Statistical Science
2004, Vol. 19, No. 1, 156–174
DOI 10.1214/088342304000000080

Incorporating Bayesian Ideas into Health-Care Evaluation

David J. Spiegelhalter

Abstract. We argue that the Bayesian approach is best seen as providing additional tools for those carrying out health-care evaluations, rather than replacing their traditional methods. A distinction is made between those features that arise from the basic Bayesian philosophy and those that come from the modern ability to make inferences using very complex models. Selected examples of the former include explicit recognition of the wide cast of stakeholders in any evaluation, simple use of Bayes theorem and use of a community of prior distributions. In the context of complex models, we selectively focus on the possible role of simple Monte Carlo methods, alternative structural models for incorporating historical data and making inferences on complex functions of indirectly estimated parameters. These selected issues are illustrated by two worked examples presented in a standardized format. The emphasis throughout is on inference rather than decision-making.

Key words and phrases: Bayes theorem, prior distributions, sceptical prior distribution, data monitoring committee, cost-effectiveness analysis, historical data, decision theory.

1. INTRODUCTION

The Bayesian approach to inference and decision-making has a tradition of controversy. In recent years, however, a more balanced and pragmatic perspective has developed, reflected in a notable increase in Bayesian publications in biostatistics in general and health-care evaluation in particular. The argument of this paper is that this perspective naturally leads to the view of Bayesian methods as adding to, rather than replacing, standard statistical techniques.

This somewhat ecumenical perspective is based on acknowledging that traditional methods of designing and analyzing studies have strongly contributed to advances in medical care, whether these comprise new drugs, devices or even organizational initiatives. Nevertheless it should be clear that the process of getting an intervention into routine practice makes demands that are not easily met by classical techniques. For example, when carrying out a clinical trial, the many sources of evidence and judgement available may be inadequately summarized by a single "alternative hypothesis," monitoring may be complicated by simultaneous publication of related studies and multiple subgroups may need to be analyzed and reported. Randomized trials may not be feasible or may take a long time to reach conclusions, and a single clinical trial will also rarely be sufficient to inform a policy decision, such as embarking on or continuing a research or drug development program, regulatory approval of a drug or device or recommendation of a treatment at an individual or population level. Since standard statistical methods are designed for summarizing the evidence from single studies or pooling evidence from similar studies, they have difficulties dealing with the pervading complexity of multiple sources of evidence. Furthermore, the Bayesian approach can be characterized as a means of rational learning from experience in the face of uncertainty, and since advances in health-care typically happen through incremental gains in knowledge rather than paradigm-shifting breakthroughs, this domain appears particularly amenable to a Bayesian perspective.

David J. Spiegelhalter is Senior Scientist, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK (e-mail: david.spiegelhalter@mrc-bsu.cam.ac.uk).

This paper presents a personal perspective based on experience of trying to communicate to "classically" trained statisticians. For such an audience it can be helpful to break the additional benefits of a Bayesian approach into two main strands: those that are inherent to the Bayesian philosophy, and those that arise through the ability provided by the "MCMC revolution" to handle complex models. For example, a recent issue of *Statistics in Medicine* (Volume 22, Number 10) comprised entirely of Bayesian analyses using Markov chain Monte Carlo (MCMC) methods: of these ten papers, all exploited the ability to handle complex models, but only one gave any attention to an informative prior distribution, and then only minimally (Hanson, Bedrick, Johnson and Thurmond, 2003).

This division is reflected in the structure of this paper. In Section 2 we focus on three selected features of basic Bayesian analysis, which are then illustrated in a fairly detailed example. We then go on in Section 3 to identify three important features of complex modelling followed again by an example. It will be clear that the emphasis throughout is firmly on inference rather than decision-making: this is primarily to avoid overlap with Berry (2004), but also reflects personal enthusiasm. This issue is briefly discussed in Section 4, which also attempts to put current developments in perspective and outlines some issues in increasing appropriate use of Bayesian methods.

Of course, this paper can only scratch the surface of a burgeoning literature, and only a sample of references is provided in the text. For basic arguments for the Bayesian approach in this context it is difficult to improve upon the classic papers by Jerome Cornfield, for example, Cornfield (1966, 1969, 1976). More recent introductions and polemics include Etzioni and Kadane (1995), Berry and Stangl (1996a) and Kadane (1995), while Spiegelhalter, Myles, Jones and Abrams (2000) systematically reviews the literature, Berry and Stangl (1996b) contains a wide range of applications, and O'Hagan and Luce (2003) is an excellent free primer. Much of the material presented in this paper is taken from Spiegelhalter, Abrams and Myles (2004), to which we refer for further detail.

2. THREE SELECTED FEATURES OF BASIC BAYESIAN ANALYSIS

A number of generic characteristics of the Bayesian paradigm make it especially suitable for application to health-care evaluations. Here we focus on a limited selection: acknowledgment of subjectivity and context, simple use of Bayes theorem and use of a "community" of prior distributions in order to assess the impact of new evidence. Of course many other important issues could be identified, including the ease of prediction, reporting probabilities of events of direct interest, use of prior distributions in sample size assessment and power calculations and so on: these features are reflected in the references given above.

