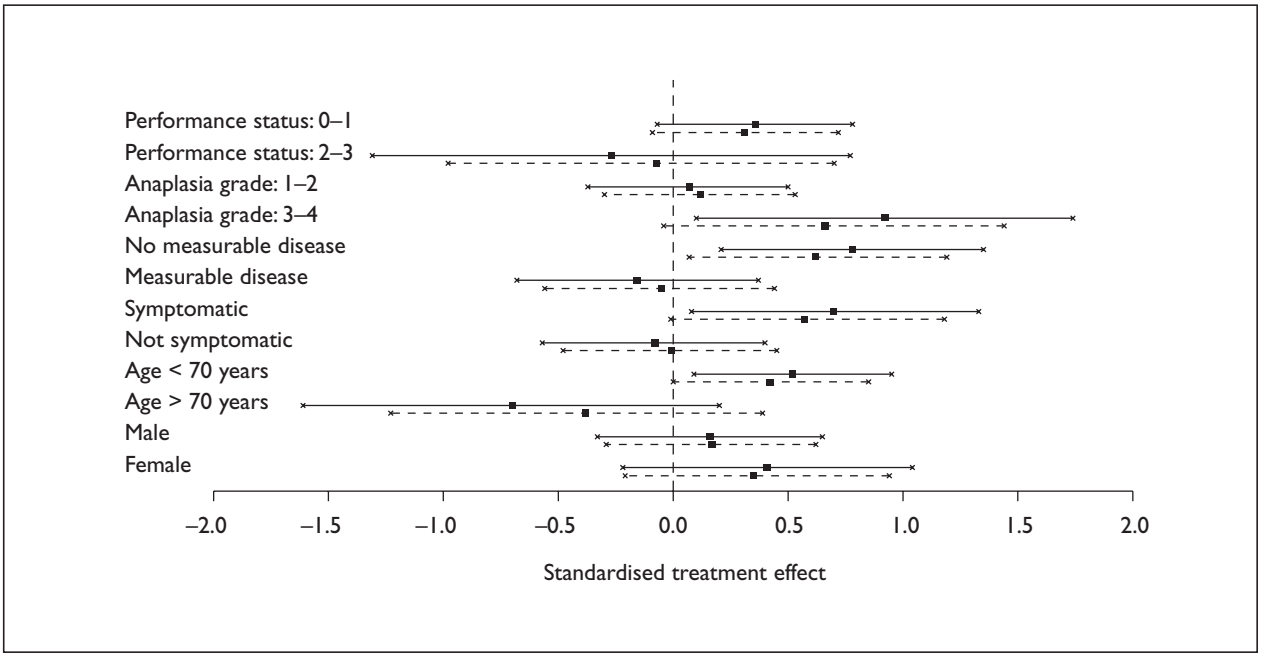


FIGURE 6 *Traditional (solid line) and Bayesian (broken line) estimates of standardised treatment effects in a cancer clinical trial. The Bayesian estimates are pulled towards the overall treatment effect by a degree determined by the empirical heterogeneity of the subset results. (Reproduced by permission of the BMJ from Spiegelhalter* et al.[426]*)*

considers the ECMO study (see page 37) with a community centred around a 'non-informative' prior but 20% 'contaminated' with a prior with minimal restrictions, such as being unimodal. The maximum and minimum posterior probability of the treatment's superiority within such a class can be plotted, providing a sensitivity analysis. Such an approach has also been explored by Greenhouse and Wasserman[206] and Carlin and Sargent.[93]

Stangl and Berry[426] emphasise the need for a fairly broad community, with sensitivity analysis not just to the spread of the prior but also its location, while they also stress that sensitivity to exchangeability and independence assumptions should be examined. In fact, the 'community' approach to prior distributions of treatment effects can also be applied to the distribution of random effects in hierarchical models: for example, Gustafson[219] considers a non-parametric approach to the effect of non-normality in a random effects distribution. Stangl and Berry[426] finish by saying that while sensitivity analysis is important, it should not serve as a substitute for careful thought about the form of the prior distribution.

There is limited experience of reporting such analyses in the medical literature, and it has been suggested by Koch,[272] Hughes[242] and Spiegelhalter *et al.*[421] that a separate 'interpretation' section is required to display how the data in a study would add to a range of currently held opinions. It would be attractive for people to be able to carry out their own sensitivity analysis to their own prior opinion: Lehmann and Shachter[290] describe a computing architecture for this, and available software and web pages are described in appendix 2.

Subset analysis

The discussion on multiplicity on page 12 has already described how multiple simultaneous inferences may be made by assuming a common prior distribution with unknown parameters, provided an assumption of exchangeability (the prior does not depend on units' identities) is appropriate. Within the context of clinical trials this has immediate relevance to the issue of estimating treatment effects in subgroups of patients.

By placing an identical prior on each subgroup, the unknown between-subgroup variability is essentially estimated from the data. As Cornfield[113] points out, this procedure "leads to 1) pooling subgroups if the differences among them appear small, 2) keeping them separate if differences appear large, and 3) providing intermediate results for intermediate situations".

Published applications are generally based on assigning a reference (uniform) prior on the overall treatment effect, and then assuming the subgroup-specific deviations from that overall effect have a common prior distribution with zero mean. This prior expresses scepticism about widely differing subgroup effects, although the variability allowed by the prior is usually estimated from the data. Thus this specification avoids the need for detailed subjective input, which may be seen as an attractive feature. Many applications consider this an empirical Bayes's procedure[123,309,365] which gives rise to traditional confidence intervals which are not given a Bayesian interpretation. Donner[139] sets out the basic ideas, and Dixon, Simon and co-authors elaborate the techniques in several examples.[134,135,136,403,405,406] Dixon and Simon[134] discuss the reasonableness of the exchangeability assumption.

Example: Bayes's theorem for subset analysis

Dixon and Simon[135] describe a Bayesian approach to dealing with subset analysis in a clinical trial in advanced colorectal cancer. The solid horizontal lines in *Figure 6* show the standardised treatment effects within a range of subgroups, using traditional methods for estimating treatment by subgroup interactions. Four of the 12 intervals exclude zero, although due to the multiple hypotheses being tested an adjustment technique such as Bonferroni might be used to decrease the apparent statistical significance of these findings.

The Bayesian approach is to assume subgroup-specific deviations from the overall treatment effect have a prior distribution centred at zero but with an unknown variability: this variability parameter is then given its own prior distribution. Since the degree of scepticism is governed by the variance of the prior distribution, the observed heterogeneity of treatment effects between subgroups will influence the degree of scepticism being imposed.

The resulting Bayesian estimates are shown as dashed lines in *Figure 6*. They tend to be pulled towards each other, due to the prior scepticism about substantial subgroup-by-treatment interaction effects. Only one 95% interval now excludes zero: the subgroup with non-measurable metastatic disease. Dixon and Simon mention that this was the conclusion of the original trial, but that the Bayesian analysis has the advantage of not relying on somewhat arbitrary adjustment techniques, being generalisable to any number of subsets, and provides a unified means of both providing estimates and tests of hypotheses.

Multicentre analysis

Methods for subset analysis (see page 35) naturally extend to multicentre analysis, in which the centre-by-treatment interaction is considered as a random effect drawn from some common prior distribution with unknown parameters. Explicit estimation of individual institutional effects may be carried out, which in turn relates strongly to the methods used for institutional comparisons of patient outcomes (see page 44).

There have been numerous examples of this procedure,[203,218,409,425,427,428] generally adopting MCMC techniques due to the intractability of the analyses. Recent case studies include Gould,[202] who provides BUGS code (see appendix 2) for Gibbs sampling analysis, and Jones *et al.*,[251] who compare estimation methods. Matsuyama *et al.*[316] allow a random centre effect both on baseline hazard and treatment effect, and examine the centres for outliers using a Student's t prior distribution for the random effects.

Senn[398] discusses the general issue of when a random-effects model for centre-by-treatment interaction is appropriate, emphasising the possible difficulty of interpreting the conclusions, particularly in view of the somewhat arbitrary definition of 'centre'.

Multiple end-points and treatments

Multiple end-points in trials can often be of interest when dealing with, say, simultaneous concern with toxicity and efficacy. This tends to occur in early phase studies, and a Bayesian approach allows one to create a two-dimensional posterior distribution over toxicity and efficacy.[137,161,451] General random effects models for more complex situations can be constructed.[288] Naturally, a two-dimensional prior is required, and particular care must be taken over the dependence assumptions.

A similar situation arises with many treatments: if one is willing to make exchangeability assumptions between treatment effects, then a hierarchical model can be constructed to deal with the multiple comparison problem. This was proposed long ago by Waller and Duncan.[468] Brant *et al.*[71] update this procedure by assuming exchangeable treatments and setting the critical values for the posterior probabilities of treatment effects by using a

decision-theoretic argument based on specifying the relative losses for type I to type II error.

Data-dependent allocation

'Bandit' problems

So far we have only covered standard randomisation designs, but a full decision-theoretic approach to trial design would consider data-dependent allocation so that, for example, in order to minimise the number of patients getting the inferior treatment, the proportion randomised to the apparently superior treatment could be increased as the trial proceeded. Such 'adaptive' designs would appear to satisfy ethical considerations for the patients under study[303] (see page 25). These designs can pose so-called 'bandit' problems, as they are analogous in theory to a gambler deciding which arm of a two-armed bandit to pull in order to maximise the expected return. An extreme example is Zelen's 'play-the-winner' rule in which the next patient is given the currently superior treatment, and randomisation is dispensed with entirely;[485] there is an extensive literature on this and other designs.[47]

There has been considerable criticism of these ideas as not being practically rooted in the realities of clinical trials. Byar *et al.*[86] identify as objections: (1) responses have to be observed without delay, (2) adaption depends on a one-dimensional response, (3) sample sizes may have to be bigger and (4) patients may not be homogeneous throughout the trial. Armitage[15] and Peto[357] add that clinicians are likely to be unhappy with adaptive randomisation, the trial will be complex and may deter recruitment, and estimation of the treatment contrast will lose efficiency. Palmer and Rosenberger[346] suggest that unbalanced allocation will make blinding difficult as clinicians may guess which treatment is 'in the lead', but point out that modern technology is making such designs more feasible. Finally, Senn[398] emphasises that future patients, greatly outnumbering those in the trial, would value a more precise treatment estimate.

A careful analysis has been carried out by Berry and Eick,[58] who conclude that such adaptive designs are more likely to yield a large improvement in the expected number of successful treatments when a large proportion of patients with the disease are likely to be in the trial. Tamura *et al.*[438] report one of the few adaptive clinical trials to take place, in patients with depressive disorder: the trial designed by Kadane[258] also adapts its allocation rules, in a somewhat complex way, to the current evidence.

Although there are specific aspects of adaptive allocation that cause practical problems, it is possibly the formulation of a trial as a decision rather than an inference that leads to most objections – see page 39.

The ECMO studies

Two studies of extracorporeal membrane oxygenation (ECMO) have been considered by many authors. ECMO is a rescue therapy for severe lung disease in newborns which carries the chance of greatly improved survival but at the risk of side-effects. An initial Zelen play-the-winner study was followed by the Harvard two-phase adaptive study designed to minimise the number of infants exposed to an inferior treatment: the design was to randomise between ECMO and conventional medical treatment (CMT) in balanced blocks of four, until one treatment had four deaths, at which point all subsequent patients were to receive the currently superior therapy. Nineteen patients were enrolled in the first phase, with 4/10 deaths under CMT and 0/9 deaths on ECMO – in the second phase,[20] patients were randomised to ECMO and one died. Interest has focused on what might have been concluded during the first phase of the study, and Bayesian analyses have been provided by Ware,[469] Berry,[45,46,51] Kass and Greenhouse,[267] Greenhouse and Wasserman,[206] and Gustafson.[219] Each of these is considered in detail in appendix 1.

Ware[469] and his discussants[46,47,267] provide a range of views on the ethics and appropriate analysis of this trial, and how existing information both about ECMO and the control therapy could have been used in formulating a prior opinion.

Trial designs other than two parallel groups

Equivalence trials

There is a large statistical literature on trials designed to establish equivalence between therapies. From a Bayesian perspective the solution is straightforward: define a region of equivalence (see page 27) and calculate the posterior probability that the treatment difference lies in this range – a threshold of 95 or 90% might be chosen to represent strong belief in equivalence. Several examples of this remarkably intuitive approach have been reported.[31,72,174,211,314,332,370,394,395] Racine *et*

al.[371] consider two-stage designs, Metzler[324] studies sample size while Lindley[308] takes a decision-theoretic approach.

Cross-over trials

The Bayesian approach to cross-over designs, in which each patient is given two or more treatments in an order selected at random, is fully reviewed by Grieve.[212] More recent references concentrate on Gibbs sampling approaches[175] and sensitivity analysis to different assumptions about carry-over effects.[11,209,210,214,215,370]

N-of-1 trials

N-of-1 studies can be thought of as repeated cross-over trials in which interest focuses on the response of an individual patient, and Zucker *et al.*[488] pool 23 *N*-of-1 studies using a hierarchical model producing the standard shrinkage of the individual conclusions towards the average result. This can be thought of as an extreme example of the subset procedure described previously, in which the subsets have been reduced to individual patients.

Factorial designs

Factorial trials, in which multiple treatments are given simultaneously to patients in a structured design, can be seen as another example of multiplicity, and hence a candidate for hierarchical models. Simon and Freedman[407] and Miller and Seaman[325] suggest suitable prior assumptions that avoid the need to decide whether interactions do or do not exist.

Other aspects of drug development

Pharmacokinetics

The 'population' approach to pharmacokinetics, in which the parameters underlying each individual's drug clearance curve are viewed as being drawn from some population, is well established and is essentially an empirical Bayes procedure. Proper Bayesian analysis of this problem has been extensively reported by Wakefield and colleagues,[372,466] emphasising MCMC methods for estimating both population and individual parameters, as well as individualising dose selection.[467]

Phase I trials

Phase I trials are conducted to determine that dosage of a new treatment which produces a level of risk of a toxic response which is deemed to be acceptable. The primary Bayesian contribution to the development of methodology for Phase I trials has been the continuous reassessment method (CRM) developed by O'Quigley and colleagues.[342] In CRM a parameter underlying a dose–toxicity curve is given a proper prior which is updated sequentially and used to find the current 'best' estimate of the dosage which would produce the acceptable risk of a toxic event if given to the next subject, as well as giving the probability of a toxic response at the recommended dose at the end of the trial.[340] High sensitivity of the posterior to the prior distribution[187] has been reported in a similar procedure. Numerous simulations and modifications of the method have been proposed.[10,102,163,197,222,274,326,329,341,445,479]

Etzioni and Pepe[161] suggest monitoring a Phase I trial with two possible adverse outcomes via the joint posterior distribution of the probabilities of the two outcomes with frequentist inference at the end of the trial.

Phase II trials

Phase II clinical trials are carried out in order to discover whether a new treatment is promising enough (in terms of efficacy) to be submitted to a controlled Phase III trial, and often a number of doses may be compared. Bayesian work has focused on monitoring and sample size determination. Monitoring on the basis of posterior probability of exceeding a desired threshold response rate has been recommended by Mehta and Cain,[322] while Heitjan[227] adapts the proposed use of sceptical and enthusiastic priors (see page 30) in Phase III studies. Korn *et al.*[276] consider a Phase II study which was stopped after three out of four patients exhibited toxicity; Bring[75] and Greenhouse and Wasserman[206] re-examine their problem from a Bayesian perspective.

Herson[231] used predictive probability calculations to select among designs with high power in regions of high prior probability. Thall and co-workers have also developed stopping boundaries for sequential Phase II studies based on posterior probabilities of clinically important events, but where the designs are selected from the frequentist properties derived from extensive simulation studies.[160,408,444,446,447,448,449,450] However, Stallard[424] has criticised this approach as being demonstrably suboptimal when evaluated using a full decision-theoretic model with a monetary loss function.

Finally, Whitehead and colleagues[81,473,474,477] have taken a full decision-theoretic approach to allocating subjects between Phase II and Phase III studies. For example, Brunier and Whitehead[81]

TABLE 6 *A brief comparison of Bayesian and frequentist methods in clinical trials*

Issue	Frequentist	Bayesian
Information other than that in the study being analysed	Informally used in design	Used formally by specifying a prior probability distribution
Interpretation of the parameter of interest	A fixed state of nature	An unknown quantity which can have a probability distribution
Basic question	'How likely is the data given a particular value of the parameter?'	'How likely is a particular value of the parameter given the data?'
Presentation of results	Likelihood functions, *P* values, confidence intervals	Plots of posterior distributions of the parameter, calculation of specific posterior probabilities of interest, and use of the posterior distribution in formal decision analysis
Interim analyses	*P* values and estimates adjusted for the number of analyses	Inference not affected by the number or timing of interim analyses
Interim predictions	Conditional power analyses	Predictive probability of getting a firm conclusion
Dealing with subsets in trials	Adjusted *P* values (e.g. Bonferroni)	Subset effects shrunk towards zero by a 'sceptical' prior.

consider the case where a single treatment with a dichotomous outcome is being evaluated for a possible Phase III trial, and use Bayesian decision theory to determine the number of subjects needed. They place a prior on the probability of success and calculate the expected cost of performing or not performing a Phase III trial, using a cost function which includes consideration of the costs to future patients if the inferior treatment is eventually used, the power of the possible Phase III trial (which they assume will be carried out by frequentist methods), and the costs of experimentation. They show how to determine, for given parameter values, the expected cost of performing a Phase II trial of any particular size, and thus the optimal size for a trial. (The work is said to correct earlier work of Sylvester and Staquet.[430,435,436,437])

When faced with selecting among a list of treatments and allocating patients, Pepple and Choi[355] have considered two-stage designs, Yao *et al.*[482] deal with screening multiple compounds and allocating patients within a programme, while Straus and Simon[432] use a prior distribution and horizon.

Phase IV – safety monitoring

A considerable literature exists on Bayesian causality assessment in adverse drug reactions: see for example Hsu *et al.*[241] and Lanctot *et al.*[279]

Commentary

Table 6 briefly summarises some major distinctions between the Bayesian and the frequentist approach to trial design and analysis.

Use of prior distributions

As well as the general issues concerning the specification and use of prior distributions (see page 22), specific questions arise in the context of clinical trials. These include:

1. **Whose prior?** Pocock[366] states that the "hardened sceptical triallist, the hopeful clinician and the optimistic pharmaceutical company will inevitably have grossly different priors", while Fisher[171] also emphasises the differing views of participants in health technology assessment.

   As mentioned on page 13, context is an essential aspect of any assessment, and chapter 7 describes the relevant perspectives of regulatory authorities and health policy makers.

2. **What about type I error?** A common objection to Bayesian methods is their apparent lack of concern with type I error, with the eventual certainty of rejecting a true null hypothesis.[249] Counter-arguments to this position were given on page 29 and 32.

3. **Stopping rule dependence.** A somewhat more subtle objection, well described by Rosenbaum and Rubin,[380] is that a Bayesian stopping rule based on posterior tail areas may be over-dependent on the precise prior distribution.[249] A possible response is that Bayesian stopping should not be based on a strict rule derived from a single prior, and instead a variety of reasonable perspectives investigated and a trial stopped only if there is broad convergence of opinion.

4. **What are the implications for the size of trials?** There is no single implication for trial size. Matthews[317] and Edwards *et al.*[153] have suggested that small, open, trials fit well into a Bayesian perspective in which all evidence contributes and there is no demand for high power to reject hypotheses.[300] Alternatively, monitoring with a sceptical prior may demand larger than standard sample sizes in order to convince an archetypal sceptic about treatment superiority.

The distinction made by Kadane and Seidenfeld[263] between 'experiments to learn' and 'experiments to prove' appears useful: trial design must naturally reflect one's objectives and different contexts may therefore demand differing designs.

Whitehead[475,476,478] argues that where a limited group of investigators are engaged on a project, for example in a drug development programme, then a full Bayesian approach with a loss function may be sensible, but that there is no place for prior opinions in publicly scrutinised Phase III studies.

As a final comment, O'Rourke[343] suggests that all methods have arbitrary aspects; a Bayesian approach has at least the simplicity of collecting all of this arbitrariness into the prior.

Use of a loss function: is a clinical trial for inference or decision?

There has been a long and intense dispute about whether a clinical trial should be considered as a decision problem, with an accompanying loss function, or as an inference.

1. **A clinical trial should be a decision.** Lindley[304] categorically states that "Clinical trials are not there for inference but to make decisions", while Berry[54] states that "deciding whether to stop a trial requires considering why we are running it in the first place, and this means assessing utilities". Healy[223] considers that "in my view the main objective of almost all trials on human subjects is (or should be) a decision concerning the treatment of patients in the future".

   Claxton[106,107] argues from an economic perspective that a utility approach to clinical trial design and analysis is necessary in order to prevent conclusions based on inferential methods leading to health or monetary losses. He points out that "Once a price per effectiveness unit has been determined, costs can be incorporated, and the decision can then be based on (posterior) mean incremental net benefit measured in either monetary or effectiveness terms". The assumptions that need to be made about rapid dissemination of superior treatments are reasonable, he claims, since we should be designing trials that are 'best' for a healthcare system: the need for incentives or arrangements to persuade decision makers is a separate issue.

   See page 52 where methods for evaluating the 'payback' of health research are discussed.

2. **A clinical trial provides an inference.** Armitage,[15] Breslow,[72] Demets,[125] Simon[402] and O'Rourke[343] all describe how it is unrealistic to place clinical trials within a decision-theoretic context, primarily because the impact of stopping a trial and reporting the results cannot be predicted with any confidence: Peto[357] states that "Bather, however, merely assumes … "it is implicit that the preferred treatment will then be used for all remaining patients" and gives the problem no further attention! This is utterly unrealistic, and leads to potentially misleading mathematical conclusions". Peto goes on to argue that a serious decision-theoretic formulation would have to model subsequent dissemination of treatment – attempts to do this will be discussed on page 52.

3. **It depends on the context.** Whitehead[478] points out that the theory of optimal decision-making only exists for a single decision-maker, and that no optimal solution exists when making a decision on behalf of multiple parties with different beliefs and utilities. He therefore argues that internal company decisions at Phase I and Phase II of drug development can be modelled as decision problems, but that Phase III trials cannot be.[476] Koch[272] also provides a non-dogmatic discussion, in which the relevant approach depends on the question being asked.

Key points

1. The Bayesian approach provides a framework for considering the ethics of randomisation.

2. Monitoring trials with a sceptical and other priors may provide a unified approach to assessing whether a trial's results would be convincing to a wide range of reasonable opinion, and could provide a formal tool for data monitoring committees.

3. Various sources of multiplicity can be dealt with in a unified and coherent way.

4. In contrast to earlier phases of development, it is generally unrealistic to formulate a Phase III trial as a decision problem, except in circumstances where future treatments can be accurately predicted.

5. An empirical basis for prior opinions in clinical trials should be investigated, but archetypal prior opinions play a useful role.

6. The structure in which trials are conducted must be recognised, but can be taken into account by specifying a range of prior opinions.

Chapter 5

A guide to the Bayesian health technology assessment literature: observational studies

Introduction

Since the probability models used in Bayesian analysis do not need to be based on actual randomisation, non-randomised studies can be analysed in exactly the same manner as randomised comparisons, perhaps with extra attention to adjusting for covariates in an attempt to control for possible baseline differences in the treatment groups with respect to uncontrolled risk factors or exposures. For example, the results from an epidemiological study, whether of a cohort or case–control design, provide a likelihood which can be combined with prior information using standard Bayesian methods. The dangers of this approach have been well described in the medical literature,[86] but nevertheless there are circumstances where randomisation is either impossible or where there is substantial valuable information available in historical data. There is, of course, a degree of subjective judgement about the comparability of groups, which fits well into the acknowledged judgement underlying all Bayesian reasoning. It is possible that the effect of unmeasured confounders can be modelled by means of a prior distribution, which could be considered as part of the explicit modelling of biases discussed on page 41. In this chapter we consider four aspects of non-randomised comparisons: case–control studies, complex epidemiological models, explicit modelling of biases, and comparisons of institutions on the basis of their outcomes. The discussion of historical controls within randomised trials (see page 27) is also relevant to the situation in which no contemporaneous controls are available.

Case–control studies

Case–control designs have been considered in detail by a number of authors,[24,313,335,486,487] generally relying on analytic approximations in order to obtain reasonably simple analyses. For example, Ashby *et al.*[24] examine two case–control studies (one being very small) and a cohort study of leukaemia following chemotherapy treatment for Hodgkin's disease, and consider the consequences of various prior distributions based on past studies, possibly downweighted. Lilford and Braunholtz's tutorial article concerns potential side effects of oral contraceptives, with likelihoods arising from case–control studies[299] (see appendix 1 for further details).

Complex epidemiological models

One approach to assessing the value of an intervention is to construct a model for the natural history of a chronic disease, and predict the consequences of implementing a specific policy. Such models can be developed by synthesising evidence from multiple sources (see chapter 6) in order to provide a 'comprehensive decision model' for cost-effectiveness analyses. However, Craig *et al.*[119] point out that predictions based on such a model require assumptions of parameter independence which do not need to be made if estimation and prediction are carried out simultaneously.

Such simultaneous analysis can be carried out if a large cohort database is available and the joint posterior distribution of the parameters of the model obtained, say through MCMC techniques. Craig *et al.*[119] describe such an analysis of a population-based cohort of patients with diabetic retinopathy in order to evaluate different screening policies. They construct a Markov model for transitions between disease states, using reasonably non-informative prior distributions and MCMC estimation.

There is also a substantial literature on Bayesian methods for complex epidemiological modelling, particular concerning problems spatial correlation,[23,40,226,377] measurement error[376] and missing covariate data.[373] Analysis is now almost universally by MCMC methods, and considerable use has been made of Bayesian graphical modelling techniques,[417] which are further explored in the Magnesium (see chapter 10) and confidence profile (see chapter 11) case studies.

Explicit modelling of biases

The fullest Bayesian analysis of non-randomised data for health technology assessment is probably the confidence profile method of Eddy and colleagues.[150] This is seen primarily as a tool for evidence synthesis (discussed in chapter 6), but also emphasises the ability to explicitly model potential biases that occur both within studies and in the attempt to generalise studies outside their target population.

They identify as 'biases to internal validity' those features of a study that may mean that the effect of interest is not being appropriately estimated within the circumstances of that study: these include dilution and contamination due to those who are offered a treatment not receiving it (see chapter 11), errors in measurement of outcomes, errors in ascertainment of exposure to an intervention, loss to follow-up, and patient selection and confounding in which the groups differ with respect to measurable features. These biases may occur singly or in combination.

'Biases to external validity' concern the ability of a study to generalise to defined populations or to be combined with studies carried out on different groups. These include 'population bias' in which the study and general population differ with respect to known characteristics, 'intensity bias' in which the 'dose' of the intervention is varied when generalised, and differences in lengths of follow-up. Finally, when combining studies it is possible to downweight the likelihood, much as historical control data can be downweighted when forming a prior (see page 27).

Eddy *et al.*[150] show how each of these biases can be given a mathematical formulation, and hence can be used to adjust the findings of any study. Two requirements are necessary. First, analytic solutions are rarely possible, and so approximations or simulations are necessary: chapter 11 shows how these are now reasonably straightforward. More serious are the necessary assumptions required concerning the extent of the biases. Data may be available on which to base accurate estimates, but there is likely to be considerable judgemental input. Any unknown quantity can, of course, be given a prior distribution, and Eddy *et al.* claim this obviates the need for sensitivity analysis. Rittenhouse[378] has argued that explicit allowance for external biases is necessary for cost-effectiveness studies.

Institutional comparisons

A classic 'multiplicity' problem arises in the use of performance indicators to compare institutions with regard to their health outcomes or use of particular procedures. Analogously to subset estimation (see page 35) and meta-analysis (see page 47), hierarchical models can be used to make inferences based on estimating a common prior or 'population' distribution.[195,333] An additional benefit of using MCMC methods (see page 12) is the ability to derive uncertainty intervals around the rank order of each institution.[312] Fully Bayesian methods have also been used in the analysis of panel agreement data on the appropriateness of coronary angiography.[27]

The case study in chapter 12 describes an analysis of success rates in *in vitro* fertilisation clinics, in which Bayesian methods are used both to make inferences on the true rank of each clinic, as well as estimating the true underlying success rates with and without an exchangeability assumption.

Commentary

To date there has been relatively little work done on Bayesian health technology assessment in an epidemiological setting compared with that in randomised controlled trials. One important reason for this is the often complex regression models that are used routinely in epidemiology to, for example, adjust for known confounders, with corresponding computational difficulties of Bayesian analysis. The advent of computer-intensive methods (see page 12) has largely overcome that problem.

In the future we can expect demands for increasingly complex analyses, such as of institutional comparisons and epidemiological data concerned with gene–environment interactions, which will place great demands on traditional hypothesis testing and estimation procedures which were essentially intended for fairly simple low-dimensional problems. This is likely to lead to increased demands for Bayesian analyses.

Key points

1. Epidemiological studies tend to demand a more complex analysis than randomised trials.

2. Computer-intensive Bayesian methods in epidemiology are becoming more common.

3. There are likely to be increased demands for Bayesian analyses, particularly in areas such as institutional comparisons and gene–environment interactions.
4. The explicit modelling of potential biases in observational data may be widely applicable but needs some evidence base in order to be convincing.

Chapter 6

A guide to the Bayesian health technology assessment literature: evidence synthesis

Meta-analysis

Bayesian meta-analysis follows the idea that has already been explored in other contexts of multiplicity, such as subset analysis (see page 35), multicentre trials (see page 36) and multiple outcomes (see page 36), in considering each trial to be a 'unit' of analysis within a hierarchical structure. The 'true' treatment effect in each trial is considered, assuming that the trials are exchangeable, as a random quantity drawn from some population distribution. This is exactly the same as the standard random-effects approach to meta-analysis,[130] but the latter tends to focus on estimating an overall treatment effect while a full Bayesian approach also concentrates on estimating trial-specific effects. The Bayesian approach also requires prior distributions to be specified for the mean effect size and for the between- and within-study variances: these will generally be default 'reference' priors, but including the uncertainty of all the parameters will tend to give wider interval estimates than a classical random effects analysis. Sutton *et al.*[434] review the whole area of meta-analysis and Bayesian methods in particular: other reviews are provided by Jones,[254] Normand[334] and Hedges.[225]

Empirical Bayes approaches have received most attention in the literature until recently, largely because of computational difficulties in the use of fully Bayesian modelling.[374,431] However, the full Bayesian hierarchical model has been investigated extensively by Dumouchel and colleagues[142,143,144,145] and Abrams and Samso[4] using analytic approximations, and also using MCMC methods.[327,412] Carlin,[94] for example, considers meta-analyses of both clinical trials and case–control studies; he examines the sensitivity to choice of reference priors, and explores checking the assumption of normal random effects. There have been many comparative studies of the full Bayesian approach, including trials,[379,433,455] and observational studies.[66,433,457] These comparative studies between different approximations show few substantial effects: the primary finding is that when there are few studies, and hence the between-study variability cannot be accurately estimated from the data alone, the prior for this parameter becomes important and the empirical Bayes approach, in which the uncertainty about the between-study variability is ignored, tends to provide intervals that are too narrow.

It is natural to use a cumulative meta-analysis as external evidence when monitoring a clinical trial,[230] and cumulative meta-analysis can also be given a Bayesian interpretation as providing a prior distribution[284] (see page 18): in this situation the Bayesian approach relies on the assumption of exchangeability of trials but avoids concerns with retaining type I error over the entire course of the cumulative meta-analysis.

Others have investigated relationship of treatment effect to underlying risk[320,393,453] – see the accompanying case study for an application of this approach (see chapter 10). Priors on the heterogeneity parameter were considered in chapter 3: Higgins and Whitehead[233] use proper priors derived from a series of meta-analyses. Another application is investigation of publication bias, which has been modelled by Begg *et al.*[34] and Givens *et al.*,[194] while Daniels and Hughes[122] pool studies in order to model a joint distribution of a surrogate end-point and eventual response.

Sutton *et al.*[434] summarise the potential advantages of the Bayesian approach to meta-analysis as including:

1. **Unified modelling.** The conflict between fixed and random effect meta-analysis is overcome by explicitly modelling between-trial variability (which could be assumed to be small), as well as allowing regression models for the treatment effect in each trial.

2. **Borrowing strength.** As in all areas in which Bayesian hierarchical modelling is adopted, the exchangeability assumption leads to each experimental unit 'borrowing' information from the other units, leading to a shrinkage of the estimate towards the overall mean, and a

reduction in the width of the interval estimate. This degree of pooling depends on the empirical similarity of the estimates from the individual units.

3. **Allowing for all parameter uncertainty.** The full uncertainty from all the parameters is reflected in the widths of the intervals for the parameter estimates.

4. **Allowing for other sources of evidence.** Other sources of evidence can be reflected in the prior distributions for parameters, or in pooling multiple types of study (see below).

5. **Allowing direct probability statements.** As with all Bayesian analyses, quantities of interest can be directly addressed, such as the probability that the true treatment effect in a typical trial is greater than 0.

6. **Predictions.** The ease of making predictions allows, for example, current meta-analyses to be used in designing future studies.

Cross-design synthesis

The previous section has dealt with meta-analysis of studies with similar basic design, but a more general approach allows mixing of different types of study. Rubin[384] emphasises pooling evidence through modelling in order to "build and extrapolate a response surface", which models the true treatment effect conditional on both the design of the study and on subgroup factors.

As noted when discussing observational studies (see chapter 5), in some circumstances randomised evidence will be less than adequate due to economic, organisational or ethical considerations.[67] Considering all the available evidence, including that from non-randomised studies, may then be necessary or advantageous. Droitcour *et al.*[140] describe the limitations of using either randomised controlled trials or databases alone, in that randomised controlled trials may be rigorous but restricted, whereas databases have a wider range but may be biased. They introduce what they term **cross-design synthesis**, an approach for synthesising evidence from different sources, with the aim "not to eliminate studies of overall low quality from the synthesis, but rather to provide the information needed to compensate for specific weaknesses".[140] Although not a Bayesian approach, they are following the explicit modelling of biases considered by the confidence profile method (see pages 44, and 49), and work on generalising the results of clinical trials for broader populations. Cross-design synthesis was outlined in a report from the US General Accounting Office,[337] but a *Lancet* editorial[152] was critical of this approach, suggesting it would deflect attention from carrying out serious controlled trials: this was denied in a subsequent reply.[101] A review by Begg[33] suggested they had underestimated the difficulty of the task, and appeared to assume that randomised trials and databases could be reconciled by statistical adjustments, whereas selection biases and differences in experimental rigour could not be eliminated so easily.