2.1 Acknowledgment of Subjectivity and Context

Bayesian analysis is rooted in probability theory, whose basic rules are generally considered as self-evident. However, as Lindley (2000) emphasizes, the rules of probability can be derived from "deeper" axioms of reasonable behavior of an individual (say *You*) in the face of Your own uncertainty. The vital point of this subjective interpretation is that Your probability for an event is a property of Your relationship to that event, and not an objective property of the event itself. This is why, pedantically speaking, one should always refer to probabilities *for* events rather than probabilities *of* events, since the probability is conditioned on the context, which includes the observer and all the observer's background knowledge and assumptions. Bayesian methods therefore 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 an understanding of the context of the study is vital:

> Bayesian statistics treats subjectivity with respect by placing it in the open and under the control of the consumer of data (Berger and Berry, 1988).

This view appears particularly appropriate to the complex circumstances in which evaluations of health-care interventions are carried out. Apart from methodological researchers, at least five different viewpoints might be identified:

- *sponsors*—for example, the pharmaceutical industry, medical charities or granting bodies such as the U.S. National Institutes of Health, and the U.K. Medical Research Council;
- *investigators*—that is, those responsible for the conduct of a study, whether industry or publicly funded;
- *reviewers*—for example, journal editors regarding publication, and regulatory bodies for approval of pharmaceuticals or devices;
- *policy makers*—for example, agencies responsible for setting health policy taking into account cost-effectiveness, such as the U.K. National Institute

for Clinical Excellence (NICE), or individual health maintenance organizations (HMOs),
- *consumers*—for example, individual patients or clinicians acting on their behalf.

Each of these broad categories can be further subdivided. Thus a characteristic of health-care evaluation is that the investigators who plan and conduct a study are generally not the same body as that which makes decisions on the basis of the evidence provided in part by that study. An immediate consequence of this complex cast of stakeholders is that it is not generally straightforward to implement a decision-theoretic approach that is based around a single decision-maker. In addition, there are a range of possible prior distributions, which may in turn be used for design but possibly not in reporting results. All this reinforces the need for an extremely flexible approach, with a clear specification of whose beliefs and values are being expressed, and the necessity of taking forward in parallel a range of possible opinions.

2.2 Simple Use of the Bayes Theorem

The use of MCMC methods can lead to the use of extravagantly complex models, but here we consider two important applications of Bayes theorem used in its simplest analytic form: the interpretation of positive trial results, and using approximate normal likelihoods and priors.

Bayes' theorem is often introduced through examples based on diagnostic testing for a disease of known prevalence. For fixed sensitivity and specificity the posterior probability after a positive test result (or the "predictive value positive") can be calculated, and the frequent conflict of this value with naive intuition can be a good educational warning to take into account the prior probability (prevalence). In the context of health-care evaluation, the equivalent of a positive test is a "significant" finding in a clinical trial, which almost inevitably receives disproportionately more publicity than a negative finding.

There have been frequent attempts to adapt Bayesian ideas as an aid to interpretation of positive clinical trial results. For example, Simon (1994) points out that if one carries out clinical trials with Type I error $\alpha = 0.05$ and power (1 − Type II error) $1 - \beta = 0.80$, then if only 10% of investigated treatments are truly effective (not an unreasonable estimate), then Bayes theorem shows that 36% of claimed "discoveries" will be false positives. This figure will be even higher if there is additional external evidence against a particular intervention, prompting Grieve (1994) to suggest that Bayes theorem provides "a yardstick against which a surprising finding may be measured." Increasing attention to "false discovery rates" (Benjamini and Hochberg, 1995), which are essentially measures of the "predictive value positive," has refocused attention on this concept within the context of classical multiple testing.

Our second example concerns the simple use of Bayes theorem when data has been analyzed using standard statistical packages. In parametric models Bayes theorem is often taught using binomial likelihoods and conjugate beta distributions, but this framework does not fit in well with comparative evaluations for which interest will generally lie with odds ratios or hazard ratios, classically estimated using logistic or Cox regression analyses provided within standard statistical packages. Practitioners will therefore generally have data summaries comprising estimates and standard errors of a log(odds ratio) or a log(hazard ratio), and these can be interpreted as providing normal likelihoods and incorporated into a Bayesian analysis.

To be specific, suppose we have a classical estimate y, with standard error s, of a true log(odds ratio) or a log(hazard ratio) θ, and this is interpreted as providing a normal likelihood based on the sampling distribution $y \sim N[\theta, \sigma^2/m]$, where $s = \sigma/\sqrt{m}$, with σ known. Then if we are willing to approximate our prior distribution for θ by $\theta \sim N[\mu, \sigma^2/n_0]$, the simplest application of Bayes theorem gives a posterior distribution

$$\theta|y \sim N\left[\frac{n_0\mu + my_m}{n_0 + m}, \frac{\sigma^2}{n_0 + m}\right].$$

The choice of σ is essentially one of convenience since we are matching three parameters (σ, m, n_0) to two specified quantities (the likelihood and prior variability), but perhaps remarkably it turns out that $\sigma = 2$ leads to a value m that is generally interpretable as the "effective number of events" (Spiegelhalter, Abrams and Myles, 2004): for example, Tsiatis (1981) shows that in a balanced trial with small treatment effect, the estimated log(hazard ratio) has approximate variance $4/m$, where m is the observed number of events. This formulation aids interpretation as one can translate both prior input and evidence from data on a common scale, as we shall see in Section 2.4.

2.3 Flexible Prior Specification

For a "classical" audience, it is important to clarify a number of possible misconceptions that may arise concerning the prior distribution. In particular, a prior is not necessarily specified beforehand: Cox (1999) 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 specified, although even then it is reasonable to carry out sensitivity analysis to alternative choices.

The prior is also not necessarily unique, since the discussion in Section 2.1 should make clear that there is no such thing as the "correct" prior. Instead, Kass and Greenhouse (1989) introduced the term "community of priors" to describe the range of viewpoints that should be considered when interpreting evidence, and therefore a Bayesian analysis is best seen as providing a mapping from a space of specified prior beliefs to appropriate posterior beliefs.

Members of this "community" may include the following:

- "*Clinical*" *priors representing expert opinion*—Elicitation methods for such priors were reviewed by Chaloner (1996), who concluded 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. Recent reports of elicitation before clinical trials include Fayers et al. (2000) and Chaloner and Rhame (2001).
- "*Evidence-based*" *priors representing a synthesis of available evidence*—Since conclusions strongly based on beliefs that cannot be "objectively" supported are unlikely to be widely regarded as convincing, it is valuable to summarize available evidence. Possible models for incorporation of past data are discussed in Section 3.2.
- "*Reference*" *priors*—It is attractive to seek a "noninformative" prior to use as a baseline analysis, and such analyses have been suggested as a way of making probability statements about parameters without being explicitly Bayesian (Burton, 1994; Shakespeare, Gebski, Veness and Simes, 2001). But the problems are well known: uniform priors on one scale are not uniform on a transformed scale, and apparently innocuous prior assumptions can have a strong impact particularly when events are rare. Special problems arise in hierarchical modelling, both with regard to appropriate priors on nuisance parameters such as baseline risks, and selection of a default prior for the between-group variability. For the latter, attention has concentrated on placing a prior directly on the degree of shrinkage (Christiansen and Morris, 1997b; Daniels, 1999; Natarajan and Kass, 2000; DuMouchel and Normand, 2000; Spiegelhalter, 2001), although a half-normal prior on the between-group standard deviation appears to be a transparent and flexible means of incorporating a degree of prior information (Spiegelhalter, Abrams and Myles, 2004).
- "*Sceptical*" *priors that express archetypal doubts about large effects*—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 (1989) 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 a shape chosen to include plausible treatment differences which determines the degree of scepticism. Spiegelhalter, Freedman and Parmar (1994) argue that a reasonable degree of scepticism may correspond to a feeling that the trial has been designed around an alternative hypothesis that is optimistic, formalized by a prior with only a small probability γ (say 5%) that the treatment effect is as large as the alternative hypothesis θ_A.

In Section 2.2 we emphasized how a "significant" positive trial result may be tempered by taking into account prior prevalence, and Matthews (2001) extended those ideas to allow for the full likelihood observed. Specifically, he derives a simple formula for working backward from the observed likelihood to the sceptical prior centered on 0 that would just

give a 95% posterior interval that included 0—if that degree of scepticism were considered plausible, then the trial results could not be considered as convincing. An example is provided in Section 2.4.

Sceptical priors have been used in a number of case studies (Fletcher et al., 1993; Parmar, Ungerleider and Simon, 1996; DerSimonian, 1996; Heitjan, 1997; Dignam et al., 1998; Cronin et al., 1999; Harrell and Shih, 2001). A senior Food and Drug Administration (FDA) biostatistician (O'Neill, 1994) has stated that he

> would like to see [sceptical priors] applied in more routine fashion to provide insight into our decision making.

- "*Enthusiastic*" *priors that express archetypal optimism*—As a counterbalance to the pessimism expressed by the sceptical prior, Spiegelhalter, Freedman and Parmar (1994) suggest an "enthusiastic" prior centered on the alternative hypothesis and with a low chance (say 5%) that the true treatment benefit is negative.

The community of prior opinions becomes particularly important when faced with the difficult issue of whether to stop a clinical trial. Kass and Greenhouse (1989) express the crucial view 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 this idea may be formalized as follows:

1. Stopping with a "positive" result (i.e., in favor of the new treatment) might be considered if a posterior based on a *sceptical* prior suggested a high probability of treatment benefit.
2. Stopping with a "negative" result (i.e., equivocal or in favor of the standard treatment) 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 they are wrong. Fayers, Ashby and Parmar (1997) provide a tutorial on such an approach, and Section 2.4 describes its application by a data monitoring committee of two cancer trials.

2.4 Example 1: Bayesian Monitoring of the CHART Trials

Here we illustrate the selected ideas described previously in the context of two clinical trials in which the data monitoring committee used Bayesian techniques to inform the decision whether to stop early. More detail is provided in Parmar et al. (2001) and Spiegelhalter, Abrams and Myles (2004).

Intervention. In 1986 a new radiotherapy technique called CHART (continuous hyper-fractionated accelerated radiotherapy) 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 twelve days). There are clearly considerable logistical problems in efficiently delivering CHART, as well as concerns about possible increased side effects.

Aim of studies. Promising nonrandomized and pilot studies led the U.K. Medical Research Council to instigate two large randomized trials to compare CHART to conventional radiotherapy in both nonsmall-cell lung and head-and-neck cancer, and in particular to assess whether CHART provides a clinically important difference in survival that compensates for any additional toxicity and problems of delivering the treatment.

Study design. The trials began in 1990, randomized in the proportion 60 : 40 in favor of CHART, 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.

Outcome measure. Full data were to become available on survival (lung) or disease-free survival (head-and-neck), with results presented in terms of estimates of the hazard ratio h, defined as the ratio of the hazard under CHART to the hazard under standard treatment. Hence hazard ratios less than 1 indicate superiority of CHART.