It is natural to take a Bayesian approach to the synthesis of multiple trial designs, and a hierarchical model can specifically allow for the quantit ative within- and between-sources heterogeneity, and for *a priori* beliefs regarding qualitative differences between the various sources of evidence. A full Bayesian version of cross-design synthesis was subsequently applied to data on breast cancer screening.[3,411] The concept of combining different types of study in a model has also been termed 'grouped meta-analysis'; Li and Begg[296] combine controlled and uncontrolled studies, Larose and Dey[283] integrate open and closed studies within a single model, while Dominici *et al.*[138] pool open and closed studies on migraine using a graphical model, different treatment contrasts and different designs. Muller *et al.*[330] combine case–control and prospective studies.

Other examples include Berry *et al.*,[62] who consider a complex synthesis of studies concerning breast cancer therapy, facing up to issues such as unplanned analyses, multiple variables, lack of exchangeability across and within studies, and the problem of convincing practitioners on the basis of such a complex analysis. Belin *et al.*[35] combine observational databases in order to evaluate interventions to increase screening rates, needing to impute missing data in some studies. Such integrated analyses naturally lead on to the 'comprehensive decision models' discussed in chapter 7.

It has yet to be established when such analyses are appropriate, as there is concern that including studies with poorer designs will weaken the analysis, though this issue is partially addressed by conducting sensitivity analyses under various assumptions. However, an example of such a synthesis is provided in the context of regulatory approval of medical devices (see page 53).

Confidence profile method

This approach was developed by Eddy and colleagues and promulgated in a book with numerous worked examples and accompanying software,[150] as well as tutorial articles.[146,148,149,151,399] They use directed conditional independence graphs to represent the qualitative way in which multiple contributing sources of evidence relate to the quantity of interest, explicitly allowing the user to discount studies due to their potential internal bias or their limited generalisability (see page 44). Their analysis is essentially Bayesian, although it is possible to avoid specification of priors and use only the likelihoods.

The software for carrying out the confidence profile method, FAST*PRO, has been used in meta-analysis of the benefits of antibiotic therapy,[28] mammography of those aged under 50 years[147] and angioplasty.[8]

The need to make explicit subjective judgements concerning the existence and extent of possible biases, and the limited capacity and friendliness of the software, has perhaps restricted the application of this technique. However, we show in chapter 11 that modern software can allow straightforward implementation of their models promulgated by Eddy and co-workers.

Key points

1. A unified Bayesian approach appears to be applicable to a wide range of problems concerned with evidence synthesis.

2. In the past, prospective evaluation of clinical interventions concentrated on randomised controlled trials, but more recent interest has focused on more diffuse areas, such as healthcare delivery or broad public health measures. This means methods that can synthesise totality of evidence are required, for example in assessing medical devices.

3. Evaluations of current technologies may often be seen as unethical subjects for randomised controlled trials, and hence modelling of available evidence is likely to be necessary.

4. Perhaps one reason for lack of uptake is that syntheses are not seen as 'clean' methods, with each analysis being context-specific, less easy to set quality markers for, easier to criticise as subjective and so on.

5. Priors for the degree of 'similarity' between alternative designs can be empirically informed by studies comparing the results of randomised controlled trials and observational data.

Chapter 7

A guide to the Bayesian health technology assessment literature: strategy, decisions and policy making

Contexts

Throughout this report we have emphasised that it was vital to take into account the context of health technology assessment is being made. The appropriate prior opinions, and the possibility of explicit loss functions, depend crucially on whose behalf any analysis is being reported or decision being made.

In this chapter we specifically address this issue using the broad categories of 'actors' introduced in chapter 2:

- **Sponsors,** for example the pharmaceutical industry, medical charities or granting agencies. In deciding whether to fund studies, they will be concerned with the potential 'payback' from research (see page 52), which within industry takes the form of a drug development programme.

- **Investigators,** that is, those responsible for the conduct of a study, whether industry or publicly funded. In previous chapters we have focused primarily on those carrying out a single study, who are primarily concerned with the accuracy of the inferences to be drawn from their work.

- **Reviewers,** for example regulatory bodies (see page 53) or journal editors. They will be concerned with the appropriateness of the inferences drawn from the studies, and so may adopt their own prior opinions and reporting standards (see page 33).

- **Consumers,** for example agencies setting health policy, clinicians or patients. Healthcare organisations may be concerned with the cost-effectiveness of an intervention (see page 51) although the sponsor or investigator may carry out this analysis on their behalf. Decisions about health policy, whether at a community or individual level, may involve explicit consideration of costs and benefits (see page 54) – here we distinguish between those based only on prior opinions (forming a decision analysis), and those requiring a full Bayesian statistical analysis.

In the remainder of this chapter we examine the potential Bayesian contribution to these different perspectives.

Cost-effectiveness within trials

The traditional tool for dealing with uncertainty in cost-effectiveness analysis has been sensitivity analysis, although there has been considerable recent work on developing classical confidence intervals for cost-effectiveness ratios. This area has recently been reviewed by Briggs and Gray,[74] who mention the use of **probabilistic sensitivity analysis**, in which prior probability distributions are placed over uncertain inputs into the analysis and the resulting distribution of potential cost-effectiveness ratios is generated by simulation. This is a particular case of a Bayesian method: the general way in which uncertainty is handled by the Bayesian approach has been emphasised by Manning *et al.*[311] and Jones.[252,253]

From a technical perspective, Grieve[216] shows how pharmaco-economics naturally gives rise to a bivariate posterior distribution of costs and effectiveness, which can be plotted and from which the probability of specific conclusions may be obtained. Heitjan *et al.*[229] illustrate this bivariate approach with a number of examples, while Briggs[73] avoids the problem of possible zero denominators in the cost-effectiveness ratio by working directly with the prior and likelihood for net benefit relative to a specified baseline ratio.

Luce and Claxton[310] present a strong argument that Bayesian methods should be applied in cost-effectiveness studies and pharmaco-economics in general. They point out that hypothesis testing is of limited relevance in economic studies, and that additional evidence outside a study is likely to be relevant. Furthermore, when a cost-effectiveness analysis is being used as one of the inputs into a

formal decision concerning drug regulation or health policy, then they recommend a full decision-theoretic approach in which an explicit loss function of the decision maker is assessed. This view has led to a Bayesian initiative in pharmacoeconomics (see appendix 2). Felli and Hazen[168,169] extend this utility perspective to sensitivity analysis, suggesting that an analysis should be considered sensitive to a particular uncertain input if the expected gain in utility from eliminating the uncertainty about that input exceeds a certain specified threshold. The use of such an **expected value of perfect information** (EVPI) approach has also been recommended when deciding research priorities (see page 52).

Finally, Rittenhouse[378] suggests that trial results may need to be adjusted to in order to generalise the cost-effectiveness analysis to other populations of interest. This essentially is concerned with the type of adjustments used in cross-design synthesis (see page 48) and the explicit modelling of biases in observational studies (see page 44).

Cost-effectiveness of carrying out trials – 'payback models'

Research planning in the public sector

Any organisation funding clinical trials must make decisions concerning the relative importance of alternative proposals, and there have been increased efforts to measure the potential 'payback' of expenditure on research. Buxton and Hanney[83] review the issues and propose a staged semi-quantitative structure, while Eddy[146] suggests a fully quantitative model based on assessing the future numbers to benefit and the expected benefit, with a subjective probability distribution over the potential benefits to be shown by the research. However, Eddy's[146] limited approach was not adopted by its sponsors, the US Institute of Medicine, who preferred a less structured model that employed weights.

It is clearly possible to extend this broad approach to increasingly sophisticated models within a Bayesian framework, and Hornberger and Eghtesady state that "by explicitly taking into consideration the costs and benefits of a trial, Bayesian statistical methods permit estimation of the value to a healthcare organisation of conducting a randomised trial instead of continuing to treat patients in the absence of more information".[238] Clearly this is a particular example of a decision-theoretic Bayesian approach, applied at the planning stage of a trial (see page 28) rather than at interim analyses (see page 32). Examples given on page 28 by Detsky,[131] Hornberger[238,239] and others explicitly calculate the expected utility of a trial in order to select sample sizes, and such calculations can also, in theory, be used to rank studies that are competing for resources, and hence to decide whether the trial is worth doing in the first place.

Detsky's early analysis[131] assumed that a trial would need to achieve statistical significance in order to have an impact on future treatments, but Claxton[106] strongly argues that dependence on such inferential methods, whether classical or Bayesian, will lead to suboptimal use of health resources. He recommends a full decision-theoretic approach to both fixed[107] and sequential[106] trials, basing his analysis on a univariate scale comprising the net benefit relative to a prespecified standard measure of effectiveness per unit cost. The EVPI must be higher than the cost of research in order to pass the first 'hurdle' for a proposed programme to overcome, and the expected value of sample information (EVSI) (essentially the EVPI allowing for the sampling error of a trial) must exceed the sample costs to overcome the hurdle for a specific proposed trial. This model allows for unbalanced allocation of patients between arms, and the ability to revise design based on interim analyses,[452] in order to optimise the expected net benefit from sampling (ENBS), which is the EVSI minus sample costs.

The standard criticisms of decision-theoretic approaches to trials apply (see page 40), particularly regarding unrealistic assumptions concerning the impact of research results (which may not even be 'significant') on clinical practice. Claxton replies that the first step should be to establish a normative framework that best meets the needs of a system, and separately to conduct studies to see how to get the research into practice.[106]

More generally, the role of decision analysis in all aspects of health services research has been emphasised by Lilford and Royston.[298]

Research planning in the pharmaceutical industry

Berry[49,59] has stressed the decision-theoretic approach to sequential trial design as being relevant to a a pharmaceutical company seeking to maximise profit. However, here we are concerned with a whole research programme in which there are multiple competing projects at different stages

of drug development. It is natural that the pharmaceutical industry would wish to organise its programme in a cost-effective way, and Bergman and Gittins[39] review quantitative approaches to planning a pharmaceutical research programme. Many of the proposed methods are sophisticated uses of bandit theory in order to allocate resources in a dynamically changing environment, but Senn[396,398] suggests a fairly straightforward scheme based on the Pearson index, which is the expected net present value divided by expected net present costs. He discusses the difficulties of eliciting suitable probabilities for the success of each stage of a drug development programme, conditional on the success of the previous stage, but suggests formal Bayesian approaches involving subjective probability assessment and belief revision should be investigated in this context.

The regulatory perspective

Regulation of pharmaceuticals

Regulatory bodies have a duty to protect the public from unsafe or ineffective therapies, and increasingly are taking on responsibility for assuring cost-effectiveness.

Opinions on the relevance of Bayesian methods to drug or device regulation cover a broad spectrum: Whitehead[478] and Koch[272] see any use of priors as being controversial and inappropriate, while on the other hand Matthews[319] claims that the use of sceptical priors "should not be optional but mandatory". Keiding[269] criticises the "ritual dances" currently prescribed for regulation, but wonders whether Bayesian methods will allow anything less ridiculous. Claxton[105] suggests that agencies take on a full decision-theoretic approach to regulation, that evaluates the expected value of further investigation in order to assess whether sufficient evidence is available to permit approval. O'Neill,[339] as a senior US FDA statistician, acknowledges the appropriate conservatism arising out of the use of sceptical priors, and considers that Bayesian methods should be investigated in parallel with other techniques.

The website of the FDA allows one to search for references to Bayesian methods among their published literature (see appendix 2). Much of the discussion concerns medical devices (see page 53). Guidelines for population pharmacokinetics are provided,[462] which can be thought of as an empirical Bayes procedure (see page 38). There is also an interesting use of a Bayesian argument in the approval of the drug enoxaparin (Lovenox®). The transcript of the Cardiovascular and Renal Drugs Advisory Committee meeting on 26 June 1997[461] shows the pharmaceutical company had been asked to make a statement about the effectiveness of enoxaparin plus aspirin as compared to placebo (aspirin alone), whereas their clinical trial had used an active control of heparin plus aspirin. They therefore used meta-analysis data comparing heparin plus aspirin versus aspirin alone in order to produce a posterior distribution on the treatment comparison of interest: an example of cross-study inference (see page 7.2). Analyses were repeated using the meta-analysis data directly, but also expressing scepticism about its relevance and reducing its influence, with results being expressed as posterior probabilities of treatment superiority over placebo. The committee welcomed this analysis and voted to approve the drug.

It is important to note that the latest international statistical guidelines for pharmaceutical submissions to regulatory agencies state that "the use of Bayesian and other approaches may be considered when the reasons for their use are clear and when the resulting conclusions are sufficiently robust".[248] Unfortunately, they do not go on to define what they mean by clear reasons and robust conclusions, and so it is still open as to what will constitute an appropriate Bayesian analysis for a pharmaceutical regulatory body.

Regulation of medical devices

The greatest enthusiasm for Bayesian methods appears to be in the FDA Center for Devices and Radiological Health (CDRH). They co-sponsored a workshop on Bayesian methods in November 1998, and are currently proposing to produce the document *Statistical Guidance on Bayesian Methods in Medical Device Clinical Trials.*[460]

Campbell recently described the potential for Bayesian methods in assessing medical devices,[87] emphasising that devices differed from pharmaceuticals in having better understood physical mechanisms, which meant that effectiveness was generally robust to small changes. Since devices tended to develop in incremental steps, a large body of relevant evidence existed, and companies did not tend to follow established phases of drug development. The fact that an application for approval might include a variety of studies, including historical controls and registries, suggests that Bayesian methods for evidence synthesis might be appropriate. However, the standard conditions apply that the source and robustness of the prior information must be

assessed, and that Bayesian analysis does not compensate for poor science and experimental design.

Campbell draws attention to the Transcan Breast Scanner®, which was approved by the CDRH in April 1999.[463] A primary 'intended use' study on 72 women was supplemented by two additional studies of differing designs, using a hierarchical multinomial logistic regression model with study introduced as a random effect. MCMC simulation methods were used by means of the BUGS software.[423]

Policy making and 'comprehensive decision modelling'

The primary advantage of a Bayesian approach is that it allows the synthesis of all available sources of evidence, whether from RCTs, databases or expert judgement, into a single model that can then be used to evaluate the cost-effectiveness of alternative policies. The approach has been termed 'comprehensive decision modelling', and can be thought of as extending the evidence synthesis methods described in chapter 6 to allow for costs in particular and utilities in general.

Parmigiani and colleagues[350,352] apply this idea to screening for breast cancer, in which many sources of evidence are brought together in a single model that predicts the consequences of alternative screening policies, while Cronin *et al.*[121] use micro-simulation at the level of the individual patient to predict the consequences of different policy decisions on lowering expected mortality from prostate cancer. Samsa *et al.*[387] consider ischaemic stroke, and construct a model for natural history using data from major epidemiological studies, and a model for the effect of interventions based on databases, meta-analysis of trials and Medicare claim records. They also use micro-simulation of the long-term consequences of different stroke prevention policies in order to compare their cost-effectiveness: Matchar, Parmigiani and colleagues[315,351,353] consider further use of the stroke prevention policy model (SPPM), developed under the auspices of the Stroke PORT (Patient Outcomes Research Team).

Key points

1. A Bayesian approach allows explicit recognition of multiple perspectives.

2. Increased attention to pharmaco-economics may lead decision-theoretic models for research planning to be explored, although this will not be straightforward.

3. There appears to be great potential for formal methods for planning in the pharmaceutical industry.

4. The regulation of devices is leading the way in establishing the role of evidence synthesis.

5. 'Comprehensive decision modelling' is likely to become increasingly important in policy making.

Chapter 8

BayesWatch: a Bayesian checklist for health technology assessment

In this chapter we present a checklist against which published accounts of Bayesian assessments of health technologies can be compared. We aim to ensure that an account which adequately contains all the points mentioned here would have the property the analysis could be replicated by another investigator who has access to the full data.

These guidelines should be seen as complementary to the CONSORT (Consolidated Standards of Reporting Trials) guidelines, in that they focus on those aspects crucial to an accountable Bayesian analysis, in addition to standard sections concerning the technology, the design and the results.

Introduction

1. **The technology.** The intervention to be evaluated must, of course, be clearly described with regard to the population of interest and so on.

2. **Objectives of study.** It is important that a clear distinction is made between desired inferences on any quantity or quantities of interest, representing the parameters to be estimated, and any decisions or recommendations for action that are to made subsequent to the inferences. The former will require a prior distribution, while the latter will require explicit or implicit consideration of a loss function/utility.

Methods

1. **Design of study.** This is a standard requirement, but when synthesising evidence, particular attention will be necessary to the similarity of studies in order to justify assumptions of exchangeability.

2. **Statistical model.** The probabilistic relationship between the parameter(s) of interest and the observed data should be explicitly described. The relationship should either be given mathematically, or its structure should be described in such a way as to allow its mathematical form to be unambiguously obtained by a competent reader. If this likelihood has been obtained by a method of model selection, whether Bayesian or not, this should be stated and the method described.

3. **Prospective analysis?** It needs to be made clear whether the prior and any loss function were constructed preceding the data collection, and whether analysis was carried out during the study.

4. **Loss function.** If an explicit method of deducing scientific consequences is decided prior to the study, this should be explicitly stated. This will often be a range of equivalence (a range of values such that if the parameter of interest lies in that range, two different technologies may be regarded as being of equal effectiveness), or a loss function whose expected value is to be minimised with respect to the posterior distribution of the parameter of interest yielding an estimated value of the parameter. If these have been obtained by an elicitation process from experts, this should be stated and the process described. Any intention to investigate the dependence of the final conclusion on the range of equivalence, etc., should be described.

5. **Prior distribution.** Explicit priors for the parameters of interest should be given, clearly showing whether an informative or 'non-informative' prior is being used. If they have been obtained by an elicitation process this should be stated and described. If it is intended to examine the effect of using different priors on the conclusion of the study, the alternative priors explicitly should be stated. Any empirical evidence underlying the prior assessment should be provided.

6. **Computations**. These need to be described to the extent that a mathematically competent reader could, if necessary, repeat all the

calculations and obtain the required results. Details of any software used to obtain the results should be given. If MCMC methods are being used, the choice of starting values, the number and length of runs and convergence diagnostics to be used should be clearly stated and justified

Results

1. **Evidence from study.** As much information about the observed data – sample sizes, measurements taken – as is compatible with brevity and data confidentiality should be given.

Interpretation

1. **Reporting.** The posterior distributions should be clearly summarised. In most cases, this should include a presentation of posterior credible intervals and a graphical presentation of the posterior distribution. If either a formal or informal loss function has been described, the results should be expressed in these terms.

   It is also essential that the likelihood can be reconstructed, usually through information given under 'evidence from study', so that subsequent users can establish the contribution from the study to, say, a meta-analysis. There should be a careful distinction between the report as a current summary for action, in which case a synthesis of all relevant sources of evidence is appropriate, and the report as a contributor of information for future action.

2. **Sensitivity analysis.** If alternative priors and/or expressions of the consequences of decisions have been given in the sections above, the results of these should be presented.

Example

The following is a single example from the literature summarised using the BayesWatch headings. A full list of 'three-star' Bayesian health technology assessment studies is provided in appendix 1.

Author. Abrams K, Ashby D and Errington D.[1]

Title. Simple Bayesian analysis in clinical trials – a tutorial.

Year. 1994.

The technology. High-energy neutron therapy against standard photon therapy in treatment of pelvic cancers.

Objectives of study. To estimate the odds ratio for 12 month survival.

Design of study. Randomised controlled trial. Separate trials were ran concurrently for cancers of the rectum, bladder, colon and cervix. Subjects were randomised in a ratio of 3:1 towards neutron therapy from 10 February 1986 until 11 January 1988 and then in a ratio of 1:1 until 12 February 1990.

Evidence from study. Twelve-month survival: 61 subjects on photon treatment, 36 alive, 25 dead; 90 subjects on neutron treatment, 44 alive, 46 dead.

Statistical model. Binomial model with 12 month mortality rate θ_p on photon treatment and θ_n on neutron treatment. Inference is on the log(odds ratio) $\log[(\theta_n(1-\theta_p)/\theta_p(1-\theta_n)]$.

Prospective analysis? Partly: prior distributions elicited prospectively, analysis performed retrospectively.

Loss function. No explicit loss function. An average of 10 clinicians demanded an odds ratio of less than 0.63 before routinely preferring neutron therapy.

Prior distribution. Ten clinicians were asked to specify their prior for 12 month mortality on neutron therapy by the roulette method assuming 12 month mortality on photon therapy to be 0.5. The arithmetic mean of these was taken and a beta prior superimposed by equating moments: $\theta_p \sim$ Beta(5.25, 6.17). The 12 month survival on proton therapy was taken to have a beta prior with a mean of 0.5 (specified by clinicians) and a variance of 0.01: $\theta_n \sim$ Beta(12, 12) (suggested by variability of previous studies).

Computations. Conjugate analysis for θ_p and θ_n, and normal approximation for the posterior of the log(odds ratio).

Reporting. Kaplan–Meier survival curves. Plots are shown of the prior, likelihood ('reference prior') and posterior. Medians, 95% credible intervals and probabilities of the odds ratio being less than 1, and less than the clinical demand (0.63), are given.

Sensitivity analysis. The analysis was repeated using a reference prior.

Comments. Likelihood conflicts with the prior distribution in the direction of effect. There is deliberate use of an analytically tractable model for tutorial purposes. This study is also considered by Spiegelhalter *et al.*[421] using a log(hazard ratio) scale.

Chapter 9

Case study 1: the CHART (lung cancer) trial

This case study briefly describes the prospective use of an informal Bayesian monitoring procedure. The information has been kindly provided by Dr Mahesh Parmar, MRC Clinical Trials Unit (Cancer Division), Cambridge. Primary references are Saunders *et al.*[390] and Parmar *et al.*[348]

The technology. In 1986 a new radiotherapy technique called CHART was introduced. Its concept was to give radiotherapy continuously (no weekend breaks), in many small fractions (three a day) and accelerated (the course completed in 12 days). It should be clear that there are considerable logistical problems in efficiently delivering CHART. Promising non-randomised and pilot studies led the UK Medical Research Council to instigate two large randomised trials for head-and-neck and lung cancer. Only the lung cancer trial[390] is considered here.

Objectives of study. To estimate the change in survival in lung cancer patients when given CHART compared with conventional radiotherapy, in particular the probability that it provides a clinically important difference in survival that compensates for any additional toxicity and problems of delivering the treatment.

Design of study. A 60:40 randomisation in favour of CHART was selected, with 90% power (with a two-sided 5% significance level) to detect a 10% improvement in 2 year survival over the 15% expected under conventional treatment. This would require around 600 patients with 470 expected events, with accrual expected to last 4 years and a further 1 year of follow-up. The alternative hypothesis of 10% is the mean of the subjective prior distribution expressed by 11 clinicians (see below).

The trial began recruitment in January 1990, with planned annual meetings of the Data Monitoring Committee (DMC) to review efficacy and toxicity data. No formal stopping procedure was specified in the protocol.

Evidence from study. The data reported at each meeting of the DMC is shown in *Table 7*.

Recruitment stopped in early 1995 after 563 patients had entered the trial. It is clear that the extremely beneficial early results were not retained as the data accumulated, although a clinically important and statistically significant difference was eventually found.

The statistical model. We assume a proportional hazards model where the hazard ratio is defined as the hazard under standard treatment to the hazard under CHART, and hence hazard ratios greater than 1 indicate the superiority of CHART. If randomisation were balanced between the arms of the trial, the likelihood for the log(hazard ratio) θ may be assumed to be normally distributed with mean $4L/m$ and variance $4/m$, where L is the log-rank statistic (observed number of deaths in the CHART group minus the expected number under the null hypothesis) and m is the total number of deaths.[421] Since the randomisation ratio is 60:40 in favour of CHART, this variance must be changed to $3.84/m$. The log(hazard ratio) θ can be transformed, under the proportional hazard assumption, to an improvement in 2 year survival rate δ through the relationship $e^{\theta} = \log p_0/\log(\delta + p_0)$, where p_0 is the 2 year survival rate under conventional treatment, assumed to be 15%. This relationship allows us to obtain estimates and intervals for δ from normal posterior distributions calculated on the θ scale.

TABLE 7 *Summary data reported at each meeting of the CHART lung cancer trial DMC*

Date	No. of patients	No. of deaths	Observed hazard ratio	2 year % survival improvement (95% CI)	Two-sided *P* value
1992	256	78	1.82	20 (5, 36)	0.007
1993	346	175	1.69	18 (7, 28)	0.0004
1994	460	275	1.43	12 (4, 20)	0.003
1995	563	379	1.33	9 (3, 16)	0.004
1996	563	444	1.32	9 (3, 15)	0.003

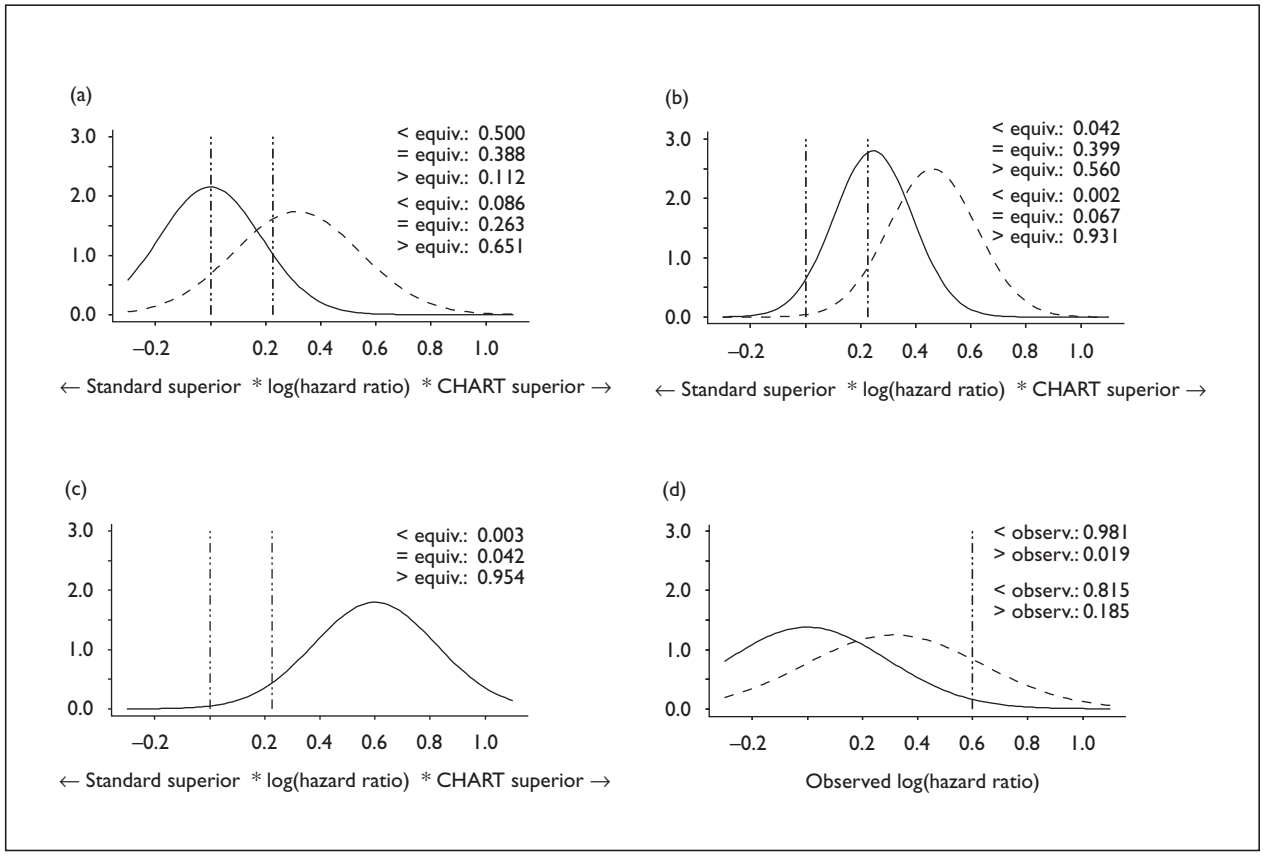


FIGURE 7 *The situation at the first interim analysis of the CHART lung cancer trial in 1992 (hazard ratio = 1.82 based on 78 events). The sceptical (solid) and clinical (dashed) (a) prior and (b) posterior distributions are shown, with probabilities of lying below, within and above the range of equivalence (corresponding to between a 0 and 7% improvement in the 5 year survival rate. (c) Likelihood (*m = 78*). (d) Predictive distribution*

Prospective analysis? The priors were elicited before the start of the trial, and the Bayesian results presented to the DMC at each of their meetings.

Loss function. No formal loss function was elicited, but a pretrial survey of 11 clinicians participating in the trial revealed that, on average, they would be willing to use CHART routinely if it conferred a 13.5% improvement in 2 year survival (from a baseline of 15%).[348]

Prior distribution. Although the participating clinicians were enthusiastic about CHART, there was considerable scepticism expressed by oncologists who declined to participate in the trial. Three forms of prior distribution are considered:

1. A **reference** prior comprising a locally uniform distribution on the log(hazard ratio) scale,
2. A **clinical** prior distribution was elicited from 11 clinicians before the trial started using the methods of Spiegelhalter *et al.*[421] (see page 17), who report the parallel elicitation exercise for the head-and-neck CHART trial. The prior distribution, when averaged over the clinicians, expressed a median anticipated 2 year survival benefit of 10%, and a 10% chance that CHART would offer no survival benefit at all. When transformed to a log(hazard ratio) scale, this subjective prior distribution had a mean of 0.314 (hazard ratio of 1.37) and a precision equivalent to a trial with 60:40 allocation in which 73 deaths had occurred (50 under CHART, 23 under standard treatment).
3. A **sceptical** prior was derived using the ideas in chapter 3: the prior mean is 0 and the precision is such that the prior probability that the true benefit exceeds the alternative hypothesis is low (5% in this case). This gives a prior equivalent to trial with 60:40 allocation with 112 events and an observed log(hazard ratio) of 0.

Both the sceptical and clinical prior distributions are displayed in *Figure 7*. The probabilities of no improvement in 2 year survival, and of an improvement greater than 7% (corresponding to a log(hazard ratio) of 0.225) are shown. The value of 7% as a clinically important difference has been subjectively chosen to be half the 13.5% originally

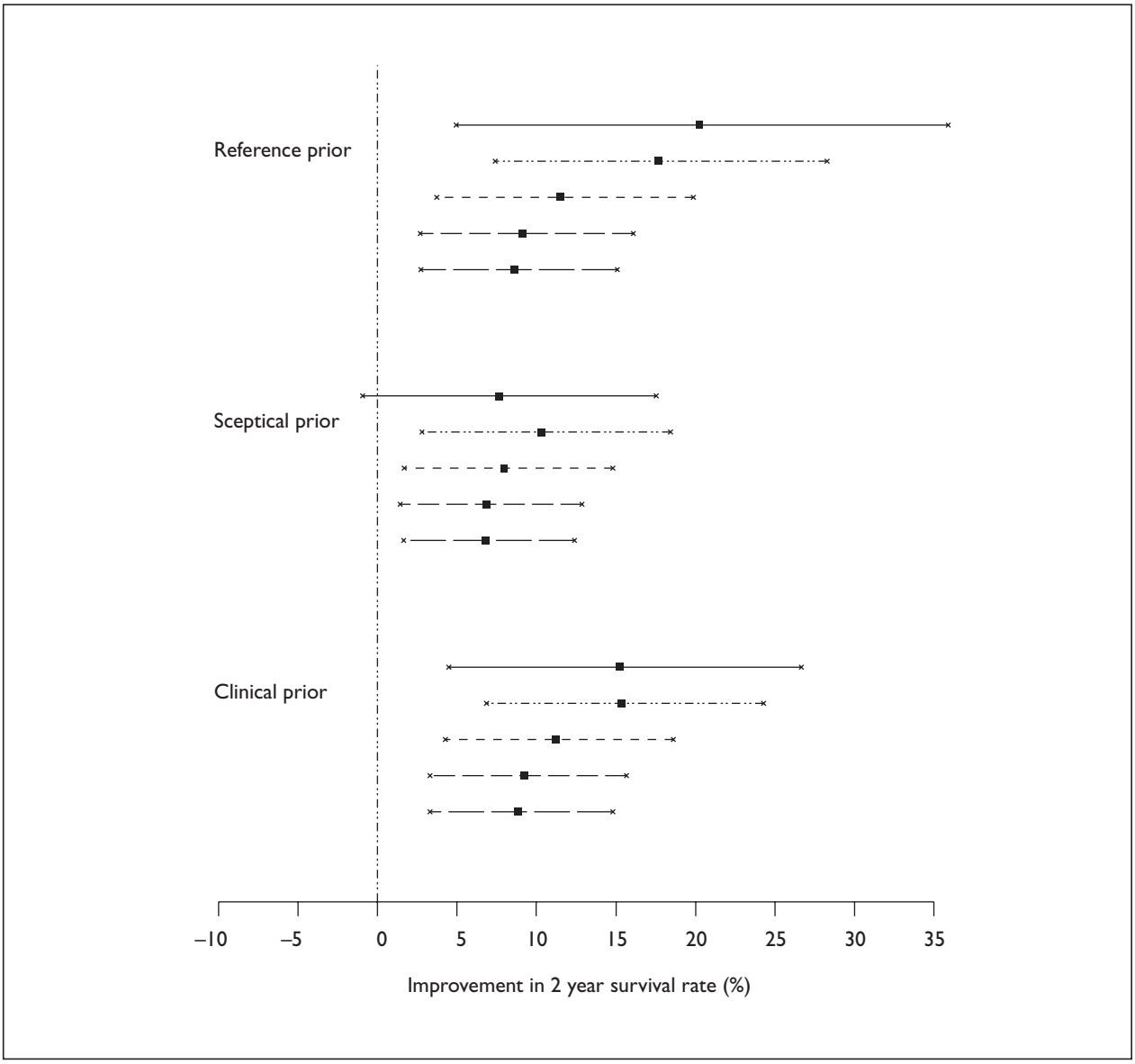


FIGURE 8 *Estimates and 95% intervals for the improvement in 2 year survival rate attributable to CHART treatment, under the reference, sceptical and clinical priors (———— , 1992; ·····, 1993; – – – – , 1994; — — — , 1995; —— —— , 1996)*

demanded by the clinicians, in view of the unexpected lack of toxicity of the new treatment.