Planned sample sizes. 600 patients were to be entered in the lung cancer trial, with 470 expected deaths, giving 90% power to detect at the 5% level a 10% improvement (15% to 25% survival). Under a proportional hazards assumption, this is equivalent to an alternative hypothesis (hazard ratio) of $h_A = \log(0.25)/\log(0.15) = 0.73$. The head-and-neck trial was to have 500 patients, with 220 expected recurrences, giving 90% power to detect at the 5% level a 15% improvement (45% to 60% disease-free survival), equivalent to an alternative hypothesis of $h_A = \log(0.60)/\log(0.45) = 0.64$.

Statistical model. A proportional hazards Cox model provides an approximate normal likelihood (Section 2.2) for $\theta = \log(h) = \log(\text{hazard ratio})$, based on

$$y_m \sim N\left[\theta, \frac{\sigma^2}{m}\right],$$

where y_m is the estimated log(hazard ratio), $\sigma = 2$ and m is the "equivalent number of events" in a trial balanced in recruitment and follow-up.

Prospective analysis? Yes, the prior elicitations were conducted before the start of the trials, and the Bayesian results presented to the DMC at each of their meetings.

Prior distributions. Although the participating clinicians were enthusiastic about CHART, there was considerable scepticism expressed by oncologists who declined to participate in the trial. Eleven opinions were elicited for the lung cancer trial and nine for the head-and-neck (Spiegelhalter, Freedman and Parmar, 1994), using a questionnaire described in detail in Parmar, Spiegelhalter and Freedman (1994). We use the arithmetic average of the distributions as a summary, since we wish to represent an "average" clinician. The prior distribution expressed a median anticipated 2-year survival benefit of 10%, and a 10% chance that CHART would offer no survival benefit at all. The histogram was then transformed to a log(hazard ratio) scale assuming a 15% baseline survival and a $N[\mu, \sigma^2/n_0]$ distribution fitted, giving $\mu = -0.28$, $\sigma = 2$, $\sigma/\sqrt{n_0} = 0.23$, which implies $n_0 = 74.3$, so the prior evidence is equivalent to that provided by a trial in which around 74 deaths had been observed, balanced equally between arms. For the head-and-neck trial, the fitted prior mean $\log(h)$ is $\theta_0 = -0.33$ with standard deviation 0.26, equivalent to $n_0 = 61.0$. Figure 1 shows the fit of the normal distributions to the transformed histograms are quite reasonable, and the similarity between the two sets of opinions is clear, each supporting around a 25% reduction in hazard, but associated with considerable uncertainty. A *sceptical prior* was also derived using the ideas in Section 2.3: 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). For the lung trial, the alternative hypothesis is $\theta_A = \log(0.73) = -0.31$, so assuming $\sigma = 2$ gives $n_0 = 110$. For the head-and-neck, the alternative hypothesis is $\theta_A = \log(0.64) = -0.45$, which gives a sceptical prior with $n_0 = 54$. These sceptical prior distributions are displayed in Figure 2 with the clinical priors derived above.

Evidence from study. For the lung cancer trial, the data reported at each of the annual meetings of the independent data monitoring committee is shown in Table 1 (Parmar et al., 2001): the final row is that of the published analysis. 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. Perhaps notable is that in 1993 the DMC recommended continuation of the trial when the 2-sided P-value was 0.001.

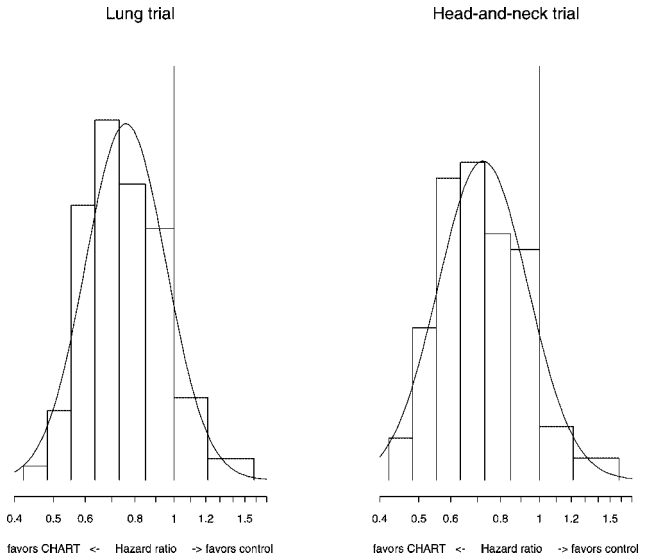

FIG. 1. *Average opinion for lung cancer and head-and-neck CHART trials with normal distributions fitted with matching mean and variance.*

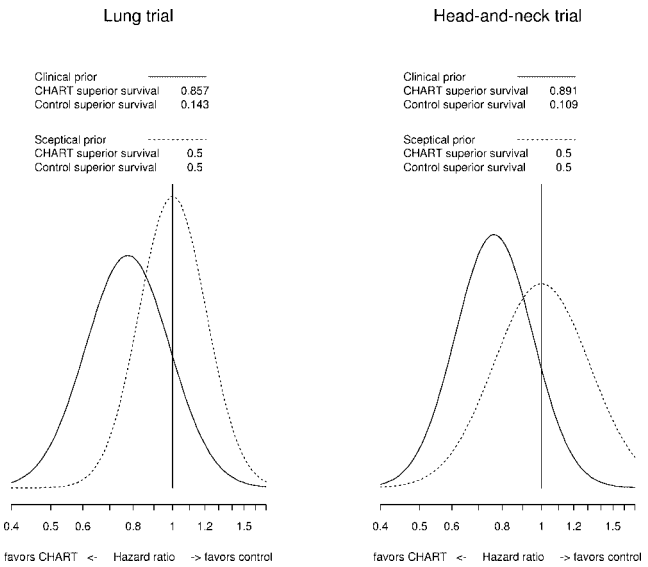

FIG. 2. *Sceptical and clinical priors for both lung and head-and-neck CHART trials, showing prior probabilities that CHART has superior survival. The sceptical priors express a* 5% *prior probability that the true benefit will be more extreme than the alternative hypotheses of* $h = 0.73$ *for the lung trial and* $h = 0.64$ *for the head-and-neck trial.*