Computations. All analysis has been carried out using S-plus BART functions previously used in Spiegelhalter *et al.*[421]

Reporting. The DMC were presented with survival curves, reference and sceptical posterior distributions and tail areas.

Sensitivity analysis. The three priors provided the sensitivity analysis.

Figure 7 shows the sceptical and clinical prior distributions, the likelihood for the results available in 1992 (equivalent to the reference posterior), the corresponding sceptical and clinical posterior distributions, and the predictive distributions for the observed log(hazard ratios) under the two priors.

Figure 8 shows the results progressing over the 5 years of the study. Under the reference prior there is substantial reduction in the estimated effect as the extreme early results are attenuated. The sceptical prior is remarkably stable, and its initial estimate in 1992 is essentially unchanged as the trial progresses.

The detailed results under the sceptical prior are shown in *Table 8*, showing the stable results over time.

TABLE 8 *Estimates presented to the CHART DMC in successive years, obtained under a sceptical prior distribution*

Year	No. of deaths	Estimates under the sceptical prior			
		Hazard ratio	2 year % survival improvement (95% CI)	*P*(improvement > 0%)	*P*(improvement > 7%)
1992	78	1.27	7 (–1,17)	0.044	0.46
1993	175	1.37	10 (3,18)	0.003	0.21
1994	275	1.28	8 (2,15)	0.006	0.40
1995	379	1.25	7 (1,13)	0.006	0.52
1996	444	1.24	7 (2,12)	0.004	0.54

Chapter 10

Case study 2: meta-analysis of magnesium sulphate following acute myocardial infarction

The technology. This example has been considered at length in the medical and statistical literature, as it features an apparent contradiction between a meta-analysis and a 'mega-trial'. An abbreviated history follows. Epidemiology, animal models and biochemical studies have suggested intravenous magnesium sulphate may have a protective effect in patients with AMI, particularly through preventing serious arrhythmias.[443] A series of small randomised trials culminated in a meta-analysis in 1991[443] which showed a highly significant ($P < 0.001$) 55% reduction in odds of death. The authors concluded that "further large scale trials to confirm (or refute) these findings are desirable", and in 1992 the Second Leicester Intravenous Magnesium Intervention Trial (LIMIT-2)[481] published results showing a 24% reduction in mortality in over 2000 patients. An editorial in *Circulation* was entitled 'An effective, safe, simple and inexpensive treatment',[484] but recommended further trials to obtain "a more precise estimate of the mortality benefit". Early results of the massive Fourth International Study of Infarct Survival (ISIS-4) trial pointed, however, to a lack of any benefit, and final publication of this trial on 58,000 patients showed a non-significant adverse mortality effect of magnesium. ISIS-4 found no effect in any subgroups, and concluded that "Overall, there does not now seem to be any good clinical trial evidence for the routine use of magnesium in suspected acute MI".[109]

There have been many responses to this apparent contradiction between meta-analysis and mega-trial, which can be summarised under four broad headings:

- **Essential scepticism about large effects.** In response to the ISIS-4 results, Yusuf, the main author of the optimistic *Circulation* editorial, claimed "since most treatments produce either no effect or at least moderate effects on major outcomes such as mortality, investigators should be sceptical if the results obtained deviate substantially from this expectation ("too good to be true")".[483] This expression of prior scepticism was echoed by Peto and colleagues,[359] who argued that the risk reduction of the initial overview was "implausibly large", and that even when combined with the LIMIT-2 data "still indicated an implausibly large reduction of one-third in mortality". However, Peto reports that the ISIS-4 steering committee was convinced there would be at least some benefit, right up until they were shown the results.

- **Criticism of the meta-analysis.** Egger and Davey-Smith[157,158] claim that the meta-analysis was flawed, as a funnel plot (of numbers of participants against observed treatment effect) suggested smaller negative studies might not have been published, and sensitivity analysis could have prevented the misleading conclusions. Using a different argument, Pogue and Yusuf[367] claim that a frequentist stopping rule applied to the meta-analysis, designed to have high power to detect a moderate effect (15% reduction in mortality), would also have led to the meta-analysis not being significant at the 1% level even taking into account the LIMIT-2 data.

- **Criticism of the mega-trial.** Woods[480] has argued that a mega-trial such as ISIS-4 will tend to bias results towards the null due to protocol violations and inaccurate data. In addition, he claims that the benefit of magnesium is to prevent reperfusion injury, and yet the ISIS-4 protocol expected all patients to be given thrombolytic therapy (which tends to induce reperfusion) before randomisation, and hence magnesium would generally be given too late to provide benefit. He claims the subgroup who did not receive thrombolytic therapy did not provide sufficient power to detect an important difference.

- **Treatment effect depending on baseline risk.** Antman[14] and others have pointed out that in ISIS-4:
  - the control group mortality was 7.2% in contrast to 11.0% observed in the data available at the time of the 1993 *Circulation* editorial
  - patients were randomised at a median of 8 hours after onset
  - 70% received thrombolytics and 94% received antiplatelets.

Antman concluded that "patients who are at low risk of mortality, at least in part due to other potent mortality-reducing therapies such as thrombolytics and aspirin, show little benefit from magnesium". The methodology for examining whether the benefit of treatment may depend on underlying risk has been recently explored in a series of papers,[320,453] emphasising that simple techniques can detect a spurious relationship due to the natural correlation between baseline and change.

Objectives of analysis. To investigate how a Bayesian perspective might have influenced the interpretation of the published evidence on magnesium sulphate in AMI available in 1993. In particular, what degree of 'scepticism' would have been necessary in 1993 not to be convinced by the meta-analysis reported by Yusuf and colleagues,[484] and is there evidence that the treatment effect depends on sample size or baseline risk?

Design of study. Meta-analysis of randomised trials, allowing for sceptical prior distributions and dependence of treatment effect on baseline risk.

Available evidence in study. We use the data shown in *Table 9* as quoted by Yusuf and colleagues,[484] which comprise the data in the 1991 meta-analysis[443] and LIMIT-2.[481]

The log(odds ratio) for the first eight trials, including LIMIT-2, using the standard Peto fixed effect analysis, would be estimated as –0.43 with standard error 0.12. The corresponding likelihood is as if a single trial had been carried out with 261 deaths in total (compared with the actual 286 deaths) and an observed log(odds ratio) of –0.43.

Statistical models. The basic sampling model is assumed to have the following form:

$$r_i^{\mathrm{C}} \sim \mathrm{Binomial}(p_i^{\mathrm{C}}, n_i^{\mathrm{C}})$$
$$r_i^{\mathrm{T}} \sim \mathrm{Binomial}(p_i^{\mathrm{T}}, n_i^{\mathrm{T}})$$
$$\mathrm{logit}(p_i^{\mathrm{C}}) = \mu_i$$
$$\mathrm{logit}(p_i^{\mathrm{T}}) = \mu_i + \delta_i$$

Different models (numbered 1 to 4 below) correspond to different assumptions about the form of the baseline risks μ_i and the treatment effects δ_i:

1. Fixed effect (pooled estimate) model:

   $$\delta_i = d$$

2. Random effects model:

   $$\delta_i \sim \mathrm{Normal}(d, \sigma^2)$$

3. Random effects model allowing effect to depend on (logarithm of) sample size:

   $$\delta_i \sim \mathrm{Normal}(d_i, \sigma^2)$$
   $$d_i = d_\delta + \beta(\log n_i - \overline{\log n_i})$$

4. Random effects model allowing effect to depend on baseline risk (this is a Bayesian analogue of the bivariate meta-analysis model[320,465]):

   $$\delta_i \sim \mathrm{Normal}(d_i, \sigma^2)$$
   $$d_i = d_\delta + \beta(\mu_i - d_\mu)$$
   $$\mu_i \sim \mathrm{Normal}(d_\mu, \sigma_\mu^2)$$
   $$\mu_0 = d_\mu - d_\delta/\beta$$

   Note that by assuming the μ_is come from a normal distribution, we are in fact assuming a bivariate normal distribution for the baseline risk and treatment effect.

   The value μ_0 is the baseline risk (on a logit scale) at which the treatment has no effect.

Prior distribution. We consider that a reasonable degree of scepticism is to think it unlikely (only 10% chance) that magnesium would change the odds on mortality by more than 25%. This can be translated into a normal prior distribution, centred on 0 and with precision equivalent to a 'trial' with 65 deaths in each group:

$$D \sim \mathrm{Normal}(0, 0.175^2)$$

(see chapter 3 for further discussion of such sceptical priors). An alternative, very sceptical prior was also examined, equivalent to having already observed 450 deaths in each group.

All other parameters are given proper minimally informative priors: Normal$(0, 10^6)$ for location parameters and Gamma$(0.001, 0.001)$ for precisions.

Loss function. No explicit loss function, but a 10% reduction in odds of death has been selected as a 'clinically important difference'.

Computations. Fixed effect analysis using the BART S-plus functions (see appendix 2), random effects analysis in BUGS).

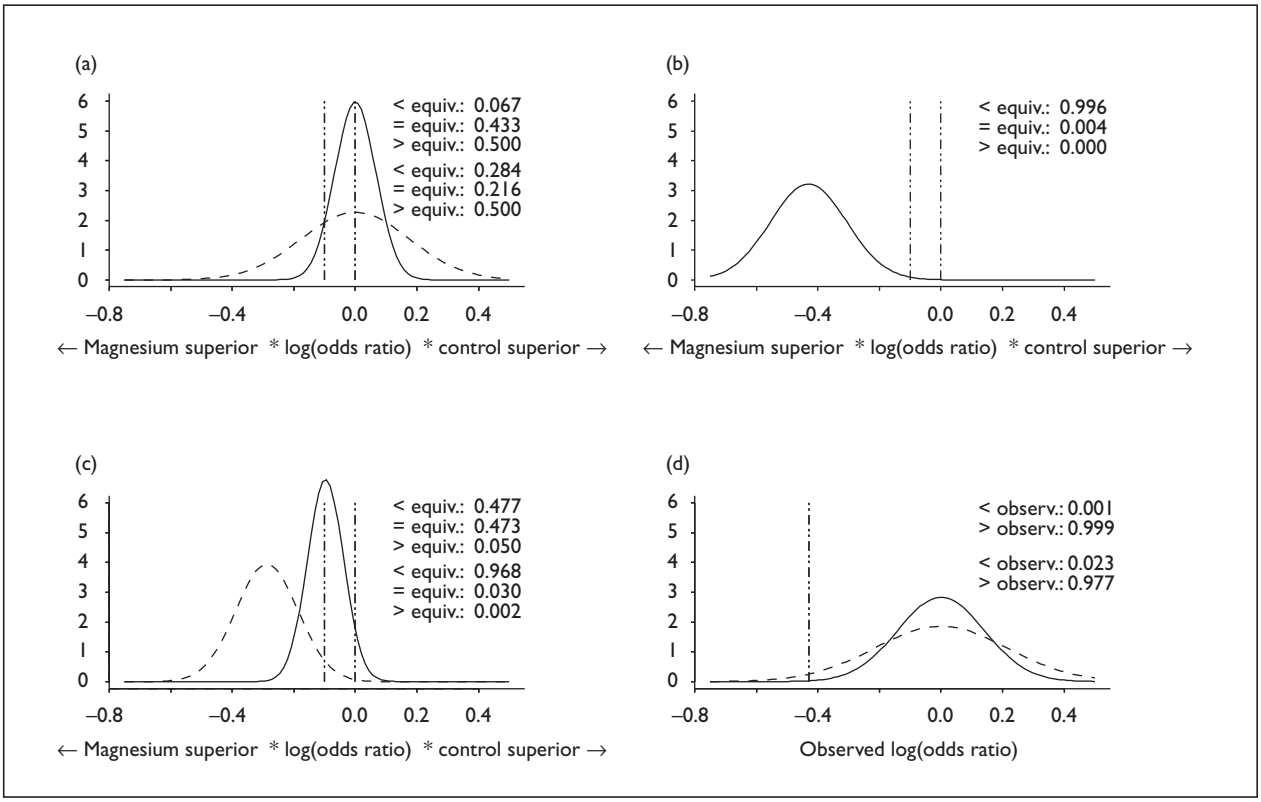


FIGURE 9 *(a) Prior, (b) likelihood (m = 261, x = –0.43) and (c) posterior distributions for two sceptical prior distributions (450 and 65 deaths): the wider is a 'reasonable' expression of scepticism (equivalent to a 'trial' with 65 deaths in each group), while the narrower prior is the scepticism necessary not to have found the meta-analysis 'significant' (equivalent to a 'trial' with 450 deaths in each group).*

TABLE 10 *Posterior probabilities of absolute and clinical superiority of magnesium, given two levels of sceptical prior*

	Magnesium superior $p(\delta < 0)$	**Magnesium clinically superior** $p(\delta < -0.1)$
Very sceptical (450 in each group)	0.95	0.48
Reasonably sceptical (65 in each group)	0.998	0.97

Results and sensitivity analysis.

Fixed effect analysis. Figure 9 shows the prior, likelihood and posterior for two sceptical priors: a 'reasonable' one and one designed to produce a posterior distribution with 5% chance that there is no benefit from magnesium.

The prior necessary not to have found the meta-analysis 'significant', even at a one-sided 5% probability, is clearly a very extreme form of scepticism. *Table 10* shows that a reasonably sceptical prior even finds the meta-analysis quite convincing concerning a clinically worthwhile improvement, in that there is 97% chance that the treatment benefit is at least 10%.

We therefore can reject Yusuf and Flather's claim that a sceptical approach applied to their analysis would have led to caution.

Random effect analysis. The random effects analysis leads to a different conclusion. *Figure 10* shows the 95% posterior credible intervals for the mortality odds ratio associated with magnesium, for both fixed and random effect analysis, and a 'flat' reference prior and the reasonably sceptical prior. The random effects analysis combined with a 'flat' reference prior, and the fixed effect analysis with a flat or sceptical prior (as also shown in *Figure 9*), all lead to highly 'significant' results. However the random effects analysis with a sceptical prior leads to a 95% interval that includes one, and hence the cautious result sought by Yusuf. Inclusion of the ISIS-4 results has a strong impact on the fixed effect analysis, but little influence on the random effect (see discussion below).

Figure 11 shows the consequences of a fully Bayesian sceptical random effects analysis on the estimates given to the **individual** trials. The LIMIT-2 results

are hardly changed, whereas the smaller studies are pulled towards a cautious conclusion.

Does effect depend on sample size? Egger and Davey-Smith[157,158] have claimed that one problem with the initial meta-analysis[443] is publication bias against smaller negative trials. Using the data available to Yusuf *et al.* in 1993, and a minimally informative reference prior, we fit the following model to see whether there is evidence that the treatment effect does depend on sample size as suggested in their funnel plot:

$$\delta_i \sim \text{Normal}(d_i, \sigma^2)$$
$$d_i = d_\delta + \beta(\log n_i - \overline{\log n_i})$$

A positive β corresponds to smaller trials having smaller expected odds ratios, corresponding to a larger treatment effect of magnesium. *Table 11* shows the estimated βs are positive but with very wide intervals, so that there is therefore only weak evidence that smaller studies had more extreme results.

Relationship to underlying risk. Figure 12 shows the apparent relationship between the observed treatment effect and underlying risk. Fitting a regression line through these points, using the appropriate bivariate model described earlier, provides the results shown in *Table 12*, with and without the ISIS-4 data, where p_0 is the risk in the control group at which the treatment has no effect.

The relationship to underlying risk is suggested before inclusion of ISIS-4, but with a very wide interval. Nevertheless, the model would have predicted that the treatment would not be effective with an underlying risk below 9.1%. Inclusion of the ISIS-4 results, whose underlying risk was 7.2%, strongly confirmed this relationship.

Discussion. We conclude that one would need to have been unreasonably extremely sceptical not to have found the 1993 meta-analysis convincing, **if one had carried out a standard Peto fixed-effect analysis**. Reasonable scepticism and a random effects meta-analysis would have led to appropriate caution. There was limited evidence available in 1993 that treatment effect was related to sample size or underlying risk, but both have been confirmed by ISIS-4.

It is important to recognise the limitations of such a statistical analysis. There is obvious heterogeneity between the studies, and it is vital to investigate the possible reasons for this using substantive knowledge, through inclusion of covariates and so on. Many sorts of sensitivity analysis are necessary (and

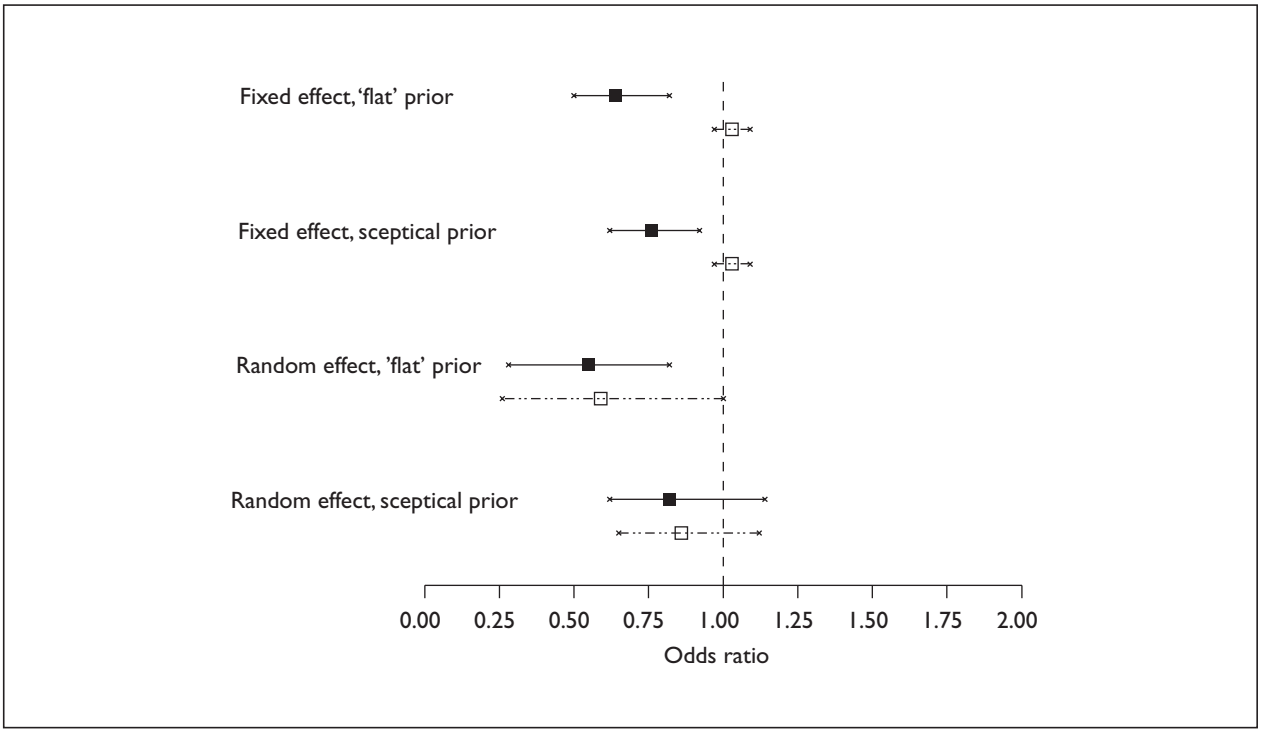


FIGURE 10 *Ninety-five per cent posterior credible intervals for the mortality odds ratio associated with magnesium, for both fixed and random effect analysis, and a 'flat' reference prior and the reasonably sceptical prior (10% chance of at least a 25% change in mortality odds) (solid line, Yusuf* et al. *(1993); broken line, Yusuf* et al. *plus ISIS-4)*

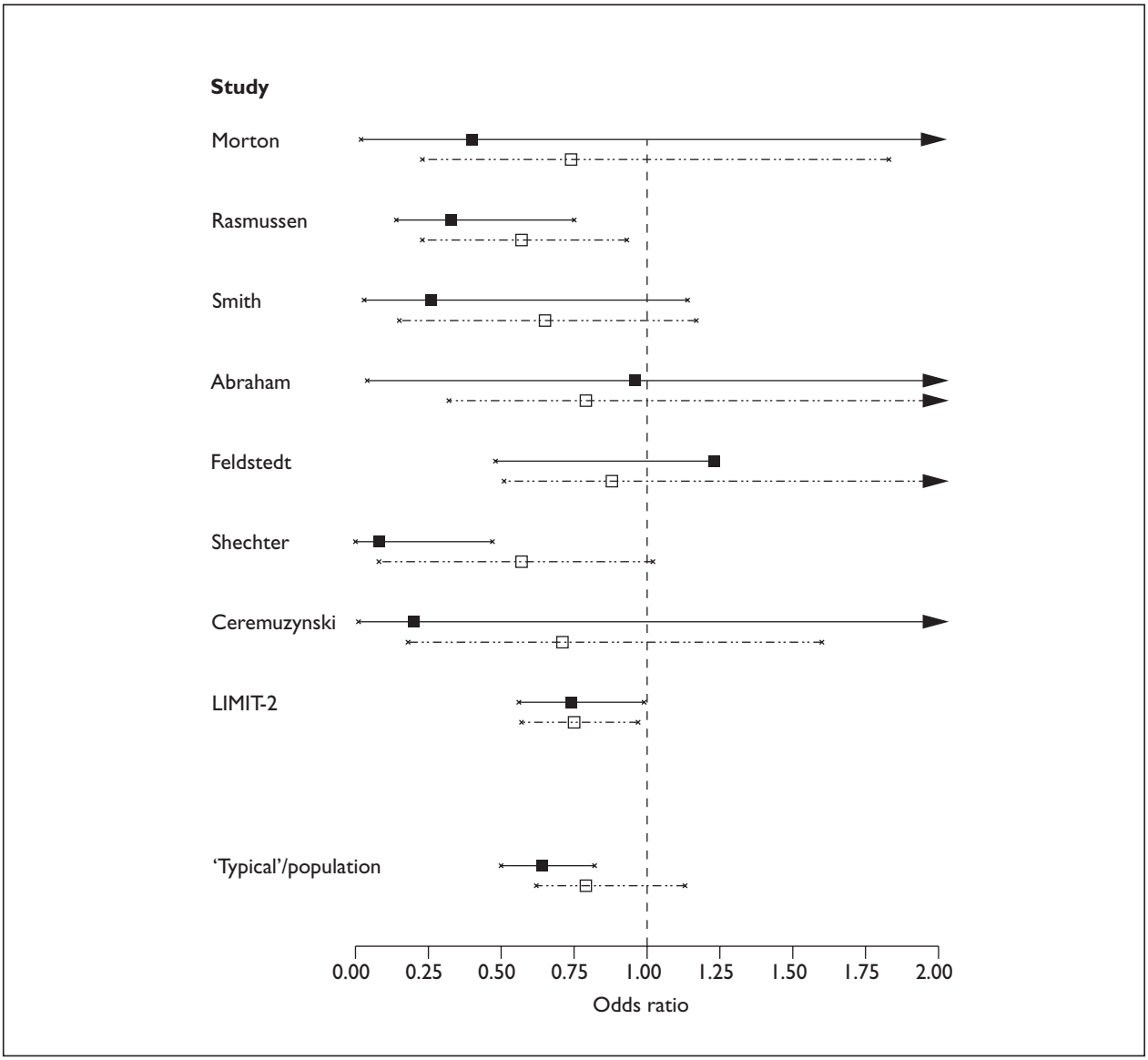


FIGURE 11 *Ninety-five per cent intervals for individual studies, from a random effects analysis with a reasonably sceptical prior (solid line, fixed effects; broken line, random effects with sceptical prior)*

TABLE 11 *Estimate and 95% interval for the influence of increased sample size on the effect of magnesium treatment*

Model	β	**95% interval**
Fixed effect	0.50	(–1.45, 2.42)
Random effect	0.29	(–0.20, 0.68)

feasible). The random effects methodology can be questioned when used on studies of very different sizes, in that the very large studies may be unreasonably downweighted. It suggests the question 'What is a "study"?'. We could always break down a large study into smaller ones to add weight: for example ISIS-4 included 31 countries and 1086 hospitals, and it would be very interesting to investigate the heterogeneity between these centres in a structured way.

TABLE 12 *The estimated influence of underlying risk on magnesium treatment effect: a negative β corresponds to the treatment effect becoming smaller as the underlying risk declines, with the treatment effect becoming zero when the underlying risk is* p_0

Model for baseline risks	β	**95% interval**	p_0 **(%)**
Without ISIS-4	–1.4	(–24, 14)	9.1 (0.5–54)
With ISIS-4	–1.2	(–2.2, –0.2)	7.2 (1.9–8.4)

The Bayesian analysis, while not necessarily providing qualitatively different conclusions to a traditional analysis, does allow subjective judgements to be formally incorporated and the sensitivity of the conclusions to those beliefs explored within a coherent framework.

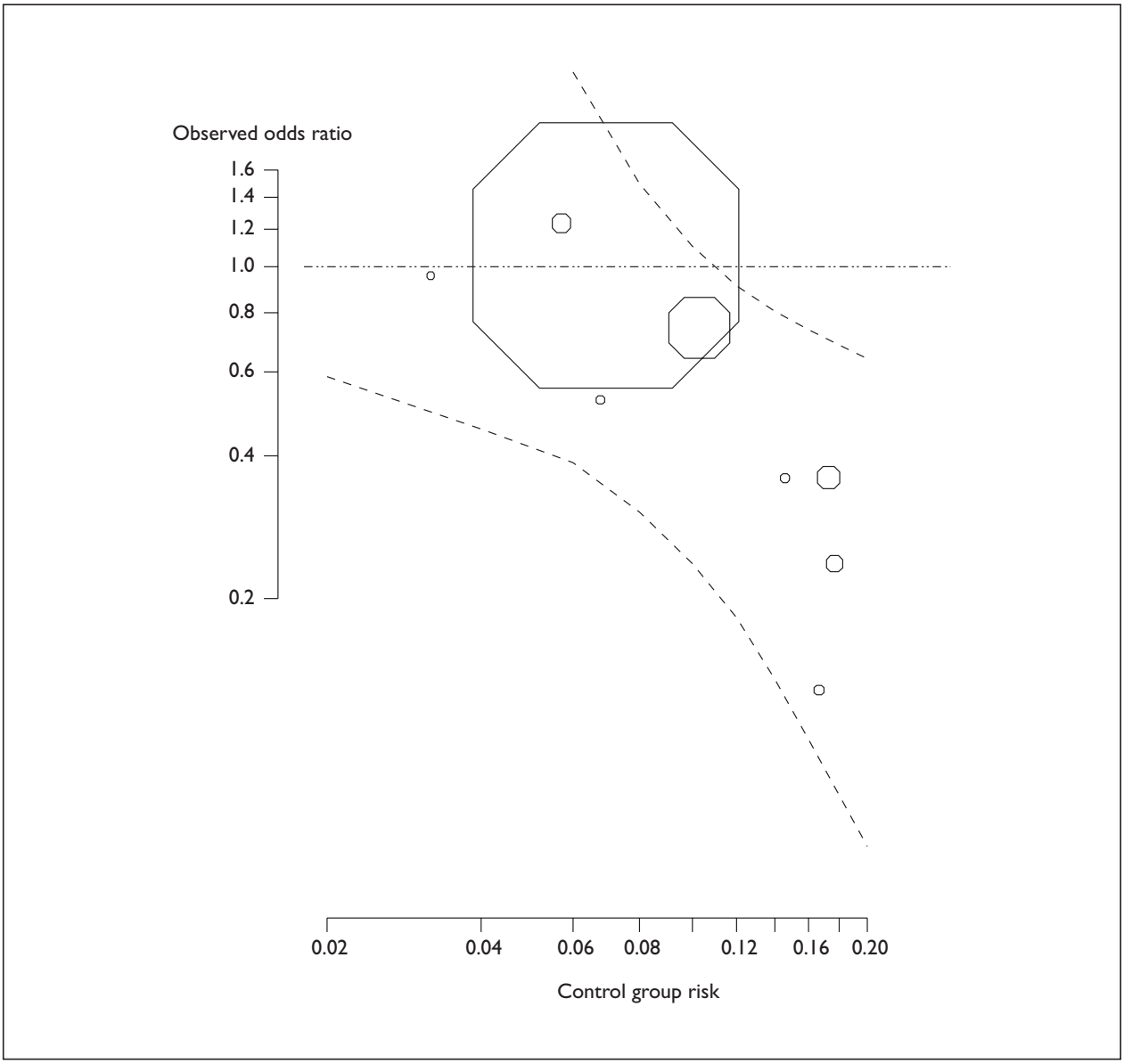


FIGURE 12 *Plot of observed mortality odds ratio associated with magnesium, for seven trials contributing to the Teo* et al. *meta-analysis, LIMIT-2 and ISIS-4. The size of each point is proportional to the precision of estimation, so that corresponding to ISIS-4 is very large. A bivariate distribution of treatment effect and baseline risk has been fitted, and a predictive distribution around the regression line plotted*

Chapter 11

Case study 3: confidence profiling revisited

Here we revisit four short examples provided by Eddy *et al.* in their text on the confidence profile method[150] (see chapter 6 for a discussion of their work and possible reasons for its lack of impact).

The confidence profile method made extensive use of graphical models for communication of the essential features of a model, although their software, FAST*PRO, was menu-driven. In contrast, more recent developments in Bayesian graphical modelling have led to software, for example WinBUGS (see appendix 2), in which the graph explicitly drives the analysis through generating the code for performing the required simulations (see chapter 2 for background on simulation-based Bayesian analysis). The re-analysis of some of their examples thus serves two purposes: to emphasise the versatility and power of their approach, and to show how current (freely available) software can make its implementation reasonably straightforward.

The four specific applications have been selected to illustrate various facets of Bayesian graphical modelling applied to evidence synthesis in health technology assessment, and each example is structured according to the BayesWatch criteria. However, they are clearly rather dated and hence should not be taken as having any substantive value, and we have simply reproduced the original description and have not attempted to carry out a full analysis to appropriate quality standards.

Analysis of surveillance of colorectal cancer patients: a modelling exercise based entirely on judgements

Reference. This example forms chapter 29 of Eddy *et al.*[150]

The technology. Surveillance of colorectal cancer patients in order to reduce the risk of liver metastases. No direct evidence for the effectiveness of this intervention is available.

Objectives of analysis. To estimate the reduction in mortality rate from liver metastasis due to introduction of surveillance, denoted δ_{surv}.

Statistical model. We assume that a death is prevented if:

1. a case has a solitary liver metastasis at the time of examination (with probability p_{exist})
2. the metastasis is detected at examination (with probability $p_{detect.if.exist}$)
3. the detected metastasis is treatable (with probability $p_{trt.if.detect}$)
4. the treatment leads to the patient surviving 5 years, who would not otherwise have survived without surveillance (survival probability has increased $\delta_{surv.if.trt}$)

The overall improvement in survival rate is given by the logical product:

$$\delta_{surv} = \delta_{surv.if.trt} \times p_{trt.if.detect} \times p_{detect.if.exist} \times p_{exist}$$

The graph in *Figure 13* shows this logical dependence, using the DoodleBUGS graph drawing facility in WinBUGS.

Prior distributions. The subjective prior distributions shown in *Table 13* are those provided by Eddy and colleagues.

Computations. Since no data are involved a forwards Monte Carlo simulation can be carried out without a burn-in stage and without concerns with convergence. A total of 100,000 iterations were carried out, taking 2 seconds on a 400 MHz personal computer.

Results. Summary statistics for the simulated posterior distributions are shown in *Table 14*.

The posterior distributions are the same as the prior distributions (up to simulation error) since no data have been observed. We estimate 0.022% increase in survival (95% interval –0.010 to 0.092%), compared with Eddy and colleagues' estimate of 0.025% (95% interval –0.031 to 0.075%). Their estimate is based on assuming a normal posterior distribution for δ_{surv}, and the

skewness of the posterior distribution displayed in *Figure 14* clearly shows this is inappropriate.

Sensitivity analysis. The whole case study can be thought of as a sensitivity analysis to the assumption of known parameter values – if the best guesses had been used the conclusion would have been an increase in survival, with no allowance for uncertainty.

Comments. This analysis illustrates the propagation of uncertainty through a logical model, akin to placing probability distributions on the inputs to a spreadsheet. The resulting skewness shows the dangers of making normal approximations to posterior distributions of logically transformed quantities.

Analysis of HIP trial of breast cancer screening: adjusting a trial's result for uncertain internal biases

Reference. This example forms chapter 19 of Eddy *et al.*[150]

The technology. Breast cancer screening offered to women aged under 50 years.

Objectives of analysis. To estimate the reduction in mortality in breast cancer mortality associated with **accepting** screening, denoted $e_d = \theta_c - \theta_t$, where θ_c is the true mortality rate in those not screened, and θ_t is the rate in those actually screened. We also require a 95% interval and the chance that the reduction is greater than 2/1000.

Design of study. Randomised controlled trial.

Available evidence in study. The HIP published in 1988 the results shown in *Table 15*. Note that the data reflect the mortality rate of those offered screening, whereas we wish to make statements about those actually taking up screening.

Statistical model. Let θ_{off} and θ_c be the underlying mortality rate of those offered and not offered screening respectively, so that $r_t \sim \text{Binomial}(n_t, \theta_{off})$, and $r_c \sim \text{Binomial}(n_c, \theta_c)$. Eddy and colleagues consider four increasingly complex models:

1. 'Intention to treat': act as if those screened have the same mortality rate as those offered, that is $\theta_{off} = \theta_t$.

2. Adjustment for dilution: assume the proportion d who do not accept the screening has the same mortality rate as those not offered screening, that is $\theta_{off} = (1 - d)\theta_t + d\theta_c$. In this example d is initially assumed to be 45%.