TABLE 1

Summary data reported at each meeting of the CHART lung trial DMC. The "effective number of deaths" m is derived from the likelihood-based 95% *interval, in that the standard error of the estimated* log(*hazard ratio*) *is assumed to be* $2/\sqrt{m}$

Date	No. patients	Actual deaths	Effective deaths m	Classical hazard ratio		2-sided P-value	"Sceptical" hazard ratio	
				Estimate	95% interval		Estimate	$P(h < 0.80)$
1992	256	78	76	0.55	(0.35–0.86)	0.007	0.79	0.56
1993	380	192	190	0.63	(0.47–0.83)	0.001	0.73	0.73
1994	460	275	253	0.70	(0.55–0.90)	0.003	0.78	0.60
1995	563	379	346	0.75	(0.61–0.93)	0.004	0.80	0.48
1996	563	444	488	0.76	(0.63–0.90)	0.003	0.81	0.52

For the head-and-neck cancer trial, the data reported at each meeting of the independent data monitoring committee showed no strong evidence of benefit shown at any stage of the study: at the final analysis the likelihood-based hazard ratio estimate was 0.95 with 95% interval 0.79 to 1.14.

Bayesian interpretation. For the lung trial, the DMC were presented with survival curves, and posterior distributions and tail areas arising from a reference prior [uniform on a log(h) scale]. Following the discussion in Section 2.3, the posterior distribution resulting from the sceptical prior was emphasized in view of the positive findings, in order to check whether the evidence was sufficient to persuade a reasonable sceptic.

Figure 3 shows the sceptical prior distributions at the start of the lung cancer trial, and the likelihood (essentially the posterior under the reference prior) and posterior for the results available in subsequent years. Under the reference prior there is substantial reduction in the estimated effect as the trial progresses, while the sceptical results are remarkably stable and Table 1 shows that the initial estimate in 1992 remains essentially unchanged.

Before the trial the clinicians were demanding a 13.5% improvement before changing treatment (Parmar et al., 2001): however, the inconvenience and toxicity were found to be substantially less than expected and so the probability of improvement of at least 7% was calculated, around half the initial demands, and equivalent to h of around 0.80. Such "shifting of the goalposts" is entirely reasonable in the light of the trial data. Figure 3 and the final column of Table 1 show that the sceptical posterior distribution is centered around these clinical demands, showing that these data should persuade even a sceptic that CHART both improves survival and, on balance, is the pragmatic treatment of choice.

Since the results for the head-and-neck trial were essentially negative, it is appropriate to monitor the trial assuming a more enthusiastic prior: we adopt the experts's clinical prior, since this expresses considerable optimism. The initial clinical demands were a 13% change in survival from 45% to 58%, but in parallel with the lung trial we have reduced this to a 7% improvement. The results are described in detail in Parmar et al. (2001): the final posterior expresses a 17% chance that CHART reduces survival, 8% chance that it reduces a clinically significant ($> 7\%$) improvement and 75% chance that the effect lies in the "grey area" in between. The data should therefore be sufficient to convince a reasonable enthusiast that, on the basis of the trial evidence, CHART is not of clinical benefit in head-and-neck cancer.

Sensitivity analysis. We can use the results of Matthews (2001) to see what degree of scepticism would have been necessary not to have found the final lung results convincing. Had the effective number of events underlying our sceptical prior been $n_0 = 701$, this would have just led the 95% posterior interval for the hazard ratio to include 1. Since this prior distribution would have restricted plausible hazard ratios to a 95% prior interval of 0.86 to 1.14, this can be considered too great a degree of reasonable scepticism and hence the trial results can be considered convincing of survival benefit.

Comments. There are two important features of the prospective Bayesian analysis of the CHART trial. First, while classical stopping rules may well have led the DMC to have stopped the lung trial earlier, perhaps in 1993 when the two-sided P-value was 0.001 this would have overestimated the benefit. The DMC allowed the trial to continue, and consequently produced a strong result that should be convincing to wide range of opinion. Second, after discovering that the secondary aspects of the new treatment were less unfavorable than expected, the DMC is allowed to "shift the goalposts" and not remain with unnecessarily strong clinical demands.

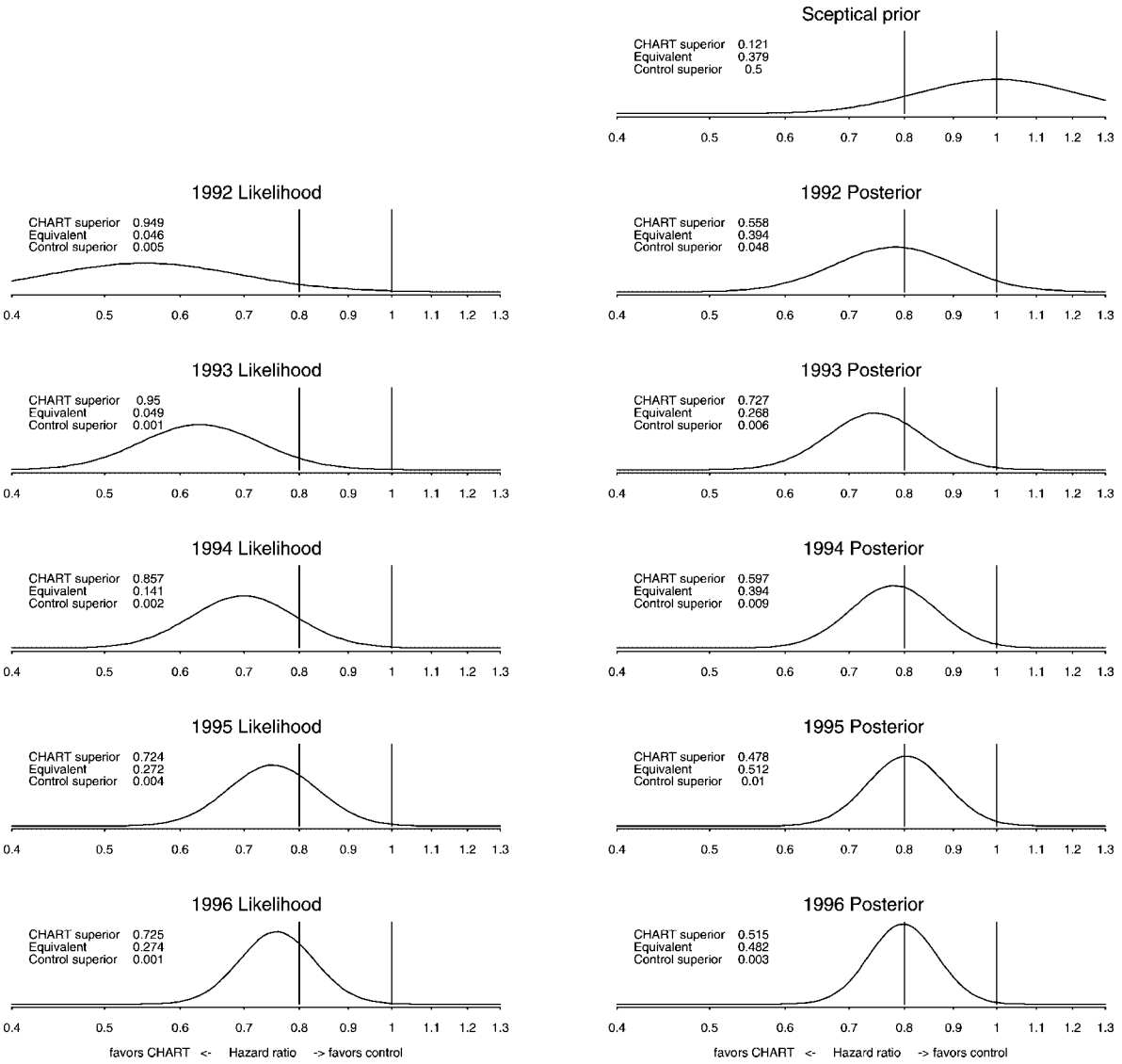

FIG. 3. *Prior, "likelihood" (posterior based on reference prior) and posterior distributions for the CHART lung cancer trial assuming a sceptical prior. The likelihood becomes gradually less extreme, providing a very stable posterior estimate of the treatment effect when adopting a sceptical prior centered on a hazard ratio of* 1. *Demands are based on a* 7% *improvement from* 15% *to* 22% 2*-year survival, representing a hazard ratio of* 0.80.

3. THREE SELECTED FEATURES OF COMPLEX BAYESIAN MODELLING

It is important to note that many of the advantages claimed for the Bayesian approach follow from the ability to handle complex models. In particular, there has been extensive use of hierarchical models in health-care evaluation: see, for example, applications in subset analysis (Dixon and Simon, 1991; Simon, Dixon and Friedlin, 1996), multicenter analysis (Gray, 1994; Stangl and Greenhouse, 1998), cluster randomized trials (Spiegelhalter, 2001; Turner, Omar and Thompson, 2001), multiple N-of-1 studies (Zucker et al., 1997), institutional comparisons (Goldstein and Spiegelhalter, 1996; Christiansen and Morris, 1997a; Normand, Glickman and Gatsonis, 1997) and meta-analysis (Sutton et al., 2000; Whitehead, 2002). However, many of these analyses minimize the role of prior information and could have been carried out using flexible likelihood methods, such as simulating the distribution of functions of maximum likelihood estimates and so on.

Here we focus on three aspects that reflect a specifically Bayesian input into complex modelling: computation; incorporation of historical information, and inference on complex functions of parameters.

3.1 Computation

It is perhaps extraordinary that the Bayesian paradigm, for so long held up as being impractical to full implement, has become through simulation methodology the easiest framework in which to carry out inference in complex models. There is no need here to describe the power of Markov chain Monte Carlo methods for approximating required integrals using simulated values from the posterior distribution: tutorial introductions are provided by Brooks (1998), Casella and George (1992) and Gilks, Richardson and Spiegelhalter (1996).

It may, however, be important to acknowledge the continuing role of simpler Monte Carlo methods in certain contexts, in which quantities are simulated from distributions expressing current uncertainty, and then complex functions of these quantities calculated, often using standard spreadsheet software. The resulting distributions of the outputs of the spreadsheet will reflect the uncertainty about the inputs. This use of Monte Carlo methods can also be termed *probabilistic sensitivity analysis* and is used extensively in the context of risk analysis and cost-effectiveness modelling. A schematic representation is shown in Figure 4(a), where it is termed the "two-stage" approach since the two stages of producing the probability distributions, and then propagating their effects, are separated. This is contrasted to the "integrated" approach in Figure 4(b) which is generally implemented using MCMC methods.