3. Adjustment for selection bias: suppose it were hypothesised that those women who would reject screening were at lower average risk than

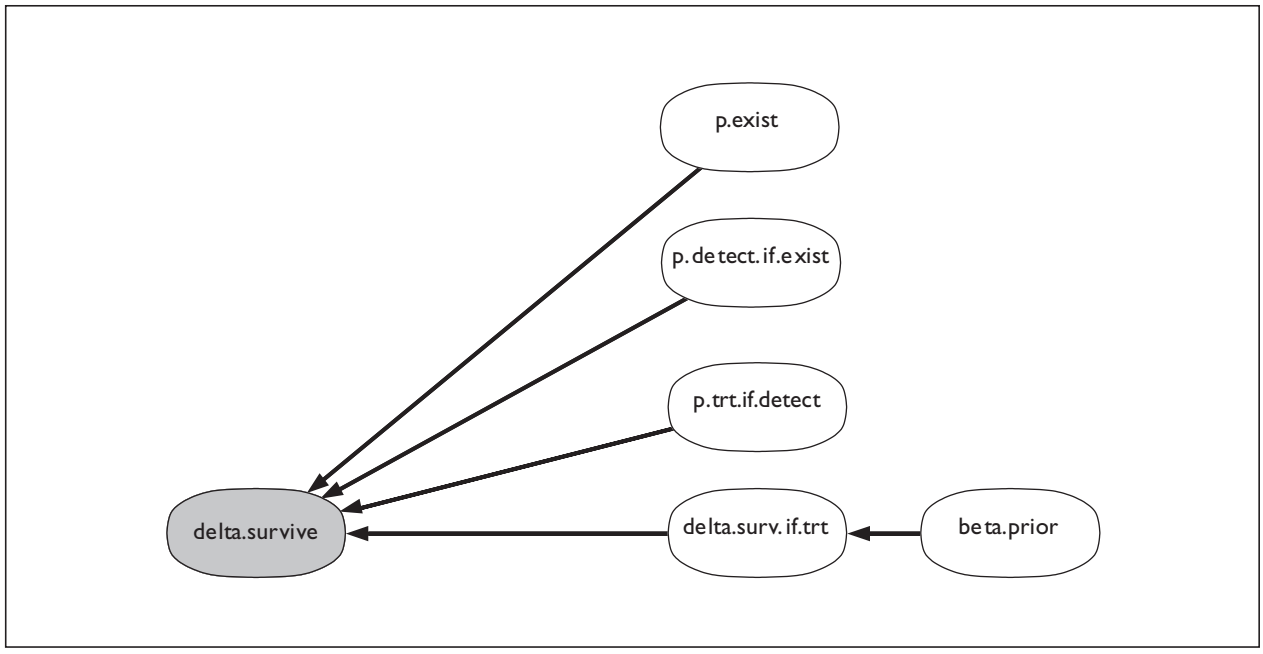


FIGURE 13 *A 'doodle' from WinBUGS illustrating the structural dependencies of the model: this graphical representation directly generates the code which carries out the simulations*

TABLE 13 *Prior distributions provided by Eddy and colleagues for cancer surveillance example*

Parameter	Best guess (%)	95% confidence range (%)	Distribution
p_{exist}	1	0.3 to 2	Beta(4.83, 488)
$p_{detect.if.exist}$	75	40 to 95	Beta(5.5, 1.83)
$p_{trt.if.detect}$	30	5 to 65	Beta(2.5, 5.83)
$\delta_{surv.if.trt}$	10	–5 to 25	(2 × Beta(96.25, 78.75)*2) – 1

those who would accept. Specifically, let p_n and p_t be the relative risks associated with the characteristics that would lead to screening being rejected, in the groups not offered and offered screening respectively. Then it can be shown (Eddy and colleagues,[150] Chapter 15) that

$$\theta_{off} = (1-d)\frac{\theta_t}{1-d+dp_t} + d\frac{\theta_c}{d+(1-d)/p_t}$$

p_n and p_t are initially assumed to be 0.9.

4. Uncertainty on the biases: informative prior distributions are now placed on the bias parameters (see below) to represent the more realistic assumption that the biases are only imprecisely known.

Figure 15 shows the model for the first analysis.

Prior distribution. The prior distributions shown in *Table 16* and provided by Eddy and colleagues are 'non-informative' (Jeffreys priors) for the primary parameters, and specify a standard deviation of 0.1 for the distributions of d, p_n and p_c. The distributions for p_n and p_c are log-normal, that is their logarithms are assumed to have a normal distribution with the appropriate mean and standard deviation.

TABLE 14 *Results for cancer surveillance example*

Parameter	Posterior mean (%)	95% credible interval
p_{exist}	1.0	(0.3, 2.0)
$p_{detect.if.exist}$	75.0	(41.0, 96.7)
$p_{trt.if.detect}$	30.0	(6.1, 62.6)
$\delta_{surv.if.trt}$	10.0	(– 4.9, 24.7)
δ_{surv}	0.022	(–0.010, 0.092)

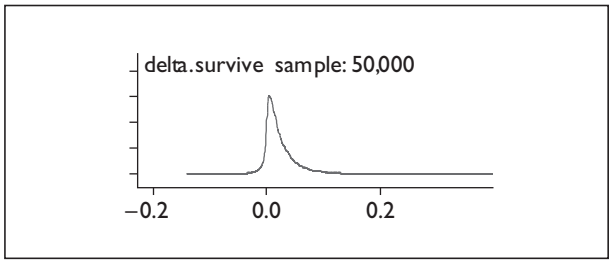


FIGURE 14 *Posterior distribution of the change in survival attributable to screening δ_{surv}, showing strongly skewed distribution*

TABLE 15 *Results of HIP screening trial published in 1988*

	Not offered screening	Offered screening
Breast cancer deaths	r_c = 65	r_t = 49
Total	n_c = 12,000	n_t = 12,000

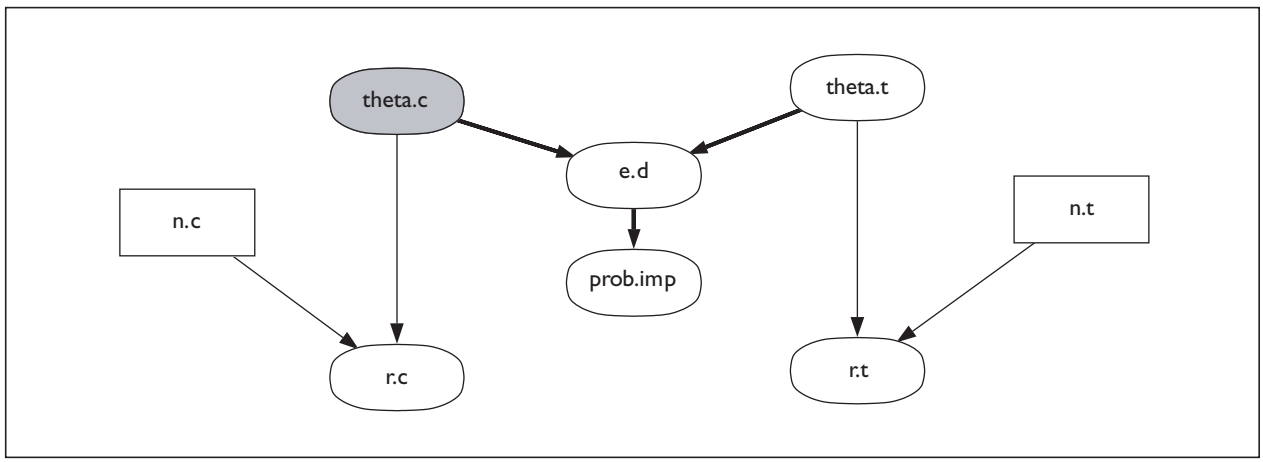


FIGURE 15 *A 'doodle' illustrating the structural dependencies of the model; thick lines represent logical dependencies, and thin lines represent stochastic dependencies*

Computations. A total of 100,000 iterations were carried out, taking 2 seconds on a 400 MHz personal computer.

Results. The posterior distribution of e_d had the properties shown in *Table 17*.

These agree closely with the results of Eddy *et al.* The crucial finding is that the results are very sensitive to the introduction of an allowance for bias (moving from model 1 to 2), but robust to specification of its precise nature.

Comments. As Eddy *et al.* point out, many additional issues might be addressed in this framework, including: varying degrees of dilution in which different proportions of women receive different numbers of examinations, the possibility of contamination in the control group, loss to follow-up, errors in measurement of outcome, and the possibility that the technology might have improved over time. It would be interesting to contrast this analysis with recent investigations of this still-controversial topic.

Analysis of screening for maple syrup urine disease (MSUD): modelling using evidence from multiple studies

Reference. This example forms chapter 27 of Eddy *et al.*[150]

TABLE 16 *Prior distributions provided by Eddy and colleagues for the cancer-screening example*

Parameter	Mean	Standard deviation	Distribution
θ_t			Beta(0.5, 0.5)
θ_c			Beta(0.5, 0.5)
d	0.45	0.1	Beta(10.69, 13.06)
log p_n	–0.105	0.0953	N(–0.105, 1/111.1)
log p_t	–0.105	0.0953	N(–0.105, 1/111.1)

TABLE 17 *Results for the cancer-screening example*

Model	Posterior mean	95% credible interval	$P(e_d < -0.002)$
1	–0.0013	(–0.0031, 0.0004)	0.23
2	–0.0027	(–0.0061, 0.0005)	0.66
3	–0.0025	(–0.0058, 0.0005)	0.63
4	–0.0028	(–0.0061, 0.0006)	0.66

TABLE 18 *Data used in the MSUD example*

Factor	Notation	Outcomes	Observations
Probability of MSUD	r	7	724,262
Probability of early detection with screening	ϕ_s	253	276
Probability of early detection without screening	ϕ_n	8	18
Probability of retardation with early detection	θ_{em}	2	10
Probability of retardation without early detection	θ_{lm}	10	10

TABLE 19 *Model and notation for the MSUD example*

Factor	Notation	Derivation
Probability of retardation for a case of MSUD who is screened	θ_{sm}	$\phi_s\theta_{em} + (1-\phi_s)\theta_{lm}$
Probability of retardation for a case of MSUD who is not screened	θ_{nm}	$\phi_n\theta_{em} + (1-\phi_n)\theta_{lm}$
Expected retardations per 100,000 newborns who are screened	$100{,}000\theta_s$	$100{,}000\theta_{sm}r$
Expected retardations per 100,000 newborns who are not screened	$100{,}000\theta_n$	$100{,}000\theta_{nm}r$
Change in retardations due to screening 100,000 newborns	e_d	$\theta_s - \theta_n$

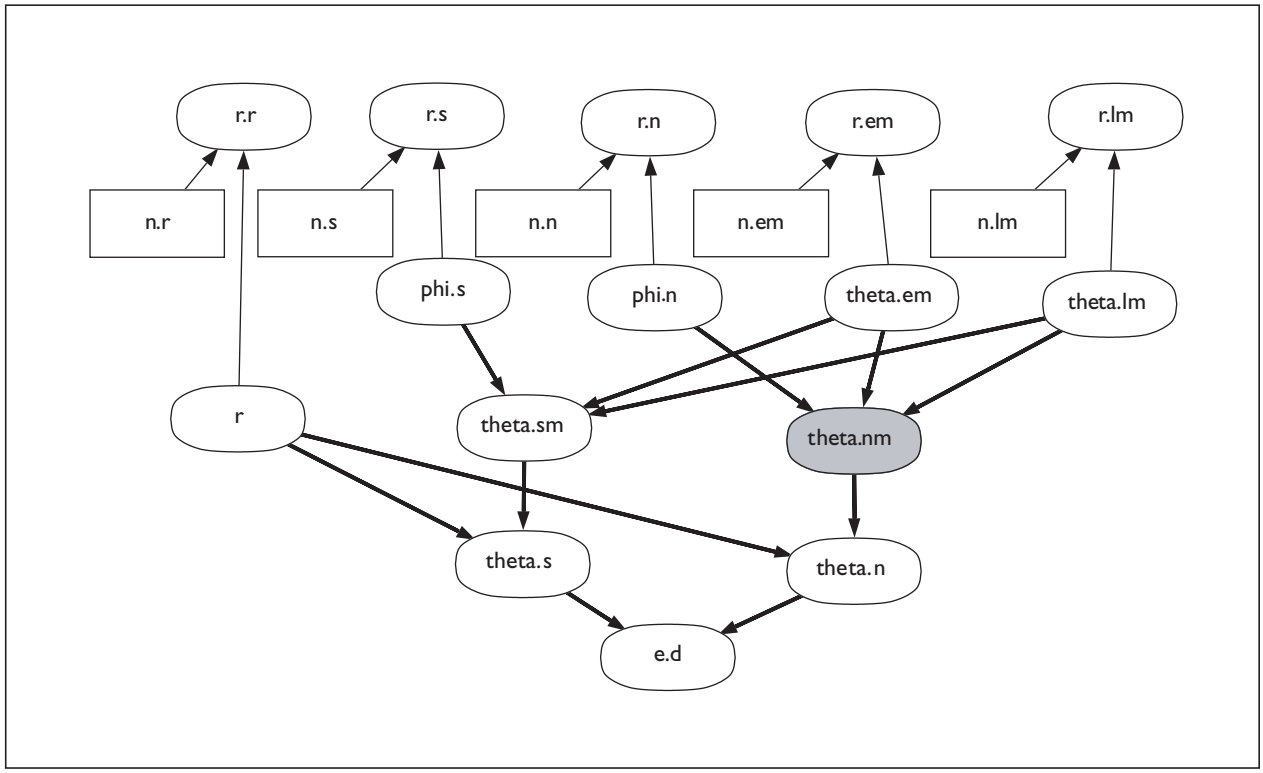


FIGURE 16 *Doodle for the MSUD example*

TABLE 20 *Results for the MSUD example*

Parameter	Notation	Posterior mean	95% credible interval
Expected retardations per 100,000 newborns who are **not** screened	θ_n	0.65	(0.25, 1.27)
Change in expected retardations due to screening 100,000 newborns	e_d	–0.35	(–0.77, –0.11)

The technology. Neonatal screening for MSUD, an inborn error in amino acid metabolism, for which early detection should lead to reduced rates of retardation.

Objectives of analysis. To estimate the probability of retardation without screening, and the change in retardation rate associated with screening. The latter is denoted $e_d = \theta_s - \theta_n$, where θ_n is the retardation rate in those not screened, and θ_s is the rate in those screened.

Design of study. Modelling exercise using results from multiple epidemiological cohort studies.

Available evidence. There was no direct evidence on the change in retardation rate in screened and unscreened populations. The data shown in *Table 18* was used (references provided by Eddy and colleagues).

Statistical model. The data described in *Table 18* are all assumed to arise from binomial distributions with the appropriate parameters. The functional relationships shown in *Table 19* then exist. The graphical model is shown in *Figure 16*.

Prior distribution. The prior distributions for all the binomial parameters provided by Eddy and colleagues are 'non-informative' (Jeffreys priors – Beta(0.5, 0.5)).

Computations. A total of 100,000 iterations were carried out, taking 3 seconds on a 400 MHz personal computer.

Results. The posterior distribution of e_d had the properties shown in *Table 20*.

Eddy and colleagues display a normal approximation to the posterior distribution for e_d, with an

estimate of –0.35 (95% interval –0.69 to –0.19). Our wider interval accurately reflects the skewed posterior distribution.

Comments. This example illustrates the synthesis of evidence from multiple studies, with appropriate allowance for the uncertainty of the parameter estimates. Further extensions could include allowance for various biases and uncertainty on the inputs to the model.

TABLE 21 Data to be observed in the colon cancer screening example

Treatment	Offered screening	Not offered screening
Colon cancer deaths	r_t	r_c
Number of cases	n_t	n_c

Analysis of colon cancer screening trial: power calculations allowing for cross-over between treatment arms

Reference. This example forms chapter 30 of Eddy *et al.*[150]

The technology. Screening for colon cancer.

Objectives of analysis. To estimate the probability of a statistically significant result in a future trial

TABLE 22 Model and notation for colon cancer screening example

Factor	Notation	Derivation
Proportion of those offered screening who cross-over	d_t	
Proportion of those not offered screening who cross-over	d_c	
Mortality rate in group offered screening	p_t	$(1 - d_t)\theta_t + d_t\theta_c$
Mortality rate in group not offered screening	p_c	$(1 - d_c)\theta_c + d_c\theta_t$
Chi-squared statistic	chisquare	$\frac{[r_t(n_c - r_c) - r(n_t - r_c)]^2(n_t + n_c)}{n_t n_c(r_t + r_c)(n_t + n_c - r_t - r_c)}$
Significant result?	Significant?	chisquare > 3.84

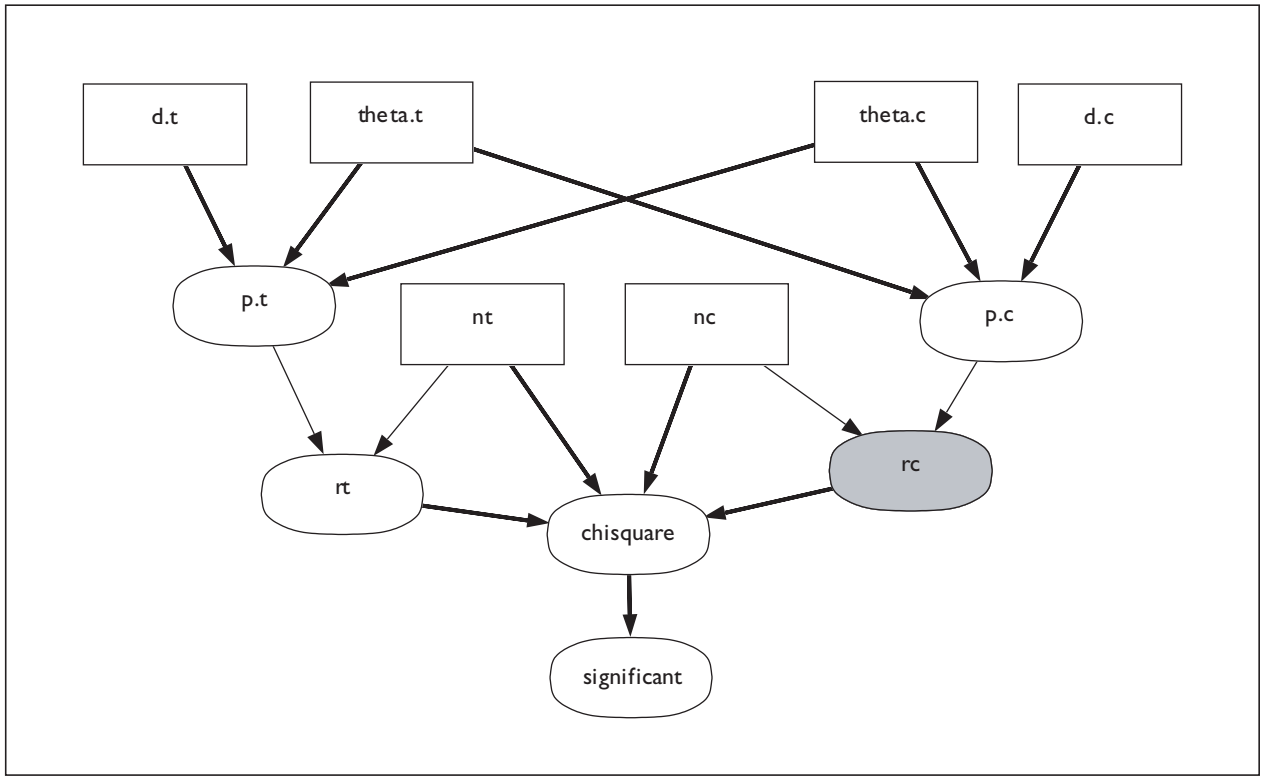


FIGURE 17 Doodle of the model for predicting the power of the colon cancer screening trial

TABLE 23 *Power of the proposed colon cancer screening trial*

Sample size per group	With dilution		Without dilution	
	WinBUGS	Eddy	WinBUGS	Eddy
50,000	0.21	**0.48**	0.66	**0.66**
150,000	0.53	**0.91**	0.98	**0.98**

(the power), assuming a two-sided type I error (α) of 0.05. This will be calculated assuming 50,000 individuals per group as an example. The analysis will be by intention-to-treat, but we want to adjust for the possibility of some of those offered screening 'crossing over' to the unscreened group, and some of those not offered screening crossing over to the screened group. Eddy terms this 'dilution'.

Design of study. Proposed randomised controlled trial.

Available evidence in study. None yet, but the final evidence will have the form shown in *Table 21*.

Statistical model. Let p_t and p_c be the underlying mortality rate of those offered and not offered screening, respectively, so that $r_c \sim \text{Binomial}(n_c, p_c)$, $r_t \sim \text{Binomial}(n_t, p_t)$. Let θ_t and θ_c be the assumed mortality rates in those actually obtaining and not obtaining screening, respectively. Under the cross-over assumption, we have the relationships shown in *Table 22*. The graph of the model is shown in *Figure 17*, from which the WinBUGS analysis is driven.

Prior distribution. Eddy and colleagues assume the following point estimates for the parameters: $\theta_c = 0.005$, $\theta_t = 0.004$, $d_t = 0.3$, $d_c = 0.2$. Thus they are attempting to detect a 20% mortality reduction, assuming 30% of those offered screening refuse, and 20% of those not offered screening are screened outside the trial.

Computations. Sampling from the binomial distribution with a large denominator is slow, and so 5000 iterations took nearly 3 minutes on a 400 MHz personal computer.

Results. These are shown in *Table 23*, with the results of Eddy *et al.* in bold; 'dilution' refers to the allowance for cross-over between intervention groups.

Comments. The results of Eddy *et al.* appear to be incorrect for the trial with dilution. It is clear that if a moderate amount of cross-over is plausible, then a clinical trial needs to be very large in order to have a reasonable chance of correctly obtaining a significant conclusion.

Commentary

The confidence profile technique can easily be transformed into the form of Bayesian graphical modelling used by the WinBUGS software. It is a very powerful method, but has perhaps been limited by its implementation, and WinBUGS appears to allow straightforward application. However, there are inevitably dangers with such a modelling exercise, which can only be as good as its structural assumptions and the quality of the data going into it.

Chapter 12

Case study 4: comparison of *in vitro* fertilisation clinics

The UK Human Fertilisation and Embryology Authority (HFEA) has a responsibility to monitor clinics in the UK licensed to carry out donor insemination (DI) and *in vitro* fertilisation (IVF). Their annual publication – *The Patients Guide to DI and IVF Clinics* – is designed to help people who are considering fertility treatment to understand the services offered by licensed clinics and to decide which clinic is best for them.[245]

One of the statistics provided regarding IVF treatment at each clinic is an adjusted live-birth rate per treatment cycle started. A **live birth** is defined as any birth event in which at least one baby is born and survives for more than 1 month, and a **treatment cycle** begins with the administration of drugs to induce superovulation. The adjustment, which is intended to take account of the mix of patients treated by the clinic by using factors such as age, duration of infertility, number of previous treatment cycles and so on,[441] varies from year to year and is based on a pooled logistic regression of all IVF treatments carried out in the UK in the relevant year. Also provided are associated 95% confidence intervals for each adjusted live-birth rate.

Marshall and Spiegelhalter[312] analyse the data published in 1996, using *Figure 18* to show the substantial range of success rates displayed by the clinics.

They then use MCMC techniques to derive posterior distributions for the ranks of the institutions: this is easily done by calculating the current rank of each institution at each iteration of the simulation, and then summarising the distribution of these calculated ranks after many thousands of iterations. *Figure 19* shows that there is considerable uncertainty in the true rank of an institution, even when they show substantial differences in performance.

We now assume the clinics are fully exchangeable (see page 21) with the true rates (on a logit scale) being drawn from a common normal distribution: if, after adjusting for case mix, we can find no other contextually meaningful way to differentiate between the institutions *a priori*, then the assumption of their exchangeability seems justified. It is clear from *Figure 20* that there is substantial 'shrinkage' towards the overall mean performance, although there are still a number of clinics that would be considered 'significantly' above or below average. It can be argued that this adjustment is an appropriate means of dealing with the problem of multiple comparisons. In addition, this shrinkage should deal with 'regression-to-the-mean', in which extreme institutions will tend back towards the overall average when they recover from their temporary run of good or bad luck.

The consequence of assuming exchangeability is to reduce the differences between clinics and hence to make their ranks even more uncertain. *Figure 21* shows this is the case to a limited extent, although since many of the extreme clinics are also fairly large, their rank is not unduly affected.

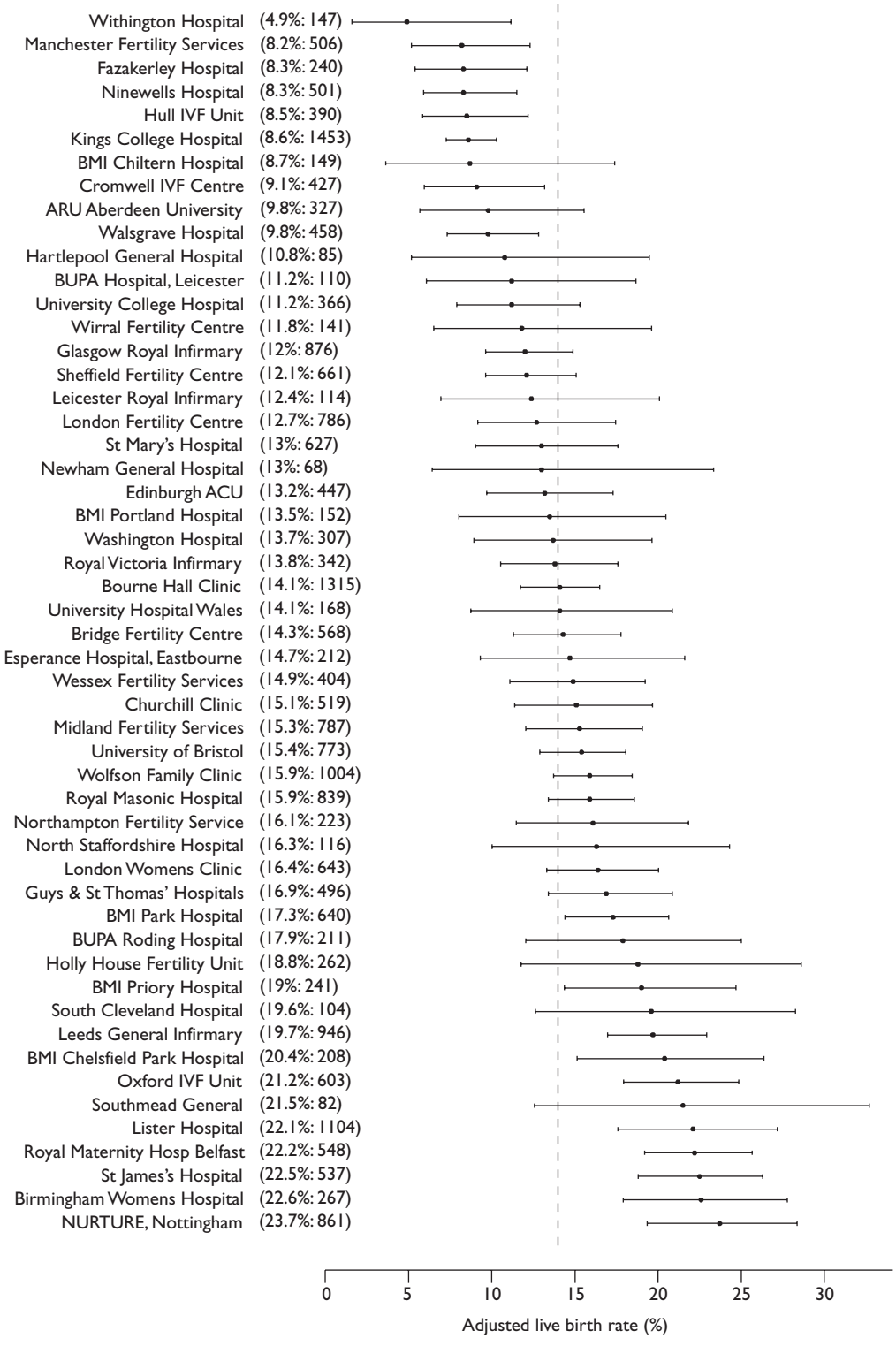


FIGURE 18 *Estimates and 95% intervals for the adjusted live-birth rate in each clinic. The vertical line represents the national average of 14%. The estimated adjusted live-birth rate for each clinic is given in parentheses, together with the number of treatment cycles started*

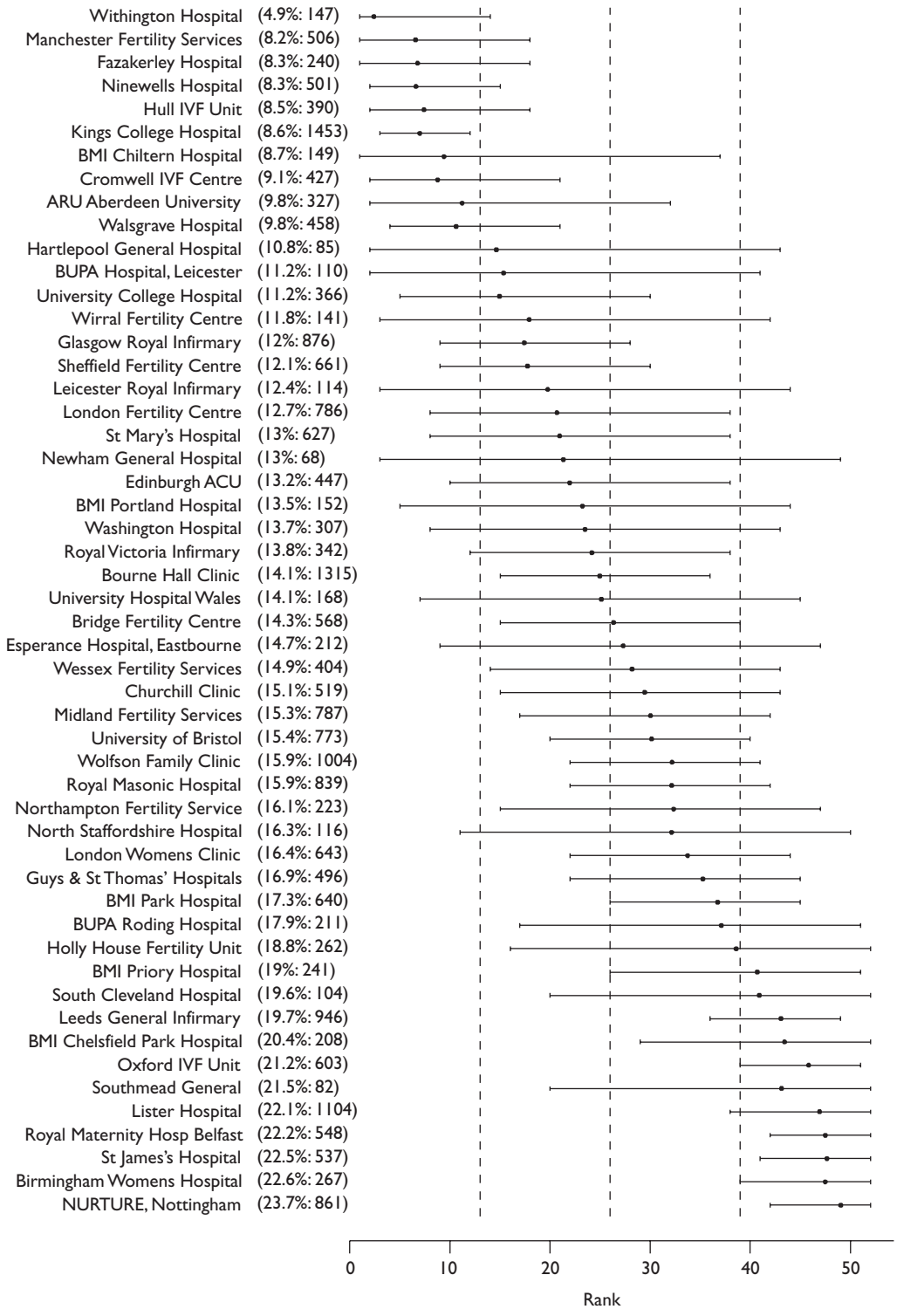


FIGURE 19 *Median and 95% intervals for the rank of each clinic. The estimated adjusted live-birth rate for each clinic is given in parentheses, together with the number of treatment cycles started. The dashed vertical lines divide the clinics into quarters according to their rank*

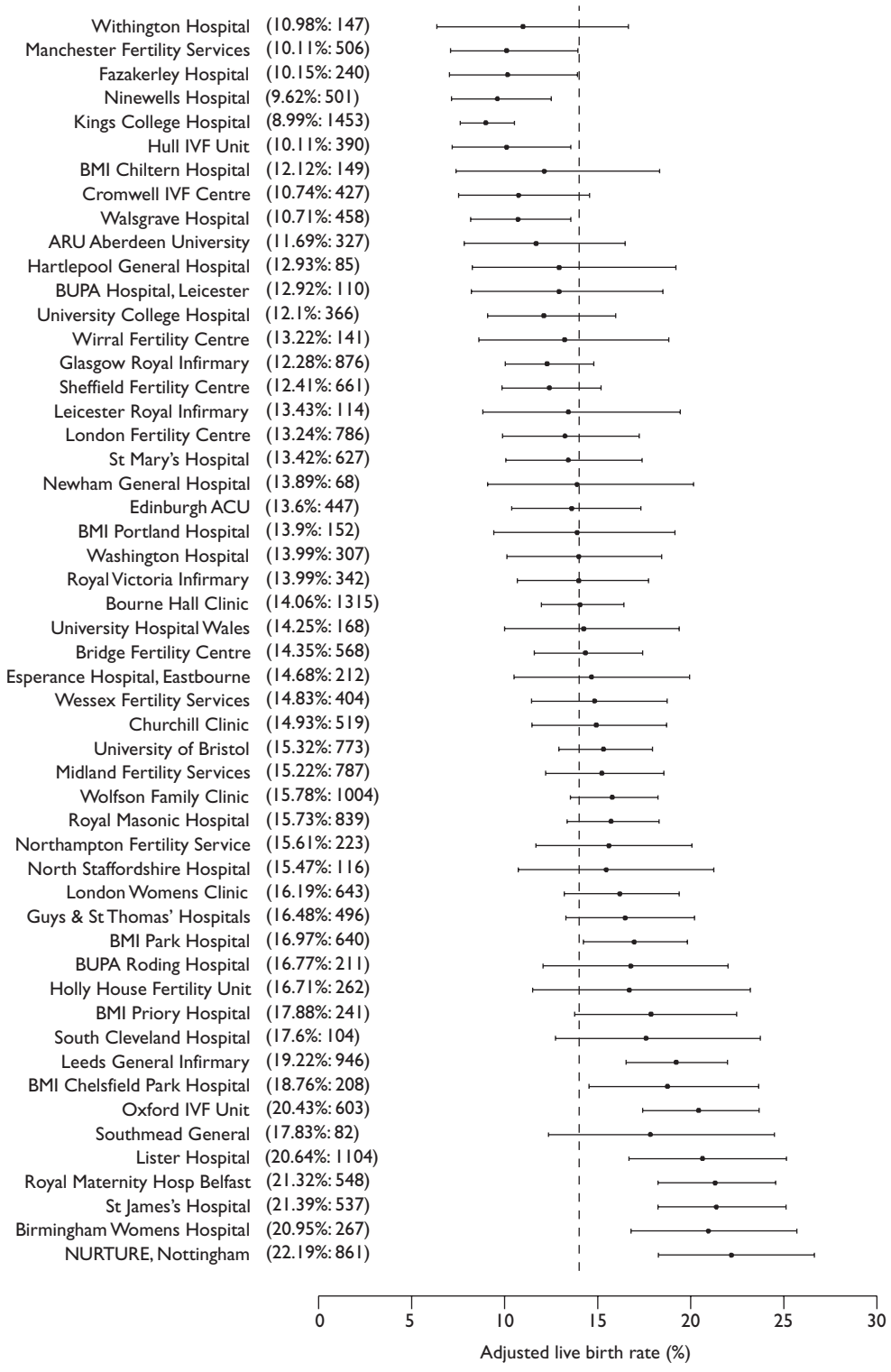


FIGURE 20 *Estimates and 95% intervals for the adjusted live-birth rate in each clinic, assuming exchangeability between clinics. The dashed vertical line represents the national average of 14%*

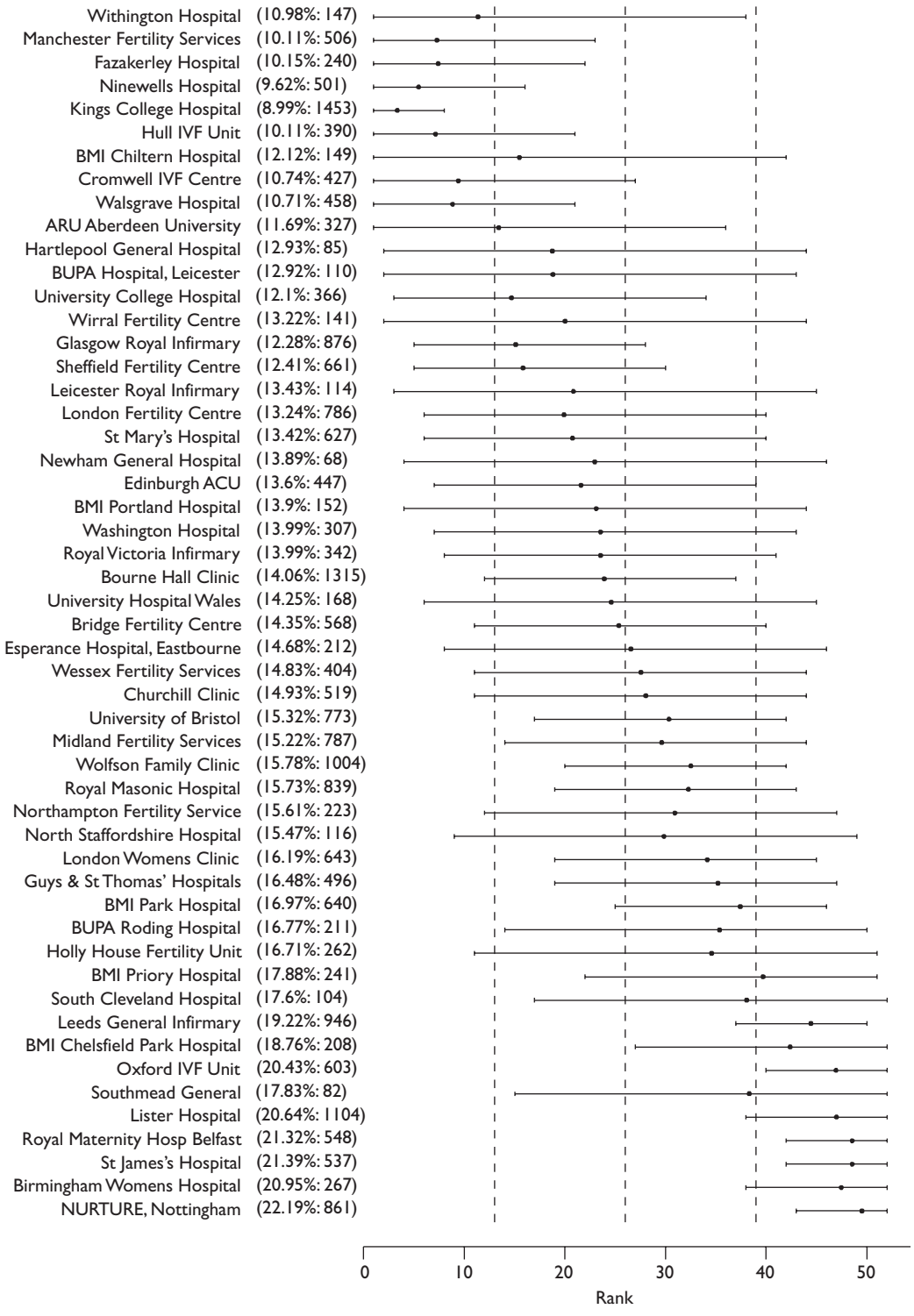


FIGURE 21 *Estimated true ranks and 95% intervals for each clinic, assuming exchangeability*

Chapter 13

Conclusions and implications for future research

Introduction

This review has described the general use of Bayesian methods in health technology assessment, and has considered a number of specific areas of application, for example, randomised controlled trials. While in many of these areas the advantages of adopting a Bayesian approach have been clearly demonstrated, a number of problems have also been identified. This chapter summarises many of these advantages and disadvantages, and makes a series of recommendations for the main participant groups in health technology assessment. The next section summarises the specific conclusions that have been drawn in the preceding chapters, after which the general advantages and problems associated with adopting a Bayesian approach are considered. The chapter concludes by summarising the areas requiring further research.

Specific conclusions

Introduction

1. Bayesian methods are defined as **the explicit quantitative use of external evidence in the design, monitoring, analysis, interpretation and reporting of a health technology assessment**.
2. Bayesian methods are a controversial topic in that they may involve the explicit use of subjective judgements in what is conventionally supposed to be a rigorous scientific exercise in health technology assessment.
3. There has been very limited use of proper Bayesian methods in practice, and relevant studies appear to be relatively easily identified.
4. The potential importance of Bayesian methods to a topic is not necessarily reflected in the volume of published literature: in particular, publications on the design and analysis of single clinical trials dominate those on the synthesis of evidence from studies of multiple designs.

Bayesian methods in the health technology assessment context

1. Claims of advantages and disadvantages of Bayesian methods are now largely based on pragmatic reasons rather than blanket ideological positions.
2. A Bayesian approach can lead to flexible modelling of evidence from diverse sources.
3. Bayesian methods are best seen as a transformation from an initial to a final opinion, rather than providing a single 'correct' inference.
4. Different contexts may demand different statistical approaches, both regarding the role of prior opinion and the role of an explicit loss function. It is vital to establish contexts in which Bayesian approaches are appropriate.
5. A decision-theoretic approach may be appropriate where the consequences of a study are predictable, such as when dealing with rare diseases treated according to a protocol, within a pharmaceutical company, or in public health policy.

The prior distribution

1. The use of a prior is based on judgement, and hence a degree of subjectivity cannot be avoided.
2. The prior is important and not unique, and so a range of options should be examined in a sensitivity analysis.
3. The intended audience for the analysis needs to be explicitly specified.
4. The quality of subjective priors (as assessed by predictions) show predictable biases in terms of enthusiasm.
5. For a prior to be taken seriously, its evidential basis must be explicitly given, as well as any assumptions made (e.g. downweighting of past data). Care must, however, be taken of bias in published results.
6. Archetypal priors may be useful for identifying a reasonable range of prior opinion.
7. Great care is required in using default priors intended to be minimally informative.
8. Exchangeability assumption should not be made lightly.

Randomised trials

1. The Bayesian approach provides a framework for considering the ethics of randomisation.
2. Monitoring trials with sceptical and other priors may provide a unified approach to assessing whether a trial's results should be convincing to wide range of reasonable opinion, and could provide a formal tool for Data Monitoring Committees.

3. Various sources of multiplicity can be dealt with in a unified and coherent way.
4. In contrast to earlier phases of development, it is generally unrealistic to formulate a Phase III trial as a decision problem, except in circumstances where future treatments can be accurately predicted.
5. An empirical basis for prior opinions in clinical trials should be investigated, but archetypal prior opinions play a useful role.
6. The structure in which trials are conducted must be recognised, but can be taken into account by specifying a range of prior opinions.

Observational studies

1. Epidemiological studies tend to demand a more complex analysis than randomised trials.
2. Computer-intensive Bayesian methods in epidemiology are becoming more common.
3. There are likely to be increased demands, particularly in areas such as institutional comparisons and gene–environment interactions.
4. The explicit modelling of potential biases in observational data may be widely applicable but needs some evidence base in order to be convincing.

Evidence synthesis

1. A unified Bayesian approach appears to be applicable to a wide range of problems concerned with evidence synthesis.
2. In the past, prospective evaluation of clinical interventions concentrated on randomised controlled trials, but more recent interest has focused on more diffuse areas, such as healthcare delivery or broad public health measures. This means methods that can synthesise totality of evidence are required, for example in assessing medical devices.
3. Evaluations of current technologies may often be seen as unethical subjects for randomised controlled trials, and hence modelling of available evidence is likely to be necessary.
4. Perhaps one reason for lack of uptake is that syntheses are not seen as 'clean' methods, with each analysis being context-specific, less easy to set quality markers for, easier to criticise as subjective and so on.
5. Priors for the degree of 'similarity' between alternative designs can be empirically informed by studies comparing the results of randomised controlled trials and observational data.

Strategy, decisions and policy making

1. A Bayesian approach allows explicit recognition of multiple perspectives.
2. Increased attention to pharmaco-economics may lead decision-theoretic models for research planning to be explored, although this will not be straightforward.
3. There appears to be great potential for formal methods for planning in the pharmaceutical industry.
4. The regulation of devices is leading the way in establishing the role of evidence synthesis.
5. 'Comprehensive decision modelling' is likely to become increasingly important in policy making.

Practical examples and case studies

1. The BayesWatch criteria may provide a basis for structured reporting of Bayesian analysis.
2. Summaries of fully fledged ('three-star') applications of Bayesian methods in health technology assessment contain few prospective analyses but provide useful guidance.
3. Four case studies show:
   a. Bayesian analyses using a sceptical prior can be useful to the data monitoring committee of a cancer clinical trial (case study 1 (chapter 9): the CHART trial).
   b. Bayesian methods can be used to temper over optimistic conclusions based on meta-analysis of small trials (case study 2 (chapter 10): magnesium sulphate after AMI).
   c. Modern graphical software can easily handle complex assessments previously analysed using the 'confidence profile' method (case study 3 (chapter 11)).
   d. Bayesian methods provide a flexible tool for performance estimation and ranking of institutions (case study 4 (chapter 12): IVF clinics).

General advantages and problems

Potential advantages of Bayesian approaches in health technology assessment

1. All evidence regarding a specific problem can be taken into account.
2. Specification of a prior distribution requires sponsors, investigators and policy makers to think carefully and be explicit about what external evidence and judgement they should include.
3. Hierarchical models, which also can be handled within a non-Bayesian framework, allow pooling of evidence and 'borrowing of strength' between multiple substudies.
4. Potential biases can be explicitly modelled, allowing the synthesis of studies of varying designs.

5. The Bayesian approach focuses on the vital question 'How should this piece of evidence change what we currently believe?'.
6. Probability statements can be made directly regarding quantities of interest, and predictive statements are easily derived.
7. Juxtaposition of current belief with clinical demands provides an intuitive and flexible mechanism for monitoring and reporting studies.
8. The inferential outputs from a Bayesian analysis feed naturally into a decision-theoretic and policy-making context.
9. Explicit recognition of the importance of context makes Bayesian methods particularly suitable for health technology assessment, in which multiple parties may well interpret the same evidence in different ways.

Generic problems

1. Unfamiliarity with Bayesian techniques, perhaps along with their perceived mathematical complexity, and some conservatism on the part of potential users, has resulted in limited use of proper Bayesian methods in health technology assessment practice to date.
2. The use of prior opinions acknowledges a subjective input into analyses, which may appear to contravene the scientific aim of objectivity.
3. Specification of priors, whether by elicitation or choice of defaults, is a contentious and difficult issue.
4. There are no established standards for the design, analysis and reporting of Bayesian studies.
5. A full decision-theoretic framework can lead to innovative but non-standard trial designs very different from those currently in use.
6. Specification of expected utilities is difficult and may require extensive assumptions about future use of technology.
7. There is no automatic measure of statistical significance and lack of model fit, such as a deviance measure and a *P* value.
8. The computational complexity of the methods has been a major issue until recently.
9. Software for implementation of the methods is still limited in availability and user-friendliness.

Many of the issues raised above have been mentioned in more specific contexts by others, for example Sutton *et al.*[434]

Future research and development

Bayesian methods could be of great value within health technology assessment. For a realistic appraisal of the methodology, it is necessary to distinguish the roles and requirements for five main participant groups in health technology assessment: methodological researchers, sponsors, investigators, reviewers and consumers. However, two common themes for all participants can immediately be identified. First, the need for an extended set of case studies showing practical aspects of the Bayesian approach, in particular for prediction and handling multiple substudies, in which mathematical details are minimised but details of implementation are provided. Second, the development of standards for the performance and reporting of Bayesian analyses, possibly derived from the BayesWatch checklist described in this report.

The potential roles for each of the participant groups are summarised below under each aspect of a Bayesian assessment: design, priors, modelling, reporting and decision-making.

1. **Methodological researchers:**
   a. **Design**. There is a need for transferable methods for sample size calculation that are not based on type I and type II error, such as targeting precision, and realistic development of payback models, with modelling of dissemination.
   b. **Priors.** Simple and reliable elicitation methods for 'non-enthusiasts' require testing, as well as demonstrations of the use of empirical data as a basis for prior distributions. Reasonable default priors in non-standard situations need to be available.
   c. **Modelling**. Methods for flexible model selection and robust MCMC analysis require development and dissemination. With regard to implementation, there is a need for user-friendly software for clinical trials and evidence synthesis.
   d. **Reporting**. It is essential to have appraisal criteria along the lines of the BayesWatch checklist, with possible reformulation as guidelines along the lines of 'How to read a Bayesian study'. It would be useful to have the term 'Bayesian' in all relevant papers in order to aid literature searches.
   e. **Decision-making**. Increased integration with a health economic and policy perspective is highly desirable, together with flexible tools for implementation.

2. **Sponsors**:
   a. **Design.** Both public sector and industry could extend their perspective beyond the classical Neyman–Pearson criteria, and in particular

investigate quantitative payback models. The pharmaceutical industry might also investigate formal project prioritisation schemes.
   b. **Priors.** All sponsors could focus on the evidential basis for assumptions made concerning alternative hypotheses and the potential gains from technology, and use empirical reviews to establish reasonable prior opinions.

3. **Investigators:**
   a. **Design.** Apart from the considerations given above under 'sponsors', there is also potential for 'open' studies in which interim results are reported to investigators.
   b. **Priors.** It would be valuable to gain experience in eliciting prior opinions from both enthusiasts and a general cross-section of the target community.
   c. **Modelling.** There is great scope, when analysing data, to go beyond the usual limited list of models and consider a range of priors and structural assumptions.
   d. **Reporting.** It is vital that any Bayesian reporting allows future users to include the evidence in their synthesis or decision. The use of BayesWatch or a similar scheme for reporting should help in this.

4. **Reviewers/regulatory bodies:**
   a. **Priors.** Regulatory bodies could establish reasonable prior opinions based on past experience in order to provide default priors.
   b. **Modelling.** Regulatory bodies could take a more flexible approach to the use of data, particularly in areas such as medical devices, and encourage efficient use of data by appropriate use of historical controls, evidence synthesis and so on.
   c. **Decision-making.** More experimental would be the explicit modelling of the consequences of decisions in order to decide evidential criteria.

5. **Consumers/policy makers**. There is a need for careful case studies in which policy makers explicitly go through the following stages in reaching a conclusion based on a full Bayesian analysis:
   a. **Priors**: specify prior opinions relevant at the time of decision-making
   b. **Modelling**: pool all available evidence into a coherent model
   c. **Reporting**: make predictive probability statements about the consequences of different policies
   d. **Decision-making**: assign costs to potential consequences and so assess (with sensitivity analysis) the expected value of different actions.

Acknowledgements

This study was commissioned by the NHS R&D HTA programme. The authors are indebted to the HTA referees for their perseverance in reading this report and the quality of their comments. The views expressed in this report are those of the authors, who are responsible for any errors.

References

1. Abrams K, Ashby D, Errington D. Simple Bayesian analysis in clinical trials – a tutorial. *Controlled Clin Trials* 1994;**15**:349–59.
2. Abrams K, Ashby D, Houghton J, Riley D. Assessing drug interactions: tamoxifen and cyclophosphamide. In: Berry and Stangl,[61] p. 3–66.
3. Abrams K, Jones DR. Meta-analysis and the synthesis of evidence. *IMA J Math Appl Med Biol* 1995;**12**:297–313.
4. Abrams K, Sansó B. Approximate Bayesian inference for random effects meta-analysis. *Stat Med* 1998;**17**:201–218.
5. Abrams KR. Monitoring randomised controlled trials. Parkinson's disease trial illustrates the dangers of stopping early. *BMJ* 1998;**316**:1183–4.
6. Abrams KR, Jones DR. Bayesian interim analysis of randomised trials. *Lancet* 1997;**349**:1911–12.
7. Achcar JA, Brookmeyer R, Hunter WG. An application of Bayesian analysis to medical follow-up data. *Stat Med* 1985;**4**:509–20.
8. Adar R, Critchfield GC, Eddy DM. A confidence profile analysis of the results of femoropopliteal percutaneous transluminal angioplasty in the treatment of lower-extremity ischemia. *J Vasc Surg* 1989;**10**:57–67.
9. Adcock CJ. The Bayesian approach to determination of sample sizes – some comments on the paper by Joseph, Wolfson and Du Berger. *J R Stat Soc Ser D* 1995;**44**:155–61.
10. Ahn C. An evaluation of Phase I cancer clinical trial designs. *Stat Med* 1998;**17**(14):1537–49.
11. Albert J, Chib S. Bayesian modelling of binary repeated measures data with application to cross-over trials. In: Berry and Stangl,[61] p. 577–600.
12. Altman DG. Bayesian approaches to randomised trials – discussion. *J R Stat Soc Ser A* 1994;**157**: 387–416.
13. Anscombe F. Sequential medical trials. *J Am Stat Assoc* 1963;**58**:365–83.
14. Antman EM. Magnesium in acute MI – timing is critical. *Circulation* 1995;**92**:2367–72.
15. Armitage P. The search for optimality in clinical trials. *Int Stat Rev* 1985;**53**:15–24.
16. Armitage P. Discussion of Breslow (1990). *Stat Sci* 1990;**5**.
17. Armitage P. Interim analysis in clinical trials. *Stat Med* 1991;**10**(6):925–37.
18. Armitage P. Letter to the editor. *Controlled Clin Trials* 1991;**12**:345.
19. Armitage P. A case for Bayesianism in clinical trials – discussion. *Stat Med* 1993;**12**:1395–404.
20. Armitage P, Colton T, editors. Encyclopaedia of biostatistics. Chichester: Wiley; 1998.
21. Armitage P, Mcpherson K, Rowe B. Repeated significance tests on accumulating data. *J R Stat Soc* A 1969;**132**:235–44.
22. Armitage PA. Inference and decision in clinical trials. *J Clin Epidemiol* 1989;**42**:293–9.
23. Ashby D, Hutton J. Bayesian epidemiology. In: Berry and Stangl,[61] p. 109–138.
24. Ashby D, Hutton J, McGee M. Simple Bayesian analyses for case–control studies in cancer epidemiology. *Statistician* 1993;**42**:385–97.
25. Atkinson EN. A Bayesian strategy for evaluating treatments applicable only to a subset of patients. *Stat Med* 1997;**16**:1803–15.
26. Avins AL. Bayesian analysis and the GUSTO trial. *J Am Med Assoc* 1995;**274**:873.
27. Ayanian J, Landrum M, Normand S, Guadagnoli E, McNeil B. Rating the appropriateness of coronary angiography – do practicing physicians agree with an expert panel and with each other? *N Eng J Med* 1998;**338**:1896–904.
28. Baraff LJ, Oslund S, Prather M. Effect of antibiotic therapy and etiologic microorganism on the risk of bacterial meningitis in children with occult bacteremia. *Pediatrics* 1993;**92**:140–3.
29. Barnett V. Comparative statistical inference. 2nd ed. Chichester: Wiley; 1982.
30. Bather JA. On the allocation of treatments in sequential medical trials. *Int Stat Rev* 1985;**53**:1–13.
31. Baudoin C, O'Quigley J. Symmetrical intervals and confidence intervals. *Biometr J* 1994;**36**:927–34.
32. Bayes T. An essay towards solving a problem in the doctrine of chances. *Phil Trans R Soc* 1763;**53**:418.
33. Begg CB. Book review of 'cross design synthesis'. *Stat Med* 1992;**12**:1627–30.

34. Begg CB, Dumouchel W, Harris J, Dobson A, Dear K, Givens GH, *et al.* Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate – comments and rejoinders. *Stat Sci* 1997;**12**:241–50.

35. Belin TR, Elashoff RM, Leung K, Nisembaum R, Bastani R, Nasseri K, *et al.* Combining information from multiple sources in the analysis of non-equivalent control group design. In: Gatsonis C, Hodges J, Kass R, Singpurwalla N, editors. Case studies in Bayesian statistics II. Berlin: Springer-Verlag; 1995. p. 241–260.

36. Berger J. Statistical decision theory and Bayesian inference. Berlin: Springer-Verlag 1985.

37. Berger J, Wolpert R. The Likelihood Principle. 2nd ed. Hayward (CA): Institute of Mathematical Statistics; 1988.

38. Berger JO, Berry DA. Statistical analysis and the illusion of objectivity. *Am Sci* 1988;**76**:159–65.

39. Bergman S, Gittins J, editors. Statistical methods for pharmaceutical research planning. New York: Marcel Dekker; 1985.

40. Bernardinelli L, Clayton D, Pascutto C, Montomoli C, Ghislandi M, Songini M. Bayesian analysis of space-time variation in disease risk. *Stat Med* 1995; **14**:2433–43.

41. Bernardo J, Berger JO, Lindley DV, Smith AFM, editors. Bayesian statistics 4. Oxford: Oxford University Press; 1992.

42. Bernardo J, DeGroot MH, Lindley DV, Smith AFM, editors. Bayesian statistics 3. Oxford: Oxford University Press; 1988.

43. Bernardo JM, Smith AFM. Bayesian theory. Chichester: Wiley; 1994.

44. Berry D, Hardwick J. Using historical controls in clinical trials: application to ECMO. In: Berger J, Gupta S, editors. Statistical decision theory and related topics V. New York: Springer-Verlag; 1993. p. 141–56.

45. Berry D, Stangl D. Bayesian methods in health-related research. In: Berry and Stangl,[61] p. 3–66.

46. Berry DA. Ethics and ECMO. Comments on 'Investigating therapies of potentially great benefit: ECMO' by J H Ware. *Stat Sci* 1989;**4**:306–10.

47. Berry DA, Hardwick J. Recent progress in clinical trial designs that adapt for ethical purposes: comments on 'Investigating therapies of potentially great benefit: ECMO' by J H Ware. *Stat Sci* 1989;**4**:327–36.

48. Berry DA. Interim analyses in clinical trials – classical vs Bayesian approaches. *Stat Med* 1985; **4**:521–6.

49. Berry DA. Interim analyses in clinical research. *Cancer Invest* 1987;**5**:469–77.

50. Berry DA. Interim analysis in clinical trials – the role of the likelihood principle. *Am Stat* 1987; **41**:117–22.

51. Berry DA. Monitoring accumulating data in a clinical trial. *Biometrics* 1989;**45**:1197–211.

52. Berry DA. Bayesian methods in Phase III trials. *Drug Information J* 1991;**25**:345–68.

53. Berry DA. A case for Bayesianism in clinical trials. *Stat Med* 1993;**12**:1377–93.

54. Berry DA. Bayesian approaches to randomised trials – discussion. *J R Stat Soc Ser A* 1994;**157**: 387–416.

55. Berry DA. Decision analysis and Bayesian methods in clinical trials. *Cancer Treatment Res* 1995;**75**: 125–54.

56. Berry DA. Statistics: a Bayesian perspective. Belmont (CA): Duxbury Press; 1996.

57. Berry DA. When is a confirmatory randomised clinical trial needed? *J Natl Cancer Inst* 1996; **88**:1606–7.

58. Berry DA, Eick SG. Adaptive assignment versus balanced randomization in clinical trials – a decision-analysis. *Stat Med* 1995;**14**:231–46.

59. Berry DA, Ho CH. One-sided sequential stopping boundaries for clinical trials – a decision-theoretic approach. *Biometrics* 1988;**44**:219–27.

60. Berry DA, Pearson LM. Optimal designs for clinical trials with dichotomous responses. *Stat Med* 1985; **4**:597–608.

61. Berry DA, Stangl DK, editors. Bayesian biostatistics. New York: Marcel Dekker; 1996.

62. Berry DA, Thor A, Cirrincione C, Edgerton S, Muss H, Marks J, *et al.* Scientific inference and predictions: Multiplicities and convincing stories: A case study in breast cancer therapy. In: Bernardo J, Berger JO, Lindley DV, Smith AFM, editors. Bayesian statistics 5. Oxford: Oxford University Press; 1996. p. 45–67.

63. Berry DA, Wolff MC, Sack D. Public health decision making: a sequential vaccine trial. In: Bernardo *et al.*,[41] p. 79–96.

64. Berry DA, Wolff MC, Sack D. Decision making during a Phase III randomised controlled trial. *Controlled Clin Trials* 1994;**15**:360–78.

65. Berry SM, Kadane JB. Optimal Bayesian randomization. *J R Stat Soc Ser B* 1997;**59**:813–19.

66. Biggerstaff BJ, Tweedie RL, Mengersen KL. Passive smoking in the workplace: classical and Bayesian

meta-analyses. *Int Arch Occupat Environ Health* 1994;**66**:269–77.

67. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;**312**:1215–18.

68. Bland JM, Altman DG. Statistics notes: Bayesians and frequentists. *BMJ* 1998;**317**:1151.

69. Box GEP. Sampling and Bayes inference in scientific modelling and robustness (with discussion). *J R Stat Soc A* 1980;**143**:383–430.

70. Box GEP, Tiao GC. Bayesian inference in statistical analysis. Reading: Addison–Wesley; 1973.

71. Brant LJ, Duncan DB, Dixon DO. K-ratio *t*-tests for multiple comparisons involving several treatments and a control. *Stat Med* 1992;**11**:863–73.

72. Breslow N. Biostatistics and Bayes. *Stat Sci* 1990; **5**(3):269–84.

73. Briggs A. A Bayesian approach to stochastic cost-effectiveness analysis. *Electronic Health Electronics Lett* 1998;**2**(4):6–13.

74. Briggs AH, Gray AM. Handling uncertainty when performing economic evaluation of healthcare interventions. *HTA* 1999;**3**(2).

75. Bring J. Stopping a clinical trial early because of toxicity – the Bayesian approach. *Controlled Clin Trials* 1995;**16**:131–2.

76. Brophy JM, Joseph L. Placing trials in context using Bayesian analysis – GUSTO revisited by Reverend Bayes. *J Am Med Assoc* 1995;**273**:871–5.

77. Brophy JM, Joseph L. Bayesian interim statistical analysis of randomised trials. *Lancet* 1997;**349**: 1166–8.

78. Brown BW, Herson J, Atkinson EN, Rozell ME. Projection from previous studies – a Bayesian and frequentist compromise. *Controlled Clin Trials* 1987;**8**:29–44.

79. Browne RH. Bayesian analysis and the GUSTO trial. *J Am Med Assoc* 1995;**274**:873.

80. Browner WS, Newman TB. Are all significant *p* values created equal? The analogy between diagnostic tests and clinical research. *J Am Med Assoc* 1987;**257**:2459–63.

81. Brunier HC, Whitehead J. Sample sizes for Phase II clinical trials derived from Bayesian decision theory. *Stat Med* 1994;**13**:2493–502.

82. Burton PR. Helping doctors to draw appropriate inferences from the analysis of medical studies. *Stat Med* 1994;**13**:1699–713.

83. Buxton M, Hanney S. Evaluating the NHS research and development programme: will the programme give value for money? *J R Soc Med* 1998;**91**:2–6.

84. Byar D. Why data bases should not replace clinical trials. *Biometrics* 1980;**36**:337–42.

85. Byar DP, 21 others. Design considerations for AIDS trials. *N Eng J Med* 1990;**323**:1343–8.

86. Byar DP, Simon RM, Friedewald WT, Schlesselman JJ, DeMets DL, Ellenberg JH, *et al.* Randomised clinical trials. perspective on some recent ideas. *N Eng J Med* 1976;**295**:74–80.

87. Campbell G. A regulatory perspective for Bayesian clinical trials. Food and Drug Administration; 1999. URL http://www.fda.gov/cdrh/present/Bayesian.pdf

88. Canner PL. Selecting one of two treatments when the responses are dichotomous. *J Am Stat Assoc* 1970;**65**:293–306.

89. Carlin BP. Bayesian approaches to randomised trials – discussion. *J R Stat Soc Ser A* 1994;**157**: 387–416.

90. Carlin BP, Chaloner K, Church T, Louis TA, Matts JP. Bayesian approaches for monitoring clinical trials with an application to toxoplasmic encephalitis prophylaxis. *Statistician* 1993;**42**: 355–67.

91. Carlin BP, Kadane JB, Gelfand AE. Approaches for optimal sequential decision analysis in clinical trials. *Biometrics* 1998;**54**:964–75.

92. Carlin BP, Louis TA. Bayes and empirical Bayes methods for data analysis. London: Chapman and Hall; 1996.

93. Carlin BP, Sargent DJ. Robust Bayesian approaches for clinical trial monitoring. *Stat Med* 1996;**15**: 1093–106.

94. Carlin JB. Meta-analysis for 2×2 tables – a Bayesian approach. *Stat Med* 1992;**11**:141–58.

95. Chalmers I. What is the prior probability of a proposed new treatment being superior to established treatments? *BMJ* 1997;**314**:74–5.

96. Chalmers TC, Lau J. Changes in clinical trials mandated by the advent of meta-analysis. *Stat Med* 1996;**15**:1263–8.

97. Chaloner K. Elicitation of prior distributions. In: Berry and Stangl,[61] p. 141–156.

98. Chaloner K, Church T, Louis TA, Matts JP. Graphical elicitation of a prior distribution for a clinical trial. *Statistician* 1993;**42**:341–53.

99. Chaloner K, Verdinelli I. Bayesian experimental design – a review. *Stat Sci* 1995;**10**:273–304.

100. Chang MN, Shuster JJ. Interim analysis for randomised clinical trials: simulating the predictive distribution of the log-rank test statistic. *Biometrics* 1994;**50**:827–33.

101. Chelimsky E, Silberman G, Droitcour J. Cross design synthesis. *Lancet* 1993;**341:**498.

102. Chevret S. The continual reassessment method in cancer Phase I clinical trials: a simulation study. *Stat Med* 1993;**12**(12):1093–108.

103. Choi SC, Pepple PA. Monitoring clinical trials based on predictive probability of significance. *Biometrics* 1989;**45**:317–23.

104. Choi SC, Smith PJ, Becker DP. Early decision in clinical trials when the treatment differences are small: experience of a controlled trial in head trauma. *Controlled Clin Trials* 1985;**6**(4):280–8.

105. Claxton K. Bayesian approaches to the value of information: implications for the regulation of new pharmaceutical. *Health Econ* 1999;**8**:269–74.

106. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *J Health Econ* 1999;**18**:341–64.

107. Claxton K, Posnett J. An economic approach to clinical trial design and research priority-setting. *Health Econ* 1996;**5**:513–24.

108. Cole P. The evolving case–control study. *J Chronic Dis* 1979;**32**:15–27.

109. Collins R, Peto R, Flather M, ISIS-4 Collaborative Group. ISIS-4 – a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium-sulfate in 58,050 patients with suspected acute myocardial infarction. *Lancet* 1995;**345**:669–85.

110. Cook RJ, Farewell VT. Multiplicity considerations in the design and analysis of clinical trials. *J R Stat Soc Ser A* 1996;**159**:93–110.

111. Cornfield J. Sequential trials, sequential analysis and the likelihood principle. *Am Stat* 1966;**20**: 18–23.

112. Cornfield J. The Bayesian outlook and its applications. *Biometrics* 1969;**25**:617–57.

113. Cornfield J. Recent methodological contributions to clinical trials. *Am J Epidemiol* 1976;**104**:408–21.

114. Coronary Drug Project Research Group. The Coronary Drug Project. Initial findings leading to a modification of its research protocol. *J Am Med Assoc* 1970;**214**:1301–13.

115. Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *J Am Med Assoc* 1975;**231**:360–81.

116. Cowles MK, Carlin BP, Connett JE. Bayesian tobit modeling of longitudinal ordinal clinical trial compliance data with non-ignorable missingness. *J Am Stat Assoc* 1996;**91**:86–98.

117. Cox DR. Discussion of paper by Lindsey. *J R Stat Soc D* 1999;**48**:30.

118. Cox DR, Farewell VT. Statistical basis of public policy – qualitative and quantitative aspects should not be confused. *BMJ* 1997;**314**:73.

119. Craig BA, Fryback DG, Klein R., Klein BEK. A Bayesian approach to modelling the natural history of a chronic condition from observations with intervention. *Stat Med* 1999;**18**:1355–72.