Advantages of the integrated approach include the following (Spiegelhalter and Best, 2003). First, there is no need to assume parametric distributional shapes for the posterior probability distributions, which may be important for inferences for smaller samples. Second, and perhaps most important, the appropriate probabilistic dependence between unknown quantities is propagated (Chessa et al., 1999), rather than assuming either independence or being forced into, for example, multivariate normality. This can be particularly vital when propagating inferences which are likely to be strongly correlated, say when considering both baseline levels and treatment differences estimated from the same studies.

Disadvantages of the integrated approach are its additional complexity and the need for full Markov chain Monte Carlo software. The "two-stage" approach, in contrast, might be implemented, for example, as macros for Excel, either from commercial software

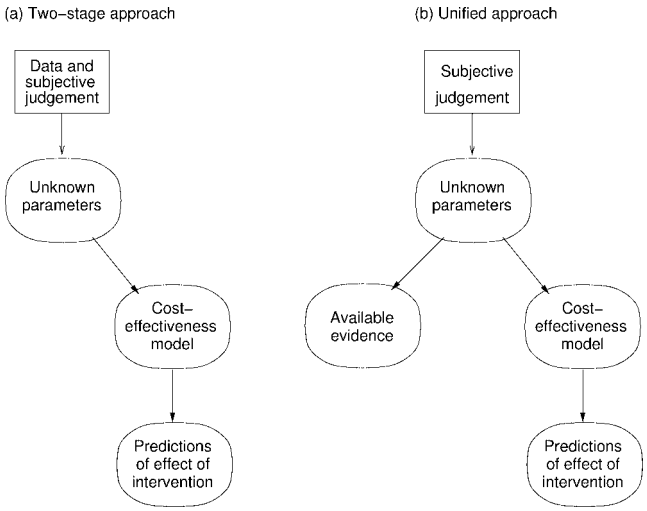

FIG. 4. *A schematic graph showing the two approaches to incorporating uncertainty about parameters into a cost-effectiveness analysis. The* (a) *two-stage approach subjectively synthesizes data and judgement to produce a prior distribution on the parameters which is then propagated through the cost-effectiveness model. The* (b) *unified or integrated approach adopts a fully Bayesian analysis*: *after taking into account the available evidence, initial prior opinions on the parameters are revised by Bayes theorem to posterior distributions, the effects of which are propagated through the cost-effectiveness model in order to make predictions. An integrated Bayesian approach ensures that the full joint uncertainty concerning the parameters is taken into account.*

such as @RISK (Palisade Europe, 2001) and Crystal Ball (Decisioneering, 2000), or self-written. However, experience with such spreadsheets suggests that they might not be particularly transparent for complex problems, due to clumsy handling of arrays and opaque formula equations.

3.2 Incorporating Historical Data

The need for using historical data has been considered in a variety of contexts, such as exploiting historical controls in randomized trials, modelling the potential biases in observational studies and pooling data from many sources in an evidence synthesis. Within the Bayesian framework all these can be formalized as a means of using past evidence as a basis for a prior distribution for a parameter of interest. Suppose, for example, we have historical data $y_1, \ldots, y_H$, each assumed to depend on a parameter θ_h, $h = 1, \ldots, H$. Numerous options are available for specifying the relationship between the θ_h's and θ, the parameter of interest, and a basic structure is provided in Figure 5:

(a) *Irrelevance*—The historical data provides no relevant information, so that each θ_h is unrelated to θ.

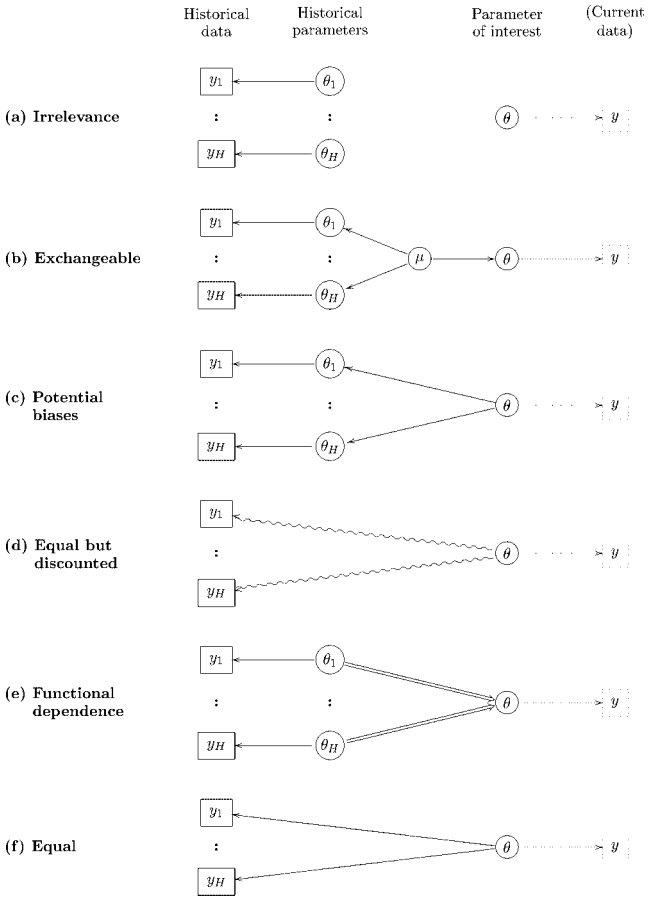

FIG. 5. *Different assumptions relating parameters underlying historical data to parameter of current interest*: *single arrows represent a distribution*; *double arrows represent logical functions*; *wobbly arrows represent discounting*.