120. Cressie N, Biele J. A sample-size-optimal Bayesian procedure for sequential pharmaceutical trials. *Biometrics* 1994;**50**:700–11.

121. Cronin KA, Legler JM, Etzioni RD. Assessing uncertainty in microsimulation modelling with application to cancer screening interventions. *Stat Med* 1998;**17**(21):2509–23.

122. Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Stat Med* 1997;**16**(17):1965–82.

123. Davis CE, Leffingwell DP. Empirical Bayes estimates of subgroup effects in clinical trials. *Controlled Clin Trials* 1990;**11**:37–42.

124. DeGroot MH. Optimal statistical decisions. Reading (MA): Addison-Wesley; 1970.

125. Demets D. Bayesian approaches to randomised trials – discussion. *J R Stat Soc Ser A* 1994;**157**: 387–416.

126. Demets DL. Stopping guidelines vs stopping rules – a practitioner's point of view. *Commun Stat Theory Methods* 1984;**13**:2395–417.

127. Dempster A, Selwyn M, Weeks B. Combining historical and randomised controls for assessing trends in proportions. *J Am Stat Assoc* 1983;**78**: 221–7.

128. Dempster AP. Bayesian methods. In: Armitage and Colton,[20] p. 263–271.

129. Dersimonian R. Meta-analysis in the design and monitoring of clinical trials. *Stat Med* 1996;**15**: 1237–48.

130. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clin Trials* 1986;**7**:177–88.

131. Detsky A. Using economic-analysis to determine the resource consequences of choices made in planning clinical-trials. *J Chronic Dis* 1985;**38**: 753–65.

132. Diamond GA, Forrester JS. Clinical trials and statistical verdicts: probable grounds for appeal. *Ann Int Med* 1983;**98**:385–94.

133. Digman JJ, Bryant J, Wieand HS, Fisher B, Wolmark N. Early stopping of a clinical trial when there is evidence of no treatment benefit: protocol B-14 of

the national surgical adjuvant breast and bowel project. *Controlled Clin Trials* 1998;**19**:575–88.

134. Dixon DO, Simon R. Bayesian subset analysis. *Biometrics* 1991;**47**:871–81.

135. Dixon DO, Simon R. Bayesian subset analysis in a colorectal cancer clinical trial. *Stat Med* 1992;**11**: 13–22.

136. Dixon DO, Simon R. Erratum: Bayesian subset analysis (*Biometrics* 1991;**47**(871–81)). *Biometrics* 1994;**50**:322.

137. Dominici F. Testing simultaneous hypotheses in pharmaceutical trials: a Bayesian approach. *J Biopharm Stat* 1998;**8**(2):283–97.

138. Dominici F, Parmigiani G, Wolpert R, Hasselblad V. Meta-analysis of migraine headache treatments: Combining information from heterogeneous designs. *J Am Stat Assoc* 1999;**94**:16–28.

139. Donner A. A Bayesian approach to the interpretation of subgroup results in clinical trials. *J Chronic Dis* 1982;**35**:429–35.

140. Droitcour J, Silberman G, Chelimsky E. Cross-design synthesis:a new form of meta-analysis for combining results from randomised clinical trials and medical-practice databases. *Int J Technol Asses Health Care* 1993;**9**:440–9.

141. DuMouchel W. A Bayesian model and graphical elicitation procedure for multiple comparisons. In: Bernardo *et al.*,[42] p. 127–145.

142. DuMouchel W. Bayesian meta-analysis. In: Berry D, editor. Statistical methodology in the pharmaceutical sciences. New York: Marcel Dekker, 1990. p. 509–29.

143. DuMouchel W, Berry DA. Meta-analysis for dose–response models. *Stat Med* 1995;**14**:679–85.

144. DuMouchel W, Waternaux C. Discussion of 'Hierarchical models for combining information and for meta-analyses'. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM, editors. Bayesian Statistics 4. Oxford: Clarendon Press; 1992. p. 338–41.

145. DuMouchel WH, Harris JE. Bayes methods for combining the results of cancer studies in humans and other species (with comment). *J Am Stat Assoc* 1983;**78**:293–308.

146. Eddy DM. The confidence profile method – a Bayesian method for assessing health technologies. *Operations Res* 1989;**37**(2):210–28.

147. Eddy DM, Hasselblad V, McGivney W, Hendee W. The value of mammography screening in women under age 50 years. *J Am Med Assoc* 1988;**259**(10): 1512–19.

148. Eddy DM, Hasselblad V, Shachter R. A Bayesian method for synthesizing evidence: the confidence profile method. *Int J Technol Assess Health Care* 1990;**6**(1):31–55.

149. Eddy DM, Hasselblad V, Shachter R. An introduction to a Bayesian method for meta-analysis – the confidence profile method. *Med Decision Making* 1990;**10**:15–23.

150. Eddy DM, Hasselblad V, Shachter R. Meta-analysis by the confidence profile method: the statistical synthesis of evidence. San Diego (CA): Academic Press; 1992.

151. Eddy DM, Wolpert RL, Hasselblad V. Confidence profiles – a Bayesian method for synthesizing evidence. *Med Decision Making* 1987;**7**:287.

152. Editorial. Cross design synthesis: A new strategy for studying medical outcomes? *Lancet* 1992;**340**:944–6.

153. Edwards S, Lilford R, Braunholtz D, Jackson J. Why 'underpowered' trials are not necessarily unethical. *Lancet* 1997;**350**:804–7.

154. Edwards SJL, Lilford RJ, Hewison J. The ethics of randomised controlled trials from the perspectives of patients, the public, and healthcare professionals. *BMJ* 1998;**317**:1209–12.

155. Edwards SJL, Lilford RJ, Jackson JC, Hewison J, Thornton J. Ethical issues in the design and conduct of randomised controlled trials. *HTA* 1998;**2**(14).

156. Edwards W, Lindman H, Savage L. Bayesian statistical inference for psychological research. *Psychol Rev* 1963;**70**:193–242.

157. Egger M, Smith GD. Magnesium and myocardial-infarction. *Lancet* 1994;**343**:1285.

158. Egger M, Smith GD. Misleading metaanalysis. *BMJ* 1995;**311**:753–54.

159. Emerson SS. Stopping a clinical trial very early based on unplanned interim analyses: a group sequential approach. *Biometrics* 1995;**51**:1152–62.

160. Estey E, Thall P, David C. Design and analysis of trials of salvage therapy in acute myelogenous leukemia. *Cancer Chemotherapy Pharmacol* 1997;**40**: S9–S12.

161. Etzioni R, Pepe MS. Monitoring of a pilot toxicity study with 2 adverse outcomes. *Stat Med* 1994;**13**:2311–21.

162. Etzioni RD, Kadane JB. Bayesian statistical methods in public health and medicine. *Annu Rev Public Health* 1995;**16**:23–41.

163. Faries D. Practical modifications of the continual reassessment method for Phase I cancer clinical trials. *J Biopharm Stat* 1994;**4**:147–64.

164. Fayers P. Bayesian approaches to randomised trials – discussion. *J R Stat Soc Ser A* 1994;**157**: 387–416.

165. Fayers P. Bayesian interim analysis of randomised trials. *Lancet* 1997;**349**:1911.

166. Fayers PM, Ashby D, Parmar MKB. Bayesian data monitoring in clinical trials. *Stat Med* 1997;**16**: 1413–30.

167. Feinstein AR. Clinical Biostatistics XXXIX: the haze of Bayes, the aerial palaces of decision analysis, and the computerised Ouija board. *Clin Pharmacol Therapeutics* 1977;**21**:482–96.

168. Felli JC, Hazen GB. Sensitivity analysis and the expected value of perfect information. *Med Decision Making* 1998;**18**:95–109.

169. Felli JC, Hazen GB. A Bayesian approach to sensitivity analysis. *Health Econ* 1999;**8**:263–8.

170. Fienberg S. A brief history of statistics in three and one-half chapters: a review essay. *Stat Sci* 1992;**7**(2): 208–25.

171. Fisher LD. Comments on Bayesian and frequentist analysis and interpretation of clinical trials – comment. *Controlled Clin Trials* 1996;**17**:423–34.

172. Fleming TR, Watelet LF. Approaches to monitoring clinical trials. *J Natl Cancer Inst* 1989;**81**:188–93.

173. Fletcher A, Spiegelhalter D, Staessen J, Thijs L, Bulpitt C. Implications for trials in progress of publication of positive results. *Lancet* 1993;**342**: 653–7.

174. Fluehler H, Grieve AP, Mandallaz D, Mau J, Moser HA. Bayesian-approach to bioequivalence assessment – an example. *J Pharm Sci* 1983;**72**: 1178–81.

175. Forster JJ. A Bayesian approach to the analysis of binary cross-over data. *Statistician* 1994;**43**:61–8.

176. Freedman B. Equipoise and the ethics of clinical research. *N Eng J Med* 1987;**317**:141–5.

177. Freedman L. Bayesian statistical methods – a natural way to assess clinical evidence. *BMJ* 1996;**313**:569–70.

178. Freedman L, Anderson G, Kipnis V, Prentice R, Wang CY, Rossouw J, *et al.* Approaches to monitoring the results of long-term disease prevention trials: Examples from the women's health initiative. *Controlled Clin Trials* 1996; **17**(6):509–25.

179. Freedman LS, Lowe D, Macaskill P. Stopping rules for clinical trials incorporating clinical opinion. *Biometrics* 1984;**40**:575–86.

180. Freedman LS, Spiegelhalter DJ. The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *Statistician* 1983;**32**:153–60.

181. Freedman LS, Spiegelhalter DJ. Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Controlled Clin Trials* 1989;**10**:357–67.

182. Freedman LS, Spiegelhalter DJ. Application of Bayesian statistics to decision making during a clinical trial. *Stat Med* 1992;**11**:23–35.

183. Freedman LS, Spiegelhalter DJ, Parmar MKB. The what, why and how of Bayesian clinical trials monitoring. *Stat Med* 1994;**13**:1371–83.

184. Freeman P. The role of *p*-values in analyzing trial results. *Stat Med* 1993;**12**:1443–52.

185. Frei A, Cottier C, Wunderlich P, Ludin E. Glycerol and dextran combined in the therapy of acute stroke. *Stroke* 1987;**18**:373–9.

186. Freirich E, Gehan E, *et al.* The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukaemia: a model for evaluation of other potentially useful therapy. *Blood* 1963;**21**:699–716.

187. Gatsonis C, Greenhouse JB. Bayesian methods for Phase I clinical trials. *Stat Med* 1992;**11**:1377–89.

188. Gelman A, Carlin J, Stern H, Rubin DB. Bayesian data analysis. New York: Chapman and Hall; 1995.

189. Gelman A, Rubin DB. Markov chain Monte Carlo methods in biostatistics. *Stat Methods Med Res* 1996; **5**:339–55.

190. Genest C, Zidek J. Combining probability distributions: a critique and an annotated bibliography (with discussion). *Stat Sci* 1986;**1**: 114–48.

191. George SL, Li CC, Berry DA, Green MR. Stopping a clinical trial early – frequentist and Bayesian approaches applied to a CALGB trial in non-small-cell lung cancer. *Stat Med* 1994;**13**:1313–27.

192. Gilbert JP, McPeek B, Mosteller F. Statistics and ethics in surgery and anesthesia. *Science* 1977; **198**:684–9.

193. Gilks WR, Richardson S, Spiegelhalter DJ. Markov chain Monte Carlo methods in practice. New York: Chapman and Hall; 1996.

194. Givens GH, Smith DD, Tweedie RL. Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Stat Sci* 1997;**12**:221–40.

195. Goldstein H, Spiegelhalter DJ. Statistical aspects of institutional performance: league tables and their limitations (with discussion). *J R Stat Soc A* 1996; **159**:385–444.

196. Goodman S, Langer A. Bayesian analysis and the GUSTO trial. *J Am Med Assoc* 1995;**274**:873–4.

197. Goodman SN, Zahurak ML, Piantadosi S. Some practical improvements in the continual

reassessment method for Phase I studies. *Stat Med* 1995;**14**:1149–61.

198. Gore SM. Biostatistics and the Medical Research Council. *Med Res Council News* 1987;**36**:19–20.

199. Gould AL. Using prior findings to augment active-controlled trials and trials with small placebo groups. *Drug Inform J* 1991;**25**(3):369–80.

200. Gould AL. Sample sizes for event rate equivalence trials using prior information. *Stat Med* 1993;**12**(21): 2009–23.

201. Gould AL. Planning and revising the sample size for a trial. *Stat Med* 1995;**14**:1039–51.

202. Gould AL. Multi-centre trial analysis revisited. *Stat Med* 1998;**17**:1779–1797.

203. Gray RJ. A Bayesian analysis of institutional effects in a multicenter cancer clinical trial. *Biometrics* 1994;**50**:244–53.

204. GREAT group. Feasibility, safety and efficacy of domiciliary thrombolysis by general practitioners: Grampian region early anisteplase trial. *BMJ* 1992; **305**:548–53.

205. Greenhouse JB. On some applications of Bayesian methods in cancer clinical trials. *Stat Med* 1992; **11**:37–53.

206. Greenhouse JB, Wasserman L. Robust Bayesian methods for monitoring clinical trials. *Stat Med* 1995;**14**:1379–91.

207. Greenland S, Robins JM. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology* 1991;**2**:244–51.

208. Grieve A. Some uses of predictive distributions in pharmaceutical research. In: Okuno T, editor. Biometry – Clinical trials and related topics. New York: Elsevier; 1988. p. 83–99.

209. Grieve A, Senn S. Estimating treatment effects in clinical cross-over trials. *J Biopharm Stat* 1998;**8**(2): 191–247.

210. Grieve AP. A Bayesian analysis of the 2-period cross-over design for clinical trials. *Biometrics* 1985; **41**:979–90.

211. Grieve AP. Evaluation of bioequivalence studies. *Eur J Clin Pharmacol* 1991;**40**:201–2.

212. Grieve AP. Bayesian analyses of two-treatment cross-over studies. *Stat Methods Med Res* 1994;**3**: 407–29.

213. Grieve AP. Bayesian approaches to randomised trials – discussion. *J R Stat Soc Ser A* 1994;**157**: 387–416.

214. Grieve AP. Extending a Bayesian analysis of the two-period cross-over to allow for baseline measurements. *Stat Med* 1994;**13**:905–29.

215. Grieve AP. Extending a Bayesian analysis of the two-period cross-over to accommodate missing data. *Biometrika* 1995;**82**:277–86.

216. Grieve AP. Issues for statisticians in pharmaco-economic evaluations. *Stat Med* 1998;**17**:1715–23.

217. Grossman J, Parmar MKB, Spiegelhalter DJ, Freedman LS. A unified method for monitoring and analysing controlled trials. *Stat Med* 1994; **13**:1815–26.

218. Gustafson P. A Bayesian analysis of bivariate survival data from a multicenter cancer clinical trial. *Stat Med* 1995;**14**:2523–35.

219. Gustafson P. Robustness considerations in Bayesian analysis. *Stat Methods Med Res* 1996;**4**:357–73.

220. Hall GH. Bayesian interim analysis of randomised trials. *Lancet* 1997;**349**:1910–11.

221. Halperin M, Lan KKG, Ware JH, Johnson NJ, DeMets DL. An aid to data monitoring in long-term clinical trials. *Controlled Clin Trials* 1982;**3**:311–23.

222. Hanauske AR, Edler L. New clinical trial designs for Phase I studies in hematology and oncology: principles and practice of the continual reassessment model. *Onkologie* 1996;**19**:404–9.

223. Healy M. New methodology in clinical trials. *Biometrics* 1978;**34**:709–12.

224. Healy MJR. Probability and decisions. *Arch Dis Child* 1994;**71**:90–4.

225. Hedges LV. Bayesian meta-analysis. In: Everitt BS, Dunn G, editors. Statistical analysis of medical data: new developments. London: Arnold; 1998. p. 251–76.

226. Heisterkamp SH, Doornbos G, Gankema M. Disease mapping using empirical Bayes and Bayes methods on mortality statistics in The Netherlands. *Stat Med* 1993;**12**:1895–913.

227. Heitjan DF. Bayesian interim analysis of Phase II cancer clinical trials. *Stat Med* 1997;**16**:1791–802.

228. Heitjan DF, Houts PS, Harvey HA. A decision-theoretic evaluation of early stopping rules. *Stat Med* 1992;**11**:673–83.

229. Heitjan DF, Moskowitz AJ, Whang W. Bayesian estimation of cost-effectiveness ratios from clinical trials. *Health Econ* 1999;**8**:191–201.

230. Henderson WG, Moritz T, Goldman S, Copeland J, Sethi G. Use of cumulative meta-analysis in the design, monitoring, and final analysis of a clinical trial: a case study. *Controlled Clin Trials* 1995;**16**(5): 331–41.

231. Herson J. Predictive probability early termination for Phase II clinical trials. *Biometrics* 1979;**35**:775–83.

232. Herson J. Bayesian analysis of cancer clinical trials – an introduction to 4 papers. *Stat Med* 1992; **11**:1–3.

233. Higgins JP, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Stat Med* 1996; **15**:2733–49.

234. Hilsenbeck SG. Early termination of a Phase II clinical trial. *Controlled Clin Trials* 1988;**9**(3):177–88.

235. Hlatky MA. Using databases to evaluate therapy. *Stat Med* 1991;**10**(4):647–52.

236. Ho CH. Some frequentist properties of a Bayesian method in clinical trials. *Biometr J* 1991;**33**:735–40.

237. Holland J. The Reverend Thomas Bayes, F. R. S (1702–61). *J R Stat Soc Ser A* 1962;**125**.

238. Hornberger J, Eghtesady P. The cost–benefit of a randomised trial to a health care organization. *Controlled Clin Trials* 1998;**19**:198–211.

239. Hornberger JC, Brown BW, Halpern J. Designing a cost-effective clinical trial. *Stat Med* 1995;**14**: 2249–59.

240. Howson C, Urbach P. Scientific reasoning: the Bayesian approach. La Salle (IL): Open Court; 1989.

241. Hsu PH, Laddu A, Jordan DC, Stoll RW, Deverka PA. Use of Bayesian drug causality assessment of adverse events in a United States pharmaceutical environment. *Clin Pharmacol Therapeutics* 1992; **51**:124.

242. Hughes MD. Practical reporting of Bayesian analyses of clinical trials. *Drug Information J* 1991;**25**:381–93.

243. Hughes MD. Reporting Bayesian analyses of clinical trials. *Stat Med* 1993;**12**:1651–63.

244. Hughes MD, Pocock SJ. Stopping rules and estimation problems in clinical trials. *Stat Med* 1988;**7**:1231–42.

245. Human Fertilisation and Embryology Authority. The patients' guide to DI and IVF clinics. 2nd ed. London: The Authority; 1996.

246. Hutton JL. The ethics of randomised controlled trials: a matter of statistical belief? *Health Care Analysis* 1996;**4**:95–102.

247. Hutton JL, Owens RG. Bayesian sample size calculations and prior beliefs about child sexual abuse. *Statistician* 1993;**42**:399–404.

248. International Conference on Harmonisation E9 Expert Working Group. Statistical principles for clinical trials: ICH harmonised tripartite guideline. *Stat Med* 1999;**18**:1905–42. URL: http://www.ich.org/ich5e.html

249. Jennison C. Discussion of Breslow (1990). *Stat Sci* 1990;**5**(3).

250. Jennison C, Turnbull BW. Statistical approaches to interim monitoring of medical trials: a review and commentary. *Stat Sci* 1990;**5**:299–317.

251. Jones B, Teather D, Wang J, Lewis JA. A comparison of various estimators of a treatment difference for a multi-centre clinical trial. *Stat Med* 1998;**17**:1767–77.

252. Jones DA. A Bayesian approach to economic evaluation of health care technologies. In: Spilker B, editor. Quality of life and pharmacoeconomics in clinical trials. New York: Raven Press; 1995.

253. Jones DA. The role of probability distributions in economic evaluations. *Br J Med Econ* 1995;**8**:137–46.

254. Jones DR. Meta-analysis: weighing the evidence. *Stat Med* 1995;**14**:137–49.

255. Joseph L, Duberger R, Belisle P. Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Stat Med* 1997;**16**:769–81.

256. Joseph L, Wolfson DB, Du Berger R. Sample size calculations for binomial proportions via highest posterior density intervals. *J R Stat Soc Ser D* 1995; **44**:143–54.

257. Joseph L, Wolfson DB, Du Berger R, Lyle RM. Change-point analysis of a randomised trial in the effects of calcium supplementation on blood pressure. In: Berry and Stangl.[61]

258. Kadane J. Bayesian methods and ethics in a clinical trial design. New York: Wiley; 1996.

259. Kadane J, Sedransk N. Towards a more ethical clinical trial. In: Bernardo J, DeGroot MH, Lindley DV, Smith AFM, editors. Bayesian statistics 1. Valencia: University Press; 1980. p. 329–38.

260. Kadane J, Wolfson L. Priors for the design and analysis of clinical trials. In: Berry and Stangl,[61] p. 157–84.

261. Kadane JB. Progress toward a more ethical method for clinical trials. *J Med Phil* 1986;**11**:385–404.

262. Kadane JB. Prime time for Bayes. *Controlled Clin Trials* 1995;**16**:313–18.

263. Kadane JB, Seidenfeld T. Randomization in a Bayesian perspective. *J Stat Planning Inference* 1990;**25**:329–45.

264. Kadane JB, Vlachos P, Wieand S. Decision analysis for a data monitoring committee of a clinical trial. In: Giron FJ, Martinez ML, editors. Proceedings of the International Workshop on Decision Analysis Applications. Boston: Kluwer, 1998. p. 115–21.

265. Kadane JB, Wolfson L. Experiences in elicitation. *Statistician* 1997;**46**:1–17.

266. Kass R, Raftery A. Bayes factors and model uncertainty. *J Am Stat Assoc* 1995;**90**:773–95.

267. Kass RE, Greenhouse JB. Comments on 'Investigating therapies of potentially great benefit: ECMO' by J H Ware. *Stat Sci* 1989;**4**:310–17.

268. Kass RE, Wasserman L. The selection of prior distributions by formal rules. *J Am Stat Assoc* 1996;**91**:1343–70.

269. Keiding N. Bayesian approaches to randomised trials – discussion. *J R Stat Soc Ser A* 1994;**157**: 387–416.

270. Kleinman KP, Ibrahim JG, Laird NM. A Bayesian framework for intent-to-treat analysis with missing data. *Biometrics* 1998;**54**:265–78.

271. Kober L, Torppedersen C, Cole D, Hampton JR, Camm AJ. Bayesian interim statistical analysis of randomised trials: the case against. *Lancet* 1997; **349**:1168–9.

272. Koch GG. Summary and discussion for 'Statistical issues in the pharmaceutical industry: analysis and reporting of Phase III clinical trials including kinetic/dynamic analysis and Bayesian analysis'. *Drug Information J* 1991;**25**:433–7.

273. Korn EL. Projection from previous studies. a caution. *Controlled Clin Trials* 1990;**11**:67–9.

274. Korn EL, Midthune D, Chen TT, Rubinstein LV, Christian MC, Simon RM. A comparison of two Phase I trial designs. *Stat Med* 1994;**13**:1799–806.

275. Korn EL, Simon R. Data monitoring committees and problems of lower than expected accrual or event rates. *Controlled Clin Trials* 1996;**17**:526–35.

276. Korn EL, Yu KF, Miller LL. Stopping a clinical trial very early because of toxicity – summarizing the evidence. *Controlled Clin Trials* 1993;**14**:286–95.

277. Lachin JM. Sequential clinical-trials for normal variates using interval composite hypotheses. *Biometrics* 1981;**37**:87–101.

278. Lan KKG, Wittes J. The *B*-value – a tool for monitoring data. *Biometrics* 1988;**44**:579–85.

279. Lanctot KL, Kwok MCO, Busto U, Naranjo CA. Bayesian evaluation of spontaneous reports of pancreatitis. *Clin Pharmacol Therapeutics* 1996;**59**: PII 4.

280. Lane N. Common sense, nonsense and statistics. *J R Soc Med* 1999;**92**:202–5.

281. Lang T, Secic M. Considering 'prior probabilities': reporting Bayesian statistical analyses. In: How to report statistics in medicine. American College of Physicians; 1997. p. 231–5.

282. Lange N, Carlin BP, Gelfand AE. Hierarchical Bayes models for the progression of HIV-infection using longitudinal CD4 T-cell numbers. *J Am Stat Assoc* 1992;**87**:615–26.

283. Larose DT, Dey DK. Grouped random effects models for Bayesian meta-analysis. *Stat Med* 1997; **16**:1817–29.

284. Lau J, Schmid CH, Chalmers TC. Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. *J Clin Epidemiol* 1995; **48**:45–57.

285. Lavori PW, Dawson R, Shera D. A multiple imputation strategy for clinical trials with truncation of patient data. *Stat Med* 1995;**14**: 1913–25.

286. Lecoutre B, Derzko G, Grouin JM. Bayesian predictive approach for inference about proportions. *Stat Med* 1995;**14**:1057–63.

287. Lee PM. Bayesian statistics: an introduction. London: Edward Arnold; 1989.

288. Legler JM, Ryan LM. Latent variable models for teratogenesis using multiple binary outcomes. *J Am Stat Assoc* 1997;**92**:13–20.

289. Lehmann HP, Nguyen B. Bayesian communication of research results over the world wide web. *M D Comput* 1997;**14**:353–9.

290. Lehmann HP, Shachter RD. A physician-based architecture for the construction and use of statistical models. *Methods Inform Med* 1994;**33**:423–32.

291. Lehmann HP, Wachter MR. Implementing the Bayesian paradigm: reporting research results over the world-wide web. In: Proceedings of the AMIA Annual Fall Symposium; 1996. p. 433–7.

292. Lewis J. Bayesian approaches to randomised trials – discussion. *J R Stat Soc Ser A* 1994;**157**:387–416.

293. Lewis R. Bayesian hypothesis testing: interim analysis of a clinical trial evaluating phenytoin for the prophylaxis of early post-traumatic seizures in children. In: Berry and Stangl,[61] p. 279–96.

294. Lewis RJ, Berry DA. Group sequential clinical trials – a classical evaluation of Bayesian decision-theoretic designs. *J Am Stat Assoc* 1994;**89**:1528–34.

295. Lewis RJ, Wears RL. An introduction to the Bayesian analysis of clinical trials. *Ann Emerg Med* 1993;**22**:1328–36.

296. Li Z, Begg CB. Random effects models for combining results from controlled and uncontrolled studies in a meta-analysis. *J Am Stat Assoc* 1994;**89**:1523–7.

297. Lilford R, Gudmundsson S, James D, Mason G, Neales K, Pearce M, *et al.* Formal measurement of clinical uncertainty: prelude to a trial in perinatal medicine. *BMJ* 1994;**308**:111–12.

298. Lilford R, Royston G. Decision analysis in the selection, design and application of clinical and health services research. *J Health Serv Res Policy* 1998;**3**:159–66.

299. Lilford RJ, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. *BMJ* 1996;**313**:603–7.

300. Lilford RJ, Jackson J. Equipoise and the ethics of randomization. *J R Soc Med* 1995;**88**:552–9.

301. Lilford RJ, Pauker SG, Braunholtz DA, Chard J. Getting research findings into practice: Decision analysis and the implementation of research findings. *BMJ* 1998;**317**(7155):405–9.

302. Lilford RJ, Thornton JG, Braunholtz D. Clinical trials and rare diseases – a way out of a conundrum. *BMJ* 1995;**311**:1621–5.

303. Lindley D. The effect of ethical design considerations on statistical analysis. *Appl Stat* 1975;**24**:218–28.

304. Lindley D. Bayesian approaches to randomised trials – discussion. *J R Stat Soc Ser A* 1994;**157**: 387–416.

305. Lindley D. The choice of sample size. *Statistician* 1997;**46**:129–38.

306. Lindley DV. Scoring rules and the inevitability of probability. *Int Stat Rev* 1982;**50**:1–26.

307. Lindley DV. Making decisions. 2nd ed. Chichester: Wiley; 1985.

308. Lindley DV. Decision analysis and bioequivalence trials. *Stat Sci* 1998;**13**:136–41.

309. Louis TA. Using empirical Bayes methods in biopharmaceutical research. *Stat Med* 1991;**10**: 811–29.

310. Luce BR, Claxton K. Redefining the analytical approach to pharmacoeconomics. *Health Econ* 1999;**8**:187–9.

311. Manning WG, Fryback FG, Weinstein MC. Reflecting uncertainty in cost-effectiveness analysis. In: Gold MR, Siegel JR, Weinstein MC, Russell LB, editors. Cost effectiveness in health and medicine. New York: Oxford University Press; 1996. p. 247–75.

312. Marshall EC, Spiegelhalter DJ. League tables of *in vitro* fertilisation clinics: how confident can we be about the rankings? *BMJ* 1998;**317**:1701–4.

313. Marshall RJ. Bayesian analysis of case–control studies. *Stat Med* 1988;**7**:1223–30.

314. Martinez ML. Bioequivalence assessment. population and individual bioequivalence. *An Real Acad Farm Inst Espana* 1997;**63**(3):493–531.

315. Matchar DB, Samsa GP, Matthews JR, Ancukiewicz M, Parmigiani G, Hasselblad V, *et al.* The stroke prevention policy model: Linking evidence and clinical decisions. *Ann Intern Med* 1997;**127**:704–11.

316. Matsuyama Y, Sakamoto J, Ohashi Y. A Bayesian hierarchical survival model for the institutional effects in a multi-centre cancer clinical trial. *Stat Med* 1998;**17**(17):1893–908.

317. Matthews JNS. Small clinical trials – are they all bad? *Stat Med* 1995;**14**:115–26.

318. Matthews R. Faith, hope and statistics. *New Sci* 1997;**156**:36.

319. Matthews R. Fact versus factions: the use and abuse of subjectivity in scientific research. Cambridge: The European Science and Environment Forum; 1998. Technical report.

320. McIntosh MW. The population risk as an explanatory variable in research synthesis of clinical trials. *Stat Med* 1996;**15**(16):1713–28.

321. Mcpherson K. On choosing the number of interim analyses in clinical trials. *Stat Med* 1982;**1**:25–36.

322. Mehta CR, Cain KC. Charts for the early stopping of pilot studies. *J Clin Oncol* 1984;**2**:676–82.

323. Meier P. Statistics and medical experimentation. *Biometrics* 1975;**31**:511–29.

324. Metzler CM. Sample sizes for bioequivalence studies. *Stat Med* 1991;**10**:961–70.

325. Miller MA, Seaman JW. A Bayesian approach to assessing the superiority of a dose combination. *Biometr J* 1998;**40**:43–55.

326. Moller S. An extension of the continual reassessment methods using a preliminary up-and-down design in a dose finding study in cancer patients, in order to investigate a greater range of doses. *Stat Med* 1995;**14**:911–22. Discussion: 923.

327. Morris CN, Normand SL. Hierarchical models for combining information and for meta-analysis. In: Bernardo *et al.*,[41] 321–44.

328. Moussa MAA. Exact, conditional and predictive power in planning clinical trials. *Controlled Clin Trials* 1989;**10**:378–85.

329. Muliere P, Walker S. A Bayesian nonparametric approach to determining a maximum tolerated dose. *J Stat Planning Inference* 1997;**61**:339–53.

330. Müller P, Parmigiani G, Schlidkraut J, Tardella L. A Bayesian hierarchical approach for combining case–control and prospective studies. *Biometrics* 1999;**55**:858–66.

331. Murphy A, Winkler R. Reliability of subjective probability forecasts of precipitation and temperature. *Appl Stat* 1977;**26**:41–7.

332. Nicolas P, Tod M, Petitjean O. Review and use of decision rules for bioequivalence studies. *Therapie* 1993;**48**:15–22.

333. Normand SL, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc* 1997;**92**:803–14.

334. Normand ST. Meta-analysis: Formulating, evaluating, combining and reporting. *Stat Med* 1999;**18**:321–59.

335. Nurminen M, Mutanen P. Bayesian analysis of case–control studies. *Stat Med* 1989;**8**:1023–4.

336. O'Brien PC. Data and safety monitoring. In: Armitage and Colton,[20] p. 1058–66.

337. Office GA. Cross design synthesis: a new strategy for medical effectiveness research. Washington, DC.: General Accounting Office, 1992.

338. O'Hagan A. Kendall's advanced theory of statistics. Vol 2B. Bayesian inference. London: Arnold; 1994.

339. O'Neill RT. Conclusions: 2. *Stat Med* 1994;**13**: 1493–9.

340. O'Quigley J. Estimating the probability of toxicity at the recommended dose following a Phase I clinical trial in cancer. *Biometrics* 1992;**48**:853–62.

341. O'Quigley J, Chevret S. Methods for dose finding studies in cancer clinical-trials – a review and results of a Monte Carlo study. *Stat Med* 1991;**10**:1647–64.

342. O'Quigley J, Pepe M, Fisher L. Continual reassessment method – a practical design for Phase I clinical trials in cancer. *Biometrics* 1990;**46**:33–48.

343. O'Rourke K. Two cheers for Bayes. *Controlled Clin Trials* 1996;**17**:350–2.

344. Palmer CR. A comparative phase-2 clinical-trials procedure for choosing the best of 3 treatments. *Stat Med* 1991;**10**:1327–40.

345. Palmer CR. Ethics and statistical methodology in clinical trials. *J Med Ethics* 1993;**19**:219–22.

346. Palmer CR, Rosenberger WF. Ethics and practice: alternative designs for Phase III randomised clinical trials. *Controlled Clin Trials* 1999;**20**:172–86.

347. Papineau D. The virtues of randomization. *Br J Phil Sci* 1994;**45**:437–50.

348. Parmar MKB, Spiegelhalter DJ, Freedman LS. The CHART trials: Bayesian design and monitoring in practice. *Stat Med* 1994;**13**:1297–312.

349. Parmar MKB, Ungerleider RS, Simon R. Assessing whether to perform a confirmatory randomised clinical trial. *J Natl Cancer Inst* 1996;**88**:1645–51.

350. Parmigiani G. Decision models in screening for breast cancer. In: Bernardo J, Berger J, Dawid A, Smith A, editors. Bayesian statistics 6. Oxford: Oxford University Press; 1999. p. 525–46.

351. Parmigiani G, Anckiewicz M, Matchar D. Decision models in clinical recommendations development: The stroke prevention policy model. In: Berry and Stangl,[61] p. 207–233.

352. Parmigiani G, Kamlet M. A cost-utility analysis of alternative strategies in screening for breast cancer. In: Gatsonis C, Hodges J, Kass R, Singpurwalla N, editors. Case studies in Bayesian statistics. Berlin: Springer-Verlag; 1993. p. 390–402.

353. Parmigiani G, Samsa GP, Ancukiewicz M, Lipscomb J, Hasselblad V, Matchar DB. Assessing uncertainty in cost-effectiveness analyses: application to a complex decision model. *Med Decision Making* 1997;**17**:390–401.

354. Pepple PA, Choi SC. Analysis of incomplete data under nonrandom mechanisms – Bayesian inference. *Comm Stat Simulation Comput* 1994;**23**: 743–67.

355. Pepple PA, Choi SC. Bayesian approach to two-stage Phase II trial. *J Biopharm Stat* 1997;**7**(2): 271–86.

356. Perneger T. What's wrong with Bonferroni adjustments. *BMJ* 1998;**316**:1236–8.

357. Peto R. "discussion of 'on the allocation of treatments in sequential medical trials' by J Bather". *Int Stat Rev* 1985;**53**:1–13.

358. Peto R, Baigent C. Trials: the next 50 years. *BMJ* 1998;**317**:1170–1.

359. Peto R, Collins R, Gray R. Large-scale randomised evidence – large, simple trials and overviews of trials. *J Clin Epidemiol* 1995;**48**:23–40.

360. Pocock S. The combination of randomised and historical controls in clinical trials. *J Chronic Dis* 1976;**29**:175–88.

361. Pocock S. Clinical trials: a practical approach. Chichester: Wiley; 1983.

362. Pocock S. Statistical and ethical issues in monitoring clinical-trials. *BMJ* 1992;**305**:235–40.

363. Pocock S, Spiegelhalter D. Domiciliary thrombolysis by general practitioners. *BMJ* 1992;**305**:1015.

364. Pocock SJ, Hughes MD. Practical problems in interim analyses, with particular regard to estimation. *Controlled Clin Trials* 1989;**10**:209S–21S.

365. Pocock SJ, Hughes MD. Estimation issues in clinical trials and overviews. *Stat Med* 1990;**9**:657–71.

366. Pocock SP. Bayesian approaches to randomised trials – discussion. *J R Stat Soc Ser A* 1994;**157**: 387–416.

367. Pogue JM, Yusuf S. Cumulating evidence from randomised trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. *Controlled Clin Trials* 1997;**18**(6):580–93.

368. Qian J, Stangl D, George S. A Weibull model for survival data: using prediction to decide when to stop a clinical trial. In: Berry and Stangl.[61]

369. Raab GM. Conflict between prior and current data. *Appl Stat* 1996;**45**(2):247–51.

370. Racine A, Grieve AP, Fluhler H, Smith AFM. Bayesian methods in practice – experiences in the pharmaceutical industry. *Appl Stat* 1986;**35**:93–150.

371. Racine-Poon A, Grieve AP, Fluhler H, Smith AFM. A 2-stage procedure for bioequivalence studies. *Biometrics* 1987;**43**:847–56.

372. Racine-Poon A, Wakefield J. Bayesian analysis of population pharmacokinetic and instantaneous pharmacodynamic relationships. In: Berry and Stangl,[61] p. 321–54.

373. Raghunathan TE, Siscovick DS. A multiple-imputation analysis of a case–control study of the risk of primary cardiac-arrest among pharmacologically treated hypertensives. *Appl Stat J R Stat Soc Ser C* 1996;**45**:335–52.

374. Raudenbush SW, Bryk AS. Empirical Bayes meta-analysis. *J Educ Stat* 1985;**10**:75–98.

375. Richards B, Blandy J, Bloom HGJ, MRC Bladder Cancer Working Party. The effect of intravesical thiotepa on tumor recurrence after endoscopic treatment of newly-diagnosed superficial bladder-cancer – a further report with long-term follow-up of a Medical-Research-Council randomised trial. *Br J Urol* 1994;**73**:632–8.

376. Richardson S, Gilks WR. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *Am J Epidemiol* 1993;**138**:430–42.

377. Richardson S, Monfort C, Green M, Draper G, Muirhead C. Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain. *Stat Med* 1995;**14**:2487–501.

378. Rittenhouse BE. Exorcising protocol-induced spirits: making the clinical trial relevant for economics. *Med Decision Making* 1997;**17**:331–9.

379. Rogatko A. Bayesian approach for meta-analysis of controlled clinical-trials. *Commun Stat Theory Methods* 1992;**21**:1441–62.

380. Rosenbaum PR, Rubin DB. Sensitivity of Bayes inference with data-dependent stopping rules. *Am Stat* 1984;**38**:106–9.

381. Rosner GL, Berry DA. A Bayesian group sequential design for a multiple arm randomised clinical trial. *Stat Med* 1995;**14**:381–94.

382. Rothman K. No adjustments are needed for multiple comparisons. *Epidemiology* 1990;**1**:43–6.

383. Rubin D. Bayesian inference for casual effects: the role of randomization. *Ann Stat* 1978;**7**:34–58.

384. Rubin DB. A new perspective. In: Rubin D, Wachter K, Straf M, editors. The future of meta-analysis. New York: Russell Sage Foundation; 1992. p. 155–65.

385. Rubin DB. More powerful randomization-based *p*-values in double-blind trials with non-compliance. *Stat Med* 1998;**17**:371–85; Discussion: 387–9.

386. Ryan L. Using historical controls in the analysis of developmental toxicity data. *Biometrics* 1993;**49**: 1126–35.

387. Samsa GP, Reutter RA, Parmigiani G, Ancukiewicz M, Abrahamse P, Lipscomb J, *et al.* Performing cost-effectiveness analysis by integrating randomised trial data with a comprehensive decision model: application to treatment of acute ischemic stroke. *J Clin Epidemiol* 1999;**52**:259–71.

388. Sargent D, Carlin B. Robust Bayesian design and analysis of clinical trials via prior partitioning (with discussion). In: Berger JO, editor. Bayesian robustness: IMS lecture notes – monograph series, 29. Hayward (CA): Institute of Mathematical Statistics; 1996. p. 175–193.

389. Sasahara A, Cole T, Ederer F, Murray J, Wenger N, Sherry S, *et al.* Urokinase pulmonary embolism trial, a national cooperative study. *Circulation* 1973;**47** (suppl 2):1–108.

390. Saunders M, Dische S, Barrett A, Harvey A, Gibson D, Parmar M, *et al.* Continuous hyper-fractionated accelerated radiotherapy CHART versus con ventional radiotherapy in non-small-cell lung cancer: a randomised multicentre trial. *Lancet* 1997;**350**:161–5.

391. Savage L. Elicitation of personal probabilities and expectations. *J Am Stat Assoc* 1971;**66**:783–801.

392. Schervish MJ. p values: what they are and what they are not. *Am Stat* 1996;**50**:203–6.

393. Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Stat Med* 1998;**17**(17):1923–42.

394. Selwyn MR, Dempster AP, Hall NR. A Bayesian approach to bioequivalence for the 2 × 2 changeover design. *Biometrics* 1981;**37**:11–21.

395. Selwyn MR, Hall NR. On Bayesian methods for bioequivalence. *Biometrics* 1984;**40**:1103–8.

396. Senn S. Some statistical issues in project prioritization in the pharmaceutical industry. *Stat Med* 1996;**15**:2689–702.

397. Senn S. Statistical basis of public policy – present remembrance of priors past is not the same as a true prior. *BMJ* 1997;**314**:73.

398. Senn S. Statistical issues in drug development. Chichester: Wiley; 1997.

399. Shachter R, Eddy DM, Hasselblad V. An influence diagram approach to medical technology assessment. In: Oliver RM, Smith JQ, editors. Influence diagrams, belief nets and decision analysis. Chichester: Wiley; 1990. p. 321–350.

400. Sheiner LB. The intellectual health of clinical drug-evaluation. *Clin Pharmacol Therapeutics* 1991;**50**:4–9.

401. Simes RJ. Application of statistical decision theory to treatment choices: implications for the design and analysis of clinical trials. *Stat Med* 1986;**4**: 1401–9.

402. Simon R. Adaptive treatment assignment methods and clinical trials. *Biometrics* 1977;**33**:743–9.

403. Simon R. Statistical tools for subset analysis in clinical trials. *Recent Results Cancer Res* 1988;**111**: 55–66.

404. Simon R. Problems of multiplicity in clinical trials. *J Stat Planning Inference* 1994;**42**:209–21.

405. Simon R. Some practical aspects of the interim monitoring of clinical trials. *Stat Med* 1994;**13**: 1401–9.

406. Simon R, Dixon DO, Friedlin B. Bayesian subset analysis of a clinical trial for the treatment of HIV infections. In: Berry and Stangl,[61] p. 555–76.

407. Simon R, Freedman LS. Bayesian design and analysis of two × two factorial clinical trials. *Biometrics* 1997;**53**:456–64.

408. Simon R, Thall PF, Ellenberg SS. New designs for the selection of treatments to be tested in randomised clinical trials. *Stat Med* 1994;**13**:417–29.

409. Skene AM, Wakefield JC. Hierarchial models for multicentre binary response studies. *Stat Med* 1990;**9**:919–29.

410. Smith T, Spiegelhalter DJ, Parmar MKB. Bayesian meta-analysis of randomised trials using graphical models and BUGS. In: Berry and Stangl,[61] p. 411–27.

411. Smith TC, Abrams KR, Jones DR. Using hierarchical models in generalised synthesis of evidence: an example based on studies of breast cancer screening. Leicester: Department of Epidemiology and Public Health, University of Leicester; 1995. Technical report 95-02.

412. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: a comparative study. *Stat Med* 1995;**14**:2685–99.

413. Souhami RL, Craft AW, Van der Eijken, Nooij M, Spooner D, Bramwell VHC, *et al.* Randomised trial of two regimens of chemotherapy in operable osteosarcoma: a study of the European osteo-sarcoma intergroup. *Lancet* 1997;**350**:911–17.

414. Spiegelhalter D, Freedman L. Bayesian approaches to clinical trials. In: Bernardo *et al.*,[42] p. 453–77.

415. Spiegelhalter D, Myles J, Jones D, Abrams K. An introduction to Bayesian methods in health technology assessment. *BMJ* 1999;**319**:508–12.

416. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Stat Med* 1986;**5**: 421–33.

417. Spiegelhalter DJ. Bayesian graphical modelling: a case-study in monitoring health outcomes. *Appl Stat J R Stat Soc Ser C* 1998;**47**:115–33.

418. Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Stat Med* 1986;**5**:1–13.

419. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: Conditional of predictive power? *Controlled Clin Trials* 1986;**7**:8–17.

420. Spiegelhalter DJ, Freedman LS, Parmar MKB. Applying Bayesian ideas in drug development and clinical trials. *Stat Med* 1993;**12**:1501–17.

421. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomised trials (with discussion). *J R Stat Soc Ser A* 1994;**157**:357–87.

422. Spiegelhalter DJ, Harris NL, Bull K, Franklin RCG. Empirical evaluation of prior beliefs about frequencies: methodology and a case study in congenital heart disease. *J Am Stat Assoc* 1994;**89**:435–43.

423. Spiegelhalter DJ, Thomas A, Best NG, Gilks WR. BUGS: Bayesian inference using Gibbs sampling, version 0.5 (version ii). Cambridge: MRC Biostatistics Unit; 1996.

424. Stallard N. Sample size determination for Phase II clinical trials based on Bayesian decision theory. *Biometrics* 1998;**54**:279–94.

425. Stangl D. Hierarchical analysis of continuous-time survival models. In: Berry and Stangl,[61] p. 429–50.

426. Stangl D, Berry DA. Bayesian statistics in medicine: where we are and where we should be going. *Sankhya Ser B* 1998;**60**:176–95.

427. Stangl DK. Prediction and decision-making using Bayesian hierarchical-models. *Stat Med* 1995;**14**: 2173–90.

428. Stangl DK, Greenhouse JB. Assessing placebo response using Bayesian hierarchical survival models. *Lifetime Data Anal* 1998;**4**:5–28.

429. Staquet MJ, Rozencweig M, Von Hoff DD, Muggia FM. The delta and epsilon errors in the assessment of cancer clinical trials. *Cancer Treatment Rep* 1979; **63**:1917–21.

430. Staquet MJ, Sylvster RJ. A decision theory approach to Phase II clinical trials. *Biomedicine* 1977;**26**:262–6.

431. Stijnen T, Van Houwelingen JC. Empirical Bayes methods in clinical trials meta-analysis. *Biometric J* 1990;**32**:335–46.

432. Strauss N, Simon R. Investigating a sequence of randomised Phase II trials to discover promising treatments. *Stat Med* 1995;**14**(13):1479–89.

433. Su XY, Po ALW. Combining event rates from clinical trials: comparison of Bayesian and classical methods. *Ann Pharmacotherapy* 1996;**30**:460–5.

434. Sutton A, Abrams KR, Jones DR, Sheldon TA, Song F. Systematic reviews of trials and other studies. *HTA* 1998;**2**(19).

435. Sylvester RJ. A Bayesian approach to sample size determination in Phase II cancer clinical trials. *Controlled Clin Trials* 1984;**5**:305.

436. Sylvester RJ. A Bayesian approach to the design of Phase II clinical trials. *Biometrics* 1988;**44**:823–36.

437. Sylvester RJ, Stquet M. Design of Phase II trials in cancer using decision theory. *Cancer Treatment Rep* 1988;**64**:519–24.

438. Tamura RN, Faries DE, Andersen JS, Heiligenstein JH. A case-study of an adaptive clinical trial in the treatment of out-patients with depressive disorder. *J Am Stat Assoc* 1994;**89**:768–76.

439. Tan SB, Smith AFM. Exploratory thoughts on clinical trials with utilities. *Stat Med* 1998;**17**: 2771–91.

440. Tarone R. The use of historical control information in testing for a trend in proportions. *Biometrics* 1982;**38**:215–20.

441. Templeton A, Morris JK, Parslow W. Factors that affect outcome on *in vitro* fertilisation treatment. *Lancet* 1996;**348**:1402–6.

442. Ten Centre Study Group. Ten centre study of artificial surfactant (artificial lung expanding compound) in very premature babies. *BMJ* 1987;**294**:991–6.

443. Teo KK, Yusuf S, Collins R, Held PH, Peto R. Effects of intravenous magnesium in suspected acute myocardial infarction: Overview of randomised trials. *BMJ* 1991;**303**:1499–503.

444. Thall PF, Estey EH. A Bayesian strategy for screening cancer treatments prior to Phase II clinical evaluation. *Stat Med* 1993;**12**:1197–211.

445. Thall PF, Lee JJ, Tseng CH, Estey EH. Accrual strategies for Phase I trials with delayed patient outcome. *Stat Med* 1999;**18**:1155–69.

446. Thall PF, Russell KE. A strategy for dose-finding and safety monitoring based on efficacy and adverse outcomes in Phase I/II clinical trials. *Biometrics* 1998;**54**:251–64.

447. Thall PF, Simon R. A Bayesian approach to establishing sample size and monitoring criteria for Phase II clinical trials. *Controlled Clin Trials* 1994; **15**:463–81.

448. Thall PF, Simon R. Practical Bayesian guidelines for Phase IIb clinical trials. *Biometrics* 1994;**50**: 337–49.

449. Thall PF, Simon RM, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Stat Med* 1995;**14**:357–79.

450. Thall PF, Simon RM, Estey EH. New statistical strategy for monitoring safety and efficacy in single-arm clinical trials. *J Clin Oncol* 1996;**14**:296–303.

451. Thall PF, Sung H. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Stat Med* 1998; **17**(14):1563–80.

452. Thompson M. Decision-analytic determination of study size. *Med Decision Making* 1981;**1**:165–79.

453. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. *Stat Med* 1997;**16**(23):2741–58.

454. Tukey J. Some thoughts on clinical trials, especially problems of multiplicity. *Science* 1977;**198**:679–84.

455. Tunis SR, Sheinhait IA, Schmid CH, Bishop DJ, Ross SD. Lansoprazole compared with histamine(2)- receptor antagonists in healing gastric ulcers: a meta-analysis. *Clin Therapeutics* 1997;**19**:743–57.

456. Tversky A. Assessing uncertainty (with discussion). *J R Stat Soc B* 1974;**36**:148–59.

457. Tweedie RL, Scott DJ, Biggerstaff BJ, Mengersen KL. Bayesian meta-analysis, with application to studies of ETS and lung-cancer. *Lung Cancer* 1996;**14**: S 171–S 194.

458. University Group Diabetes Program. A study of the effects of hypoglycemic agents on vascular complications in patients with adult onset diabetes. *Diabetes* 1970;**19** (suppl 2):747–830.

459. Urbach P. The value of randomization and control in clinical trials. *Stat Med* 1993;**12**:1421–31.

460. US Food and Drug Administration. Semiannual guidance agenda. *Federal Register* 1998;**63**(212): 59317–26.

461. US Food and Drug Administration. Transcript of Cardiovascular and Renal Drugs Advisory Committee meeting, 26 June 1997. URL: http://www.fda.gov/ohrms/dockets/ac/97/transcpt/3320t1.pdf; 1998.

462. US Food and Drug Administration. Guidance for industry: population pharmacokinetics. URL: http://www.fda.gov.cder/guidance/index.htm; 1999.

463. US Food and Drug Administration. Summary of Safety and effectiveness data for T-scan breast scanner. URL: http://www.fda.gov/cdrh/pdf/p970033b.pdf; 1999.

464. Vanhouwelingen HC. The future of biostatistics: expecting the unexpected. *Stat Med* 1997;**16**: 2773–84.

465. Vanhouwelingen HC, Zwinderman K, Stijnen T. A bivariate approach to meta-analysis. *Stat Med* 1993; **12**:2272–84.

466. Wakefield J, Bennett J. The Bayesian modeling of covariates for population pharmacokinetic models. *J Am Stat Assoc* 1996;**91**:917–27.

467. Wakefield J, Walker S. A population approach to initial dose selection. *Stat Med* 1997;**16**:1135–49.

468. Waller SE, Duncan D. A Bayes rule for the symmetric multiple comparison problem. *J Am Stat Assoc* 1969;**64**:1484–508.

469. Ware J. Investigating therapies of potentially great benefit: ECMO (with discussion). *Stat Sci* 1989;**4**:298–340.

470. Ware JH, Muller JE, Braunwald E. The futility index: an approach to the cost-effective termination of randomised clinical trials. *Am J Med* 1985;**78**: 635–43.

471. Weinstein M, Fineberg H. Clinical decision analysis. Philadelphia: Saunders; 1980.

472. Westfall PH, Johnson WO, Utts JM. A Bayesian perspective on the Bonferroni adjustment. *Biometrika* 1997;**84**:419–27.

473. Whitehead J. Designing Phase II studies in the context of a program of clinical research. *Biometrics* 1985;**41**:373–83.

474. Whitehead J. Sample sizes for Phase-II and Phase-III clinical trials – an integrated approach. *Stat Med* 1986;**5**:459–64.

475. Whitehead J. Letter to the editor. *Controlled Clin Trials* 1991;**12**:340–4.

476. Whitehead J. The case for frequentism in clinical trials. *Stat Med* 1993;**12**:1405–19.

477. Whitehead J. Bayesian decision procedures with application to dose-finding studies. *Int J Pharm Med* 1997;**11**(4):201–7.

478. Whitehead J. The design and analysis of sequential clinical trials. 2nd ed. Chichester: Wiley; 1997.

479. Whitehead J, Brunier H. Bayesian decision procedures for dose determining experiments. *Stat Med* 1995;**14**:885–93.

480. Woods KL. Mega-trials and management of acute myocardial-infarction. *Lancet* 1995;**346**:611–14.

481. Woods KL, Fletcher S, Roffe C, Haider Y. Intravenous magnesium sulphate in suspected acute myocardial infarction: results of the second Leicester intravenous magnesium intervention trial (LIMIT-2). *Lancet* 1992;**339**(8809):1553–8.

482. Yao TJ, Begg CB, Livingston PO. Optimal sample size for a series of pilot trials of new agents. *Biometrics* 1996;**52**:992–1001.

483. Yusuf S, Flather M. Magnesium in acute myocardial-infarction. *BMJ* 1995;**310**:751–2.

484. Yusuf S, Teo K, Woods K. Intravenous magnesium in acute myocardial infarction: an effective, safe, simple and inexpensive treatment. *Circulation* 1993;**87**:2043–6.

485. Zelen M. Play the winner rule and the controlled clinical trial. *J Am Stat Assoc* 1969;**64**:131–46.

486. Zelen M. Discussion of Breslow (1990). *Stat Sci* 1990;**5**(3).

487. Zelen M, Parker RA. Case control studies and Bayesian inference. *Stat Med* 1986;**5**:261–9.

488. Zucker DR, Schmid CH, McIntosh MW, Agostino RB, Selker HP, Lau J. Combining single patient (*n*-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. *J Clin Epidemiol* 1997;**50**:401–10.

Appendix 1

Three-star applications

TABLE 24 *Summary of three-star applications with respect to BayesWatch criteria*

Author(s)	Year	Design	Model	Prospective study	Elicited prior from experts	Loss function used	Computations	Sensitivity
Abrams *et al.*[1 *]	1994	RCT	B	✗	✓	✓	C	✓
Abrams *et al.*[2]	1996	RCT	S	✗	✗	✗	L	✓
Berry[51]	1989	RCT	B	✗	✗	✗	NI	✗
Berry *et al.*[64]	1994	RCT	P	✗	✗	✓✓	D	✓✓
Brophy and Joseph[76]	1995	RCT	NA	✗	✗	✗	C	✓✓
Brophy and Joseph[77]	1997	RCT	NA	✗	✗	✗	C	✗
Carlin *et al*[90]	1993	RCT	S	✗	✓	✓	NA/NI	✓✓
DerSimonian[129]	1996	RCT	NA	✓	✗	✗	C	✓✓
Digman *et al.*[133]	1998	RCT	NA	✓	✗	✗	C	✓✓
Fayers *et al.*[166]	1997	RCT	NA	✓	✗	✓	C	✓✓
Fletcher *et al.*[173]	1993	RCT	NA	✗	✗	✓	CNR	NR
Freedman and Spiegelhalter[182]	1992	RCT	NA	✗	✗	✓	C	✓✓
Freedman *et al.*[183]	1994	RCT	NA	✗	✗	✓	C	✓✓
George *et al.*[191]	1994	RCT	S	✗	✗	✗	MCMC	✓✓
Greenhouse and Wasserman[206]	1995	RCT	B	✗	✗	✗	NI	✓✓
Gustafson[219]	1996	RCT	B	✗	✗	✗	NES	✓✓
Hughes[242]	1991	RCT	NA	✓	✓	✗	C	✓✓
Kadane[258]	1996	ETH	N	✓	✓	✗	C	✓✓
Kass and Greenhouse[267]	1989	RCT	B	✗	✗	✗	NES	✓✓
Lewis[293]	1996	RCT	B	✓	✗	✓✓	C	✓✓
Lilford and Braunholtz[299]	1996	M-A	NA	✗	✓	✗	C	✓✓
Parmar *et al.*[348]	1994	RCT	NA	✓	✓	✓	C	✓✓
Parmar *et al.*[349]	1996	RCT	NA	✗	✗	✗	C	✓✓
Pocock and Hughes[364]	1989	RCT	NA	✗	✗	✗	C	✓✓
Pocock and Spiegelhalter[363]	1992	RCT	NA	✗	✓	✗	C	✗
Sasahara *et al.*[389]	1973	RCT	NA	✓	✗	✗	C	✗
Spiegelhalter *et al.*[420]	1993	RCT	NA	✓	✓	✓	C	✓✓
Spiegelhalter *et al.*[421]	1994	RCT	NA	✗	✗	✓	C	✓✓
Stangl[427]	1995	RCT	S	✗	✗	✗	MCMC	✗
Ware[469]	1989	RCT	B	✗	✗	✗	A	✗

Key: design (RCT, randomised controlled trial; ETH, 'ethical' study; M-A, meta-analysis); model (B, binomial; NA, normal approximation; N, normal; P, Poisson; S, Survival); prospective analysis (✓, yes; ✗, no); elicited prior from experts (✓, yes; ✗, no); loss function used (✓✓, yes; ✓, just demands; ✗, no); computations (A, analytical/closed-form; C, conjugate; D, dynamic programming; L, Laplace; NA, normal approximations; NI, numerical integration; NES, not explicitly stated); sensitivity analysis (✓✓, full; ✓, reference only; ✗, no; NR, none reported)

[*] *This three-star application is provided earlier in this report in chapter 8 (see page 56)*

What is a three-star application

In chapter 1 we defined 'three-star' Bayesian health technology assessment studies as those

1. intending to confirm the value of a technology
2. using an informative, carefully considered prior distribution for the primary quantity of interest and
3. updating, or planning to update, this prior distribution by Bayes's theorem.

Table 24 summarises the three-star applications identified by the review with respect to the BayesWatch criteria outlined in chapter 8, although we have placed 'evidence from study' earlier in the list. Out of the 30 studies identified, all but two considered the application of Bayesian methods in a randomised controlled trial setting, with 17 studies adopting a normal approximation to the appropriate likelihood and a corresponding conjugate analysis. This is particularly interesting since the majority of papers appeared after 1993 yet only two used a MCMC technique, and highlights the wide applicability of a normal–normal conjugate model. Perhaps more disappointingly only nine studies conducted the analysis prospectively, of which four also undertook elicitation of subjective prior beliefs from experts, though an additional four studies undertook some form of elicitation exercise before conducting a retrospective analysis. Although 11 of the 30 studies considered some form of clinical/policy demand, only two of these did so using a formal loss function. Particularly encouraging was the fact that all but five of the studies undertook some form of sensitivity analysis.

Three-star applications

Author. Abrams K, Ashby D, Houghton J and Riley D.[2]

Title. Assessing drug interactions: tamoxifen and cyclophosphamide.

Year. 1996.

The technology. Tamoxifen and cyclophosphamide as treatments in early breast cancer.

Objectives of study. To compare overall and disease-free survival with each drug separately and in combination.

Design of study. Randomised controlled trial: 2 × 2 factorial study with 2230 women randomised.

Evidence from study. Survival curves and estimated hazard ratios.

Statistical model. Proportional hazards with a fully parametric exponential model.

Prospective analysis? No.

Loss function. No.

Prior distribution. Priors for main effects obtained from trial evidence available at the start of the trial (1980) – uniform reference prior on interaction.

Computations. Laplace approximation.

Reporting. Posterior distributions and probability that effects are less than 0, for survival and disease-free survival.

Sensitivity analysis. Reference and data-based prior.

Comments. Model checking of proportional hazards assumption is carried out. Problem of using a clinical prior for the interaction is discussed. This is an example of the direct use of a previous trial's results to provide a prior, although strictly speaking the parameter addressed is not the same in the two studies – the current model includes another main effect and an interaction term.

Author. Berry DA.[51]

Title. Monitoring accumulating data in a clinical trial.

Year. 1989.

The technology. ECMO.

Objectives of study. To determine the probability of superiority and the difference in expected mortality between the ECMO and CMT treatment groups.

Design of study. Adaptive randomised controlled trial.

Evidence from study. Four deaths out of 10 controls, and zero deaths out of nine ECMO patients.

Statistical model. Models probability of death in each group via a logistic function.

Prospective analysis? No.

Loss function. No.

Prior distribution. Arbitrary prior assumed.

Computations. Numerical integration.

Reporting. Reports as the posterior probability of patient receiving superior, in terms of mortality and probability of death related to initial prognostic score.

Sensitivity analysis. Not explicitly reported.

Comments. See also Berry and Stangl,[45] Greenhouse and Wasserman[206] and Kass and Greenhouse.[267]

Author. Berry DA, Wolff MC and Sack D.[64]

Title. Decision making during a Phase III randomised controlled trial.

Year. 1994.

The technology. The paper describes a trial of the effectiveness of a vaccine, that links the HIB capsular polysaccharide to the outer-membrane protein complex (OMPC) of *Neisseria meningitidis* serogroup B, in preventing HIB infection.

Objectives of study. To minimise the expected number of cases of HIB amongst Navajo children in the next 20 years.

Design of study. Randomised controlled trial.

Evidence from study. A total of 5190 children in Navajo were vaccinated at 2 and 4 months, with either the vaccine being tested or a placebo. Evidence is the number of children who contract HIB in the vaccinated and unvaccinated group each month', where a 'month' is the length of time taken to enrol 105 children in each group.

Statistical model. Poisson event model.

Prospective analysis? No, retrospective analysis, with prior assessed after trial had taken place.

Loss function. Assumed to be linear in the number of HIB cases.

Prior distribution. λ_v, the rate of vaccinated children affected by HIB per child month was given a gamma prior with parameters (1, 3200) and λ_p, the rate for placebo children, was given a gamma prior with parameters (5, 10,700). This was based on the judgement of one of the authors, derived from published background information.

Computations. Dynamic programming.

Reporting. After each month, the authors calculated the expected number of cases of HIB under the following assumptions:

1. The probability of the vaccine being accepted by the regulatory authorities, following a subjectively assessed model by one of the authors.
2. The time taken for the vaccine to become available once the vaccine is approved (if it is) is given by the smaller of 1 year and $(2 \times g)$ years, where g is the probability of accepting the vaccine as given above.
3. If the vaccine is rejected, it is assumed that one of 50% efficiency will be developed in 10 years. (Efficiency is defined as 1 – [(incidence among vaccinated)/(incidence among non-vaccinated)].)

Sensitivity analysis. The excess number of HIB cases in the case of stopping as opposed to not stopping is plotted against month for:

1. different horizons, that is, different lengths of time over which the expected number of HIB cases is to be minimised (5, 10, 40 and 80 years)
2. different priors for λ_v and λ_p
3. different values for the parameter s.

Comments. A rare example of a full decision-theoretic analysis, but a retrospective study. Mathematical details are described by Berry *et al.*[63]

Author. Brophy JM and Joseph L.[76]

Title. Placing trials in context using Bayesian analysis – GUSTO revisited by Reverend Bayes.

Year. 1995.

The technology. Thrombolytic therapy following myocardial infarction.

Objectives of study. To estimate the mortality and stroke rate difference between tissue plasminogen activator (t-PA) and streptokinase.

Design of study. Randomised controlled trial (GUSTO).

Evidence from study. Event rate data from GUSTO.

Statistical model. Binomial.

Prospective analysis? No, carried out after publication of GUSTO results.

Loss function. No explicit loss.

Prior distribution. Based on pooled data from two previous trials (Gruppo Italiano per lo Studio della Sopravvivenza dell'Infarto Miocardico 2 (GISSI-2) and Third International Study of Infarct Survival (ISIS-3)), downweighted to 50 and 10%, respectively.

Computations. Conjugate normal approximations.

Reporting. Posterior plots and probabilities of net benefit.

Sensitivity analysis. Reference, full prior and discounted prior.

Comments. Letters: Avins[26] says null hypothesis should be shifted to a 1% difference, Browne[79] says Bayesian methods are not appropriate, and Goodman and Langer[196] warn about assuming exchangeability.

Author. Brophy JM, Joseph L.[77]

Title. Bayesian interim statistical analysis of randomised trials.

Year. 1997.

The technology. Angiotensin-converting enzyme (ACE) inhibitors for congestive heart failure.

Objectives of study. To estimate mortality benefit over placebo.

Design of study. Randomised controlled trial (Trandolapril Cardiac Evaluation, TRACE).

Evidence from study. Event rate data from TRACE.

Statistical model. Binomial.

Prospective analysis? No, carried out after publication of TRACE results.

Loss function. No explicit loss.

Prior distribution. Based on pooled data from two previous trials (Survival And Ventricular Enlargement (SAVE) and Acute Infarction Ramipril Efficacy (AIRE) studies).

Computations. Conjugate normal approximations.

Reporting. Posterior plots and probabilities of net benefit.

Sensitivity analysis. None.

Comments. The authors suggest that TRACE could have ended earlier owing to consistency with previous findings. In a reply, Kober *et al.*[271] say previous trials were not sufficiently relevant. Letters: Hall[220] points out a simple normal approximation, Fayers[165] suggests monitoring using a sceptical prior, and Abrams and Jones[6] point out that DMC can also make predictions.

Author. Carlin BP, Chaloner K, Church T, Louis TA and Matts JP.[90]

Title. Bayesian approaches for monitoring clinical trials with an application to toxoplasmic encephalitis prophylaxis.

Year. 1993.

The technology. The use of the drug pyrimethamine in preventing toxoplasmic encephalitis among HIV-positive patients.

Objectives of study. To estimate log(hazard ratio) associated with the treatment for time until development of toxoplasmic encephalitis.

Design of study. Randomised controlled trial.

Evidence from study. Survival data available at three interim analyses, with 12 events at the final analysis, when the trial was stopped early on an informal stopping rule.

Statistical model. Cox regression with two covariates: treatment and baseline CD4 counts.

Prospective analysis? Priors elicited prospectively, and analyses carried out retrospectively and not used in monitoring study.

Loss function. No explicit loss function, but 25% reduction in hazard used as the lower bound of range of equivalence.

Prior distribution. Priors elicited from five AIDS experts (three are clinicians) using techniques described in Chaloner *et al.*[98] Beliefs about 2 year survival were transformed into a 31-point histogram on treatment coefficient. Priors were not given, but the prior of one of the experts was plotted.

Computations. Normal likelihood, exact posterior by discretisation, and normal approximation to posterior.

Reporting. Number of events at each reporting date; plots of likelihood, prior, posterior and normal approximation to posterior for the first two experts; and plots of probability of 25 and 50% reductions in hazard rates at each reporting stage with priors of the first two experts with likelihood, exact posterior and normal approximation to posterior.

Sensitivity analysis. Five priors from different experts.

Comments. Marked conflict between the optimism of the prior distribution and the ineffectiveness of treatment (see page 22). Authors discuss possible reasons. This trial is further discussed by Carlin and Sargent,[93] where the use of robust monitoring schemes (identifying what class of priors would lead to specific conclusions) are explored retrospectively.

Author. Dersimonian R.[129]

Title. Meta-analysis in the design and monitoring of clinical trials.

Year. 1996.

The technology. Calcium supplementation in the prevention of pre-eclampsia in pregnant women.

Objectives of study. To estimate the odds ratio associated with treatment with regard to prevention of pre-eclampsia.

Design of study. Randomised controlled trial of maximum size 4500 with interim analyses.

Evidence from study. None: trial is being designed.

Statistical model. Normal approximation to likelihood for log(odds ratio).

Prospective analysis? Yes.

Loss function. No explicit loss function.

Prior distribution. 'Enthusiastic' normal prior derived from the meta-analysis of previous studies, and sceptical prior centred at 0 with same precision as the enthusiastic prior, equivalent to an experiment in which 8% of the intended sample size had been entered with no observed treatment effect.

Computations. Conjugate analysis.

Reporting. Stopping boundaries plotted under all three priors, as well as those obtained under Pocock and O'Brien–Fleming rules.

Sensitivity analysis. Boundaries for priors with four combinations of mean and precision are compared.

Comments. None.

Author. Digman JJ, Bryant J, Wieand HS, Fisher B and Wolmark N.[133]

Title. Early stopping of a clinical trial when there is evidence of no treatment benefit: protocol B-14 of the National Surgical Adjuvant Breast and Bowel Project.

Year. 1998.

The technology. Tamoxifen therapy for prevention of recurrence of breast cancer.

Objectives of study. To estimate the disease-free survival benefit from tamoxifen over placebo.

Design of study. Sequential randomised controlled study (protocol B-14 of the National Surgical Adjuvant Breast and Bowel Project) using O'Brien–Fleming stopping boundaries.

Evidence from study. At the third interim analysis there were 32 events on placebo and 56 on tamoxifen, giving a normalised likelihood with a mean of –0.53 and standard deviation of 0.21.

Statistical model. Proportional hazards.

Prospective analysis? Yes.

Loss function. No explicit loss.

Prior distribution. An 'optimistic' prior centred on a 40% hazard reduction and a 5% chance of no effect, equivalent on the log(hazard ratio) scale to a normal prior with a mean of 0.51 and standard deviation of 0.31.

Computations. Conjugate normal.

Reporting. Posterior distribution, probability of benefit from tamoxifen, and 95% interval for the hazard ratio.

Sensitivity analysis. Range of prior distributions with means varying between optimistic and sceptical.

Comments. Analysis suggested that an optimist would have a 13% belief of treatment benefit, and therefore would not rule out further trials. But a predictive calculation suggests that continued follow-up would almost certainly not lead to evidence of benefit for tamoxifen.

Author. Fayers PM, Ashby D and Parmar MKB.[166]

Title. Bayesian data monitoring in clinical trials.

Year. 1997.

The technology. Pre-operative chemotherapy for resectable oesophageal cancer.

Objectives of study. To estimate log(hazard ratio) associated with treatment with respect to survival, and to suggest stopping criteria for a sequential trial.

Design of study. Randomised controlled trial, with interim analyses.

Evidence from study. Survival in each arm at interim analyses.

Statistical model. Proportional hazards model.

Prospective analysis? Yes: methods are being used for monitoring, but the study is continuing and data are confidential, so imaginary data were used in this tutorial.

Loss function. No explicit loss function, although the chances of 5 and 10% improvement in 2 year survival were used as quantities of interest.

Prior distribution. Archetypal reference, and sceptical and enthusiastic priors, following the suggestions of Spiegelhalter *et al.*[421]

Computations. Conjugate analysis using normal approximation to the likelihood.

Reporting. Observed log(hazard ratio), and posterior probability of 0, 5 and 10% improvements in 2 year survival.

Sensitivity analysis. Results reported for reference, sceptical and enthusiastic priors.

Comments. This study suggests monitoring sequential trials according to whether a sceptic is 'convinced' of efficacy, or an enthusiast is convinced of inefficacy.

Author. Fletcher A, Spiegelhalter D, Staessen J, Thijs L and Bulpitt C.[173]

Title. Implications for trials in progress of publication of positive results.

Year. 1993.

The technology. The treatment of isolated systolic hypertension with the aim of preventing coronary heart disease and stroke in the elderly.

Objectives of study. To estimate the risk reduction associated with treatment and hence judge whether a confirmatory trial is justifiable.

Design of study. Randomised controlled trial (Systolic Hypertension in the Elderly Program, SHEP).

Evidence from study. Thirty-seven per cent reduction in fatal strokes ($P = 0.0004$; 95% confidence interval, 18 to 51%). Other reductions seen for fatal stroke, myocardial infraction or coronary heart disease deaths.

Statistical model. Normal approximation to likelihood for risk reduction.

Prospective analysis? No, retrospective analysis.

Loss function. No explicit loss function, but range of equivalence of 0–15% risk reduction selected as being 'reasonable'.

Prior distribution. Archetypal sceptical prior ("only to give a flavour of the approach") normal, mean 0 and $\Pr(|x|) > 33\% = 0.05$.

Computations. Conjugate normal.

Reporting. Observed values with confidence intervals and *P* values, plots of likelihood and posteriors for strokes and coronary heart disease deaths, probabilities of any effect greater than zero and effect greater than upper point of range of equivalence, and probability of future trials yielding a significant result.

Sensitivity analysis. Not mentioned.

Comments. None.

Author. Freedman LS and Spiegelhalter DJ.[182]

Title. Application of Bayesian statistics to decision making during a clinical trial.

Year. 1992.

The technology. The study described was intended to compare six treatments of colorectal cancer. This paper concentrates on comparisons between:

1. subjects receiving the standard treatment, the chemopheraputics agent 5-fluorouracil
2. subjects receiving the standard treatment plus high-dose leucovorin
3. subjects receiving the standard treatment plus high-dose cisplatinum.

Objectives of study. Estimation of log(hazard ratio).

Design of study. Randomised controlled trial.

Evidence from study. 5-Fluorouracil versus 5-fluorouracil plus leucovorin after 70 subjects had entered each arm: 82 deaths had occurred and the estimated log(hazard ratio) was 1.65. 5-Fluorouracil versus 5-fluorouracil plus cisplatinum after 70 subjects had entered each arm: 96 deaths had occurred and the estimated log(hazard ratio) was –0.01.

Statistical model. Proportional hazards model.

Prospective analysis? No, retrospective analysis.

Loss function. No explicit loss function, but the range of equivalence was log(1)–log(1.5).

Prior distribution. As a prior distribution the authors use a mixture distribution with a mass of p_0 on 0, and a mass of $1 - p_0$ on a normal distribution with a mean of 0 and a variance of $4/n_0$, truncated at 0. The authors use values of $p_0 = 0.25$ and $n_0 = 25$ as giving a 5% chance that the experimental regimen doubles survival time, with a small chance of no benefit "in view of the observed regression of tumours".

Computations. Closed form normal approximations.

Reporting. Plots of likelihood and posterior, mention of likelihood/prior agreement, *z* values, and posterior probabilities critical to stopping decisions.

Sensitivity analysis. In the 5-fluorouracil plus leucovorin trial, posterior probabilities are calculated for four different sets of values of parameter values for the prior.

Comments. Further analyses of this study are provided by Greenhouse[205] and Dixon and Simon.[135]

Author. Freedman LS, Spiegelhalter DJ and Parmar MKB.[183]

Title. The what, why and how of Bayesian clinical trials monitoring.

Year. 1994.

The technology. The treatment of Duke's C colorectal cancer with levamisole plus 5-fluorouracil.

Objectives of study. To estimate log(hazard ratio).

Design of study. Randomised controlled trial.

Evidence from study. Data from published study.

Statistical model. Proportional hazards, and approximate normal likelihood for log(hazard ratio).

Prospective analysis? No.

Loss function. No, but range of equivalence assessed: 0–log(1.33) on log(hazard ratio) scale.

Prior distribution. A sceptical prior with a mean of 0 and a 5% chance that log(hazard ratio) is greater than the alternative hypothesis of 0.3, and an enthusiastic prior with the same precision but a mean of 0.3.

Computations. Conjugate normal.

Reporting. Number of deaths in each group; estimated log(hazard ratio); plots of prior, likelihood and posterior; posterior mean and standard deviation; and the probability of log(hazard ratio) being in range of equivalence range and on either side.

Sensitivity analysis. Reference, sceptical and enthusiastic priors.

Comments. Discussed further by Spiegelhalter *et al.*[421]

Author. George SL, Li CC, Berry DA and Green MR.[191]

Title. Stopping a clinical trial early – frequentist and Bayesian approaches applied to a CALGB trial in non-small-cell lung cancer.

Year. 1994.

The technology. The addition of two cycles of induction chemotherapy prior to thoracic radiation in subjects with Phase II non-small cell cancer.

Objectives of study. To compare survival by estimating log(hazard ratio).

Design of study. Sequential randomised controlled study (CALGB 8433) using O'Brien–Fleming stopping boundaries.

Evidence from study. Stopped at the fifth interim analysis, there had been 56 deaths and the survival difference had an adjusted *P* value of 0.0015.

Statistical model. Exponential survival model.

Prospective analysis? No, prior distributions chosen after the trial.

Loss function. No explicit loss.

Prior distribution. A gamma prior for the rate on standard treatment assessed from explicit evidence, and a sceptical prior distribution for log(hazard ratio) with standard deviation of 1, giving 16% chance that the true effect exceeds the alternative hypothesis.

Computations. Gibbs sampling.

Reporting. Posterior distribution, probability of log(hazard ratio) being less than –0.25 and –0.5.

Sensitivity analysis. Analysis by proper and sceptical priors.

Comments. The authors simulate and display predictive distributions of what demand probabilities would have been estimated had the trial run its full course given the data at the final analysis. This trial is also analysed using a more complex Weibull model, but a matching prior opinion, by Qian *et al.*[368]

Author. Greenhouse JB and Wasserman L.[206]

Title. Robust Bayesian methods for monitoring clinical trials.

Year. 1995.

The technology. ECMO for premature babies.

Objectives of study. To compare survival with ECMO (new) and without (conventional regimen CMT).

Design of study. Sequential adaptive randomised controlled trial.

Evidence from study. Six out of 10 survivors on CMT, and nine out of nine on ECMO.

Statistical model. Binomial.

Prospective analysis? No.

Loss function. No.

Prior distribution. Uniform prior on survival rates as the baseline; ε contaminated the prior around this.

Computations. Numerical integration.

Reporting. Upper and lower bounds on "*P*(ECMO superior to CMT)", over all possible priors.

Sensitivity analysis. Search over all priors with ε = 0.1, 0.2, 0.3 and 0.4.

Comments. The authors also consider the single-armed Phase II study of Korn *et al.*,[276] in which 3/4 patients showed toxicity. They again use a robust prior based on past trial evidence (with and without discounting), reporting the bounds on the posterior probability of at least 20% toxicity.

Author. Gustafson P.[219]

Title. Robustness considerations in Bayesian analysis.

Year. 1996.

The technology. ECMO.

Objectives of study. To estimate the probability of survival better for ECMO patients, that is, $P(\theta_E > \theta_C|\text{Data})$.

Design of study. Adaptive randomised controlled trial.

Evidence from study. Four deaths out of 10 controls, and zero deaths out of nine ECMO patients.

Statistical model. Model/likelihood: (θ_C, θ_E), the survival probabilities for control and ECMO patients, respectively, are assumed independent.

Prospective analysis? No.

Loss function. No.

Prior distribution. Arbitrary, initial Beta(1.25, 1.25) prior distributions are assumed for both θ_C and θ_E.

Computations. Not explicitly reported.

Reporting. Ranges of posterior probabilities for the three different classes of prior distributions as data accumulates.

Sensitivity analysis. Uses three different classes of unrestricted, restricted and density-bounded contamination classes of prior distributions, assuming θ_C and θ_E are independent, and assuming that the prior for one parameter is fixed throughout.

Comments. None.

Author. Hughes MD.[242]

Title. Practical reporting of Bayesian analyses of clinical trials.

Year. 1991.

The technology. Beta-blocker treatment for prevention of bleeding oesophageal varices.

Objectives of study. To compare the incidence of bleeding and survival with a beta-blocker (new regimen) and placebo (existing regimen).

Design of study. Randomised controlled trial.

Evidence from study. No evidence available at the time of analysis: hypothetical data used.

Statistical model. Approximate normal likelihood for log(odds ratio).

Prospective analysis? Yes.

Loss function. No.

Prior distribution. Reference prior, data-based prior obtained from simple pooling of previous studies, and subjective prior elicited from six participating clinicians as a histogram with a risk-difference scale. Also considers Bayes factor analysis in which a lump of prior is placed at 0.

Computations. Conjugate normal analysis.

Reporting. Posterior distributions and probabilities of benefit.

Sensitivity analysis. Reference, clinical and data-based posterior distributions. For 'lump' prior, the study explores the sensitivity of the "posterior probability of no effect" versus "prior probability of no effect".

Comments. See Hughes[243] for more details.

Author. Kadane JB.[258]

Title. Bayesian methods and ethics in a clinical trial design.

Year. 1996.

The technology. Drug treatment to control hypertension following open-heart surgery.

Objectives of study. To compare verapamil with nitroprusside with respect to post-operative blood pressure.

Design of study. 'Ethical' design, in which a patient will only receive a treatment if at least one member of a team of experts (represented by a model of his or her current posterior beliefs) considers that the treatment is optimal for the patient's covariates. If more than one member considers a treatment to be optimal, then the treatment is allocated to balance prognostic factors between groups.[259,261]

Evidence from study. A total of 49 patients were enrolled, of which 30 could be studied: 12 on nitroprusside, 18 on verapamil. As the trial proceeded, there was an imbalance towards verapamil as the experts' opinions changed.

Statistical model. Normal linear model for blood pressure depending on four covariates and treatment.

Prospective analysis? Yes.

Loss function. No.

Prior distribution. Pilot data on five patients were available. Prior opinions were elicited from five participating clinicians through hour-long interactive computer sessions.

Computations. Conjugate normal/*t* analysis.

Reporting. Prior and posterior distributions were presented for each of the five experts, and for each of 16 types of patient. However, the primary results of the trial (see chapter 12) were presented using a non-Bayesian analysis.

Sensitivity analysis. Results for each expert were displayed.

Comments. The study was passed by the internal review board of the Johns Hopkins Hospital. Further elicitations were necessary midway through the study when the safety criterion being updated was changed. The elicitation procedure is discussed in detail and difficulties acknowledged. A bug in the allocation program was found midway, which had meant that some of the early patients had not in fact been allocated to the appropriate treatment, although it is claimed that no adverse effects resulted.

Author. Kass RE and Greenhouse JB.[267]

Title. Comments on 'Investigating therapies of potentially great benefit: ECMO' by J H Ware.

Year. 1989.

The technology. ECMO.

Objectives of study. To estimate the log(mortality odds ratio) associated with ECMO patients.

Design of study. Adaptive randomised controlled trial.

Evidence from study. Four deaths out of 10 controls, and zero deaths out of nine ECMO patients.

Statistical model. Model/likelihood: (δ, γ) where $\delta = n_C - n_E$ and $\gamma = (n_C + n_E)/2$ and n_C and n_E are log(odds ratio) of death in the control and ECMO patients, respectively.

Prospective analysis? No.

Loss function. No, but log(odds ratio) of 0.4 considered important, not based on clinical demands.

Prior distribution. Historical clinical series of 13 patients receiving conventional medical therapy, of whom two survived.

Computations. Not specifically reported, but possibly used Laplace approximations to the posterior distribution.

Reporting. Posterior probabilities, but also consider the use of Bayes factors, based on each of the five prior-to-posterior analyses, in order to determine the evidence for and against a treatment effect, that is, on the log(odds ratio) scale, of zero.

Sensitivity analysis. Five different prior distributions used in the analysis reported, with the historical evidence downweighted in each.

Comments. See also Berry and Stangl[45] Berry[51] and Greenhouse and Wasserman.[206]

Author. Lewis RJ.[293]

Title. Bayesian hypothesis testing: interim analysis of a clinical trial evaluating phenytoin for the prophylaxis of early post-traumatic seizures in children.

Year. 1996.

The technology. Phenytoin for the prophylaxis of early post-traumatic seizures in children.

Objectives of study. To estimate the difference in 48 hour seizure rates with and without phenytoin.

Design of study. Randomised controlled trial, with 95% power to detect a 50% reduction in the seizure rate.

Evidence from study. First interim analysis, 0/7 versus 2/7 seizures on phenytoin.

Statistical model. Binomial model.

Prospective analysis? Yes.

Loss function. Yes, an implicit loss function is found that would lead to a decision-theoretic design with good type I and type II error rates.

Prior distribution. "Wide, pessimistic and optimistic" priors assessed.

Computations. Conjugate binomial analysis.

Reporting. Posterior distributions and probability that rate difference is greater than 12.5% and 0.

Sensitivity analysis. Results for the three priors are given.

Comments. The decision to continue was made, in spite of initially unencouraging results.

Author. Lilford RJ and Braunholtz D.[299]

Title. The statistical basis of public policy: a paradigm shift is overdue.

Year. 1996.

The technology. Third-generation contraceptive pill.

Objectives of study. To assess the risk of venous thrombosis associated with the third-generation contraceptive pill.

Design of study. Meta-analysis of case–control studies.

Evidence from study. Odds ratio of 2 and 95% interval of 1.4 to 2.7.

Statistical model. Normal likelihood for log(odds ratio).

Prospective analysis? No.

Loss function. No.

Prior distribution. Priors were subjectively assessed by two experts.

Computations. Conjugate normal.

Reporting. Posterior distributions and 95% intervals.

Sensitivity analysis. Different priors examined, as well as discounting likelihood by 30% additional dispersion (undirected bias) and allowing for the possibility of a 30% overestimate (directed bias).

Comments. Critical letters included one by Cox and Farewell,[118] who pointed out that bias could be explored through sensitivity analysis and that the authors were setting up an unrealistic "straw man" of traditional statistics as relying heavily on significance testing.

Author. Parmar MKB, Spiegelhalter DJ and Freedman LS.[348]

Title. The CHART trials: Bayesian design and monitoring in practice.

Year. 1994.

The technology. CHART radiotherapy regimen for non-small-cell lung and head-and-neck cancer.

Objectives of study. To compare survival with (new regimen) and without (existing regimen) CHART .

Design of study. Randomised controlled trial.

Evidence from study. No evidence available at the time of analysis: hypothetical data used.

Statistical model. Proportional hazards and approximate normal likelihood for log(hazard ratio).

Prospective analysis? Yes.

Loss function. No, but ranges of equivalence assessed.

Prior distribution. Elicited from 11 participating clinicians.

Computations. Conjugate normal analysis.

Reporting. Posterior distributions and probabilities relative to ranges of equivalence.

Sensitivity analysis. Reference, clinical and sceptical priors.

Comments. See chapter 9 for details of this study. The eventual results of these studies are discussed on page 4.7.

Author. Parmar MKB, Ungerleider RS and Simon R.[349]

Title. Assessing whether to perform a confirmatory randomised clinical trial.

Year. 1996.

The technology. Adjunct chemotherapy for non-small-cell lung cancer.

Objectives of study. To compare survival with (new regimen) and without (existing regimen) chemotherapy

Design of study. Randomised controlled trial conducted between 1984 and 1987.

Evidence from study. Hazard ratio of 1.63 (1.14–2.33).

Statistical model. Proportional hazards and approximate normal likelihood for log(hazard ratio).

Prospective analysis? No.

Loss function. No.

Prior distribution. Default reference and sceptical priors.

Computations. Conjugate normal analysis.

Reporting. Posterior distributions and probabilities relative to median survival improvements of 3, 4 and 5 months.

Sensitivity analysis. Reference, clinical and sceptical priors.

Comments. See page 33 for details of this study. This paper also considers another chemotherapy study using the same prior distribution.

Author. Pocock SJ and Hughes MD.[364]

Title. Practical problems in interim analyses, with particular regard to estimation.

Year. 1989.

The technology. Anisoylated plasminogen streptokinase activator complex (APSAC) thrombolytic therapy after myocardial infarction.

Objectives of study. To compare 30 day mortality under APSAC (new regimen) and placebo (existing regimen).

Design of study. Randomised controlled trial (APSAC Intervention Mortality Study, AIMS).

Evidence from study. At the second interim analysis: 32 versus 61 deaths with 502 patients in each arm.

Statistical model. Approximate normal likelihood for log(risk ratio).

Prospective analysis? No.

Loss function. No.

Prior distribution. Normal on log(risk ratio) scale, with a median risk ratio of 0.8, a 7% chance of a risk ratio above 1, with a 10% chance of a risk ratio below 0.67 that "seems plausible given earlier clinical trials of other thrombolytic agents".

Computations. Conjugate normal analysis.

Reporting. Median and 95% intervals for risk ratio.

Sensitivity analysis. Four choices of prior median and standard deviation explored.

Comments. Simulation exercise to see how the prior generates results, and what kind of biases would occur with sequential trials.

Author. Pocock SJ and Spiegelhalter DJ.[363]

Title. Domiciliary thrombolysis by general practitioners.

Year. 1992.

The technology. Home thrombolytic therapy after myocardial infarction.

Objectives of study. To compare the 30 day mortality rate under antistreplase (new regimen) and placebo (existing regimen).

Design of study. Randomised controlled trial (GREAT).

Evidence from study. At interim analysis: 23/148 on control versus 13/163 deaths on new treatment.

Statistical model. Approximate normal likelihood for log(odds ratio).

Prospective analysis? No.

Loss function. No.

Prior distribution. Elicited from an expert, who based his opinion on published and unpublished data.

Computations. Conjugate normal analysis.

Reporting. Posterior distribution of risk ratio, mean and 95% interval.

Sensitivity analysis. None.

Comments. See chapter 2 for further discussion of this example.

Author. Sasahara AA, Cole TM, Ederer F, Murray JA, Wenger NK, Sherry S and Stengle JM.[389]

Title. Urokinase Pulmonary Embolism Trial, a national cooperative study.

Year. 1973.

The technology. Urokinase treatment in pulmonary embolism.

Objectives of study. To compare thrombolytic capability on urokinase (new regimen) with heparin (standard regimen).

Design of study. Randomised controlled trial.

Evidence from study. Multiple end-points on 160 patients entered between 1968 and 1970.

Statistical model. Normal model.

Prospective analysis? Yes.

Loss function. No.

Prior distribution. Point mass on null hypothesis, with remainder normally distributed around 0, with standard deviation such that the expected effect, were it to be present, would be equal to the alternative hypothesis (see page 19). Alternative hypotheses were "based on what appeared reasonable from previous experience with thrombolytics".

Computations. Conjugate normal analysis.

Reporting. "Relative betting odds", that is, posterior odds on null hypothesis.

Sensitivity analysis. None.

Comments. No *P* values were provided in the analysis.

Author. Spiegelhalter DJ, Freedman LS and Parmar MKB.[420]

Title. Applying Bayesian ideas in drug development and clinical trials.

Year. 1993.

The technology. Chemotherapy for osteosarcoma.

Objectives of study. **To c**ompare survival in multi-drug (new) and two-drug (existing) regimens.

Design of study. Randomised controlled trial.

Evidence from study. None available at time of analysis.

Statistical model. Proportional hazards and approximate normal likelihood for log(hazard ratio).

Prospective analysis? Yes.

Loss function. No, but a range of equivalence of 0–10% improvement in 5 year survival.

Prior distribution. Elicited priors from seven participating oncologists, averaged to give a median

hazard ratio of 1.11, and 95% interval (0.66–1.83) in favour of new therapy. Sceptical prior with the same precision but centred on 0.

Computations. Conjugate normal analysis.

Reporting. Posterior distributions and probabilities relative to range of equivalence.

Sensitivity analysis. Reference, clinical and sceptical priors.

Comments. Final results are now available[413] showing no evidence of benefit: hazard ratio = 0.94 (0.69–1.27) (see page 4.7).

Author. Spiegelhalter DJ, Freedman LS and Parmar MKB.[421]

Title. Bayesian approaches to randomised trials (with discussion).

Year. 1994.

The technology. Misonidazole as adjunct chemotherapy for head and neck cancer.

Objectives of study. To compare primary control with (new regimen) and without (existing regimen) misonidazole.

Design of study. Randomised controlled trial.

Evidence from study. Third interim analysis: 108 events and a hazard ratio of 0.9 (0.62–1.31).

Statistical model. Proportional hazards and approximate normal likelihood for log(hazard ratio).

Prospective analysis? No.

Loss function. No, but a range of equivalence of 0–0.414 in log(hazard ratio), corresponding to a rise from 25 to 40% improvement in 2 year primary control.

Prior distribution. Default reference, sceptical and enthusiastic priors.

Computations. Conjugate normal analysis.

Reporting. Posterior distributions and probabilities relative to a range of equivalence.

Sensitivity analysis. Reference, clinical and sceptical priors.

Comments. The probability of clinical superiority was low even with an enthusiastic prior. In fact, the trial was terminated at this third analysis. This paper also considers 'three-star' analyses of the neutron study[1] and the levamisole plus 5-fluorouracil study.[183]

Author. Stangl DK.[427]

Title. Prediction and decision-making using Bayesian hierarchical-models.

Year. 1995.

The technology. Imipramine hydrochloride for prevention of or delaying a return to depression.

Objectives of study. To compare the time to recurrent depression with (new regimen) and without (existing regimen) imipramine.

Design of study. Five-centre randomised controlled trial.

Evidence from study. Survival data from five centres.

Statistical model. Exponential survival model, with the option of a change point to allow for a non-constant hazard.

Prospective analysis? No.

Loss function. No.

Prior distribution. Gamma priors were obtained by considering the first gamma parameter to be a sample size, and making the second have a gamma prior which itself has a second parameter equal to the maximum likelihood estimate. Three different sets of priors were used.

Computations. MCMC (Gibbs sampling).

Reporting. The expectation of the time to recurrence for patient in each treatment group, and the expected difference between the treated and untreated group and at each clinic, were provided under each model and prior, and the predictive distribution of the difference at each clinic under one model and prior was plotted.

Sensitivity analysis. Three different priors were explored.

Comments. The use of decision theory is explored in deciding which treatment to give to a patient.

This study was also analysed using Laplace approximations[425] in which further sensitivity analysis to prior assumptions is carried out.

Author. Ware J.[469]

Title. Investigating therapies of potentially great benefit: ECMO (with discussion).

Year. 1989.

The technology. ECMO.

Objectives of study. To estimate the difference in mortality associated with ECMO patients.

Design of study. Adaptive randomised controlled trial.

Evidence from study. Four deaths out of 10 controls, and zero deaths out of nine ECMO patients.

Statistical model. Model/likelihood: (p_C, p_E), the survival probabilities in the control and ECMO groups, respectively, such that a beta prior is assumed for p_C, and the conditional distribution of p_E, given p_C, is such that $P(p_C < p_E) = P(p_C = p_E) = P(p_C > p_E) = 1/3$.

Prospective analysis? No.

Loss function. No.

Prior distribution. Historical clinical series of 13 patients receiving conventional medical therapy, of whom two survived.

Computations. Due to the model formulation, posterior probabilities could be obtained in a closed form.

Reporting. Posterior probabilities: $P(p_C < p_E)$, $P(p_C = p_E)$, $P(p_C > p_E)$.

Sensitivity analysis. Both an informative data-based prior and a vague prior for p_C were used.

Comments. None.

Appendix 2

Websites and software

Here we provide a selection of sites that currently provide useful material on Bayesian methods applicable to health technology assessment and lists of links. This list is not exhaustive but should provide some entry into the huge range of material available on the Internet.

This sites below were all functioning in November 2000.

Bayesian methods in health technology assessment

http://www.fda.gov/cdrh/
This is the home page for the US Food and Drug Administration's Center for Devices and Radiological Health, which contains a number of items relating to Bayesian methods. To identify these, use the search facility with the keyword 'Bayesian'.

http://www.bayesian-initiative.com/
The Bayesian Initiative in Health Economics and Outcome Research provides useful background material on Bayesian approaches to pharmacoeconomics, but does not appear to have been updated for some time.

Bayesian software

http://www.shef.ac.uk/~st1ao/1b.html
The First Bayes software is freely available, and features good graphical presentation of conjugate analysis of basic data sets. It is suitable for teaching, and is strong on predictive distributions.

http://www.mrc-bsu.cam.ac.uk/bugs/
The BUGS software is designed for analysis of complex analysis using MCMC methods. The WinBUGS version features an interface for specifying models as graphs. The software assumes familiarity with Bayesian methods and MCMC computation.

http://www.epi.mcgill.ca/~web2/Joseph/software.html
Lawrence Joseph's Bayesian Software site provides downloadable code for a wide variety of sample size calculations using prior opinion.

http://omie.med.jhmi.edu/bayes/
The Bayesian Communication website is hosted by Harold Lehmann, and features a prototype example in which a Bayesian analysis can be carried out on-line.[289–291]

http://www.research.att.com/~volinsky/bma.html
The Bayesian Model Averaging web page provides S-plus and Fortran software for carrying out model averaging, as well as featuring reprints and links.

http://www.palisade.com/
The Palisade Corporation markets the @RISK® software, which is an add-on to spreadsheet packages that allows probability distributions to be placed over the inputs to spreadsheets. Predictive distributions over the outputs are then obtained by simulation. Demonstrator versions are available for downloading.

http://www-math.bgsu.edu/~albert/mini_bayes/info.html
This site is an adjunct to Jim Albert's book *Bayesian Computation Using Minitab*, and features macros for carrying out a variety of analyses.

General Bayesian sites

http://bayes.stat.washington.edu/bayes_people.html
The Bayesian Statistics Personal Web Pages site has links to the home pages of many researchers in Bayesian methods. These provide a vast array of lecture notes, reprints and slide presentations.

http://www.bayesian.org/
The International Society for Bayesian Analysis provides information on its activities and useful links.

http://www.stat.ucla.edu/~jsanchez/sbssnews/sbssnews.html
The American Statistical Association Section on Bayesian Statistical Sciences (SBSS) has a preprint archive and links to other sites.

http://www.isds.duke.edu/sites/bayes.html
This web page hosted from Duke University provides a list of Bayesian sites.

Methodology Group

Members

Methodology Progamme Director
Professor Richard Lilford
Director of Research and Development
NHS Executive – West Midlands, Birmingham

Chair
Professor Martin Buxton
Director, Health Economics Research Group
Brunel University, Uxbridge

Professor Douglas Altman
Professor of Statistics in Medicine
University of Oxford

Dr David Armstrong
Reader in Sociology as Applied to Medicine
King's College, London

Professor Nicholas Black
Professor of Health Services Research
London School of Hygiene & Tropical Medicine

Professor Ann Bowling
Professor of Health Services Research
University College London Medical School

Professor David Chadwick
Professor of Neurology
The Walton Centre for Neurology & Neurosurgery
Liverpool

Dr Mike Clarke
Associate Director (Research)
UK Cochrane Centre, Oxford

Professor Paul Dieppe
Director, MRC Health Services Research Centre
University of Bristol

Professor Michael Drummond
Director, Centre for Health Economics
University of York

Dr Vikki Entwistle
Senior Research Fellow, Health Services Research Unit
University of Aberdeen

Professor Ewan B Ferlie
Professor of Public Services Management
Imperial College, London

Professor Ray Fitzpatrick
Professor of Public Health & Primary Care
University of Oxford

Dr Naomi Fulop
Deputy Director, Service Delivery & Organisation Programme
London School of Hygiene & Tropical Medicine

Mrs Jenny Griffin
Head, Policy Research Programme
Department of Health
London

Professor Jeremy Grimshaw
Programme Director
Health Services Research Unit
University of Aberdeen

Professor Stephen Harrison
Professor of Social Policy
University of Manchester

Mr John Henderson
Economic Advisor
Department of Health, London

Professor Theresa Marteau
Director, Psychology & Genetics Research Group
Guy's, King's & St Thomas's School of Medicine, London

Dr Henry McQuay
Clinical Reader in Pain Relief
University of Oxford

Dr Nick Payne
Consultant Senior Lecturer in Public Health Medicine
ScHARR
University of Sheffield

Professor Joy Townsend
Director, Centre for Research in Primary & Community Care
University of Hertfordshire

Professor Kent Woods
Director, NHS HTA Programme, & Professor of Therapeutics
University of Leicester

HTA Commissioning Board

Members

Programme Director
Professor Kent Woods
Director, NHS HTA Programme, &
Professor of Therapeutics
University of Leicester

Chair
Professor Shah Ebrahim
Professor of Epidemiology of Ageing
University of Bristol

Deputy Chair
Professor Jon Nicholl
Director, Medical Care Research Unit
University of Sheffield

Professor Douglas Altman
Director, ICRF Medical Statistics Group
University of Oxford

Professor John Bond
Director, Centre for Health Services Research
University of Newcastle-upon-Tyne

Ms Christine Clark
Freelance Medical Writer
Bury, Lancs

Professor Martin Eccles
Professor of Clinical Effectiveness
University of Newcastle-upon-Tyne

Dr Andrew Farmer
General Practitioner &
NHS R&D Clinical Scientist
Institute of Health Sciences
University of Oxford

Professor Adrian Grant
Director, Health Services Research Unit
University of Aberdeen

Dr Alastair Gray
Director, Health Economics Research Centre
Institute of Health Sciences
University of Oxford

Professor Mark Haggard
Director, MRC Institute of Hearing Research
University of Nottingham

Professor Jenny Hewison
Senior Lecturer
School of Psychology
University of Leeds

Professor Alison Kitson
Director, Royal College of Nursing Institute, London

Dr Donna Lamping
Head, Health Services Research Unit
London School of Hygiene & Tropical Medicine

Professor David Neal
Professor of Surgery
University of Newcastle-upon-Tyne

Professor Gillian Parker
Nuffield Professor of Community Care
University of Leicester

Dr Tim Peters
Reader in Medical Statistics
University of Bristol

Professor Martin Severs
Professor in Elderly Health Care
University of Portsmouth

Dr Sarah Stewart-Brown
Director, Health Services Research Unit
University of Oxford

Professor Ala Szczepura
Director, Centre for Health Services Studies
University of Warwick

Dr Gillian Vivian
Consultant in Nuclear Medicine & Radiology
Royal Cornwall Hospitals Trust
Truro

Professor Graham Watt
Department of General Practice
University of Glasgow

Dr Jeremy Wyatt
Senior Fellow
Health Knowledge Management Centre
University College London

Current and past membership details of all HTA 'committees' are available from the HTA website (see inside front cover for details)

Feedback

The HTA programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (http://www.ncchta.org) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 23 8059 5639 Email: hta@soton.ac.uk
http://www.ncchta.org

ISSN 1366-5278