(b) *Exchangeable*—Current and past studies are judged "similar" to the extent that θ_h, $h = 1, \ldots, H$, and θ are assumed exchangeable: for example, θ_h, $\theta \sim N[\mu, \tau^2]$. Exchangeability is a strong assumption, but if this is judged reasonable, then it provides a simple model to use databases to provide prior distributions (Gilbert, McPeek and Mosteller, 1977). For example, Lau, Schmid and Chalmers (1995) and DerSimonian (1996) use cumulative random-effects metaanalysis to provide a prior distribution for a subsequent trial, while Gould (1991) suggests using past trials to augment current control group information, by assuming exchangeable control groups.

Models can become more complex when we wish to synthesize evidence from different study "types," say randomized, case-control or cohort studies: this naturally leads to hierarchical exchangeability assumptions, which can specifically allow for the quantitative within- and between-"study-type" heterogeneity, and incorporate prior beliefs regarding qualitative differences between the various sources of evidence. Examples of this approach include Prevost, Abrams and Jones (2000), who pool randomized and nonrandomized studies on breast cancer screening, and Larose and Dey (1997), who similarly assume open and closed studies are exchangeable, while Dominici, Parmigiani, Wolpert and Hasselblad (1999) examine migraine trials and pool open and closed studies of a variety of designs in a four-level hierarchical model. There is a clearly a difficulty in making such exchangeability assumptions, since there are few study-types and hence little information on the variance component.

(c) *Potential biases*—Past studies may be biased, either through lack of quality (internal bias) or because the setting is such that the studies are not precisely measuring the underlying quantity of interest (external bias), or both: Eddy, Hasselblad and Shachter (1992) identify a range of such sources of bias and argue that their magnitudes may be modelled and the historical results appropriately adjusted. A common choice is the existence of a simple bias δ_h so that $\theta_h = \theta + \delta_h$, and a number of choices may be made about the distribution of δ_h. For example, Brophy and Joseph (2000) consider possible sources of bias when using past trials to create a prior for the GUSTO trial, while Pocock (1976) assumes a bias with prior mean 0 when incorporating a group of historical controls into a clinical trial.

Such models are clearly potentially controversial, and careful sensitivity analysis is essential. However, we note the increasing research concerning the quantitative bias of observational studies: see, for example, Kunz and Oxman (1998), Britton et al. (1998), Benson and Hartz (2000), Ioannidis et al. (2001), Reeves et al. (2001) and Sanderson et al. (2001).

(d) *Equal but discounted*—Past studies may be assumed to be unbiased, but their precision decreased in order to "discount" past data. In the context of control groups, Kass and Greenhouse (1989) 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." Ibrahim and Chen (2000) suggest the "power" prior, in which we assume $\theta_h = \theta$, but discount

the historical evidence by taking its likelihood $p(y_h|\theta_h)$ to a power α. For example, Greenhouse and Wasserman (1995) downweight a previous trial with 176 subjects to be equivalent to only 10 subjects: Fryback, Stout and Rosenberg (2001) also discounted past trials to create a prior for the GUSTO analysis. We note, however, that Eddy, Hasselblad and Shachter (1992) are very strong in their criticism of this method, as it does not have any operational interpretation and hence no clear means of assessing a suitable value for α.

(e) *Functional dependence*—The current parameter of interest is a logical function of parameters estimated in historical studies: this option is further explored in Section 3.3.

(f) *Equal*—Past studies are measuring precisely the parameters of interest and data can be directly pooled—this is equivalent to assuming exchangeability of individuals.

Various combinations of these techniques are possible. For example, Berry and Stangl (1996a) assume a fixed probability p that each historical patient is exchangeable with those in the current study [i.e., either option (f) (complete pooling) with probability p or option (a) (complete irrelevance) with probability $1 - p$], while Racine, Grieve, Fluhler and Smith (1986) assume a certain prior probability that the entire historical control group exactly matches the contemporaneous controls and hence can be pooled.

Given the wide range of options concerning the way in which historical data may be incorporated into a model, there is clearly a need for both qualitative and quantitative input into the modelling, based on both judgements and substantive knowledge.

3.3 Inference on Complex Functions

This section expands on option (e) of Figure 5: where we establish a functional relationship between the parameter of interest and past data.

This could arise in the following context. Suppose that a number of experimental interventions are investigated in a series of studies, where each study compares a subset of the interventions with a control group. We would like to draw inferences on the treatment effects compared with control and possibly also make comparisons between treatments that may well have not ever been directly compared "head-to-head." We can call these *indirect* comparisons, although the term *mixed* comparisons has also been used. Higgins and Whitehead (1996) and Hasselblad (1998) consider a range of hierarchical models for this problem, while Song, Altman, Glenny and Deeks (2003) carry out an empirical investigation and report that such comparisons arrive at essentially the same conclusions as head-to-head comparisons. A specific application arises in the context of "active control" studies. Suppose an established treatment C exists for a condition, and a new intervention T is being evaluated. The efficacy of T would ideally be estimated in randomized trial with a placebo P as the control group, but because of the existence of C this may be considered unethical. Hence C may be used as an "active control" in a head-to-head clinical trial, and inferences about the efficacy of T may have to be estimated indirectly, using past data on C-versus-P comparisons.

A more complex situation is as follows. Suppose we are interested in drawing inferences on a quantity f about which no direct evidence exists, but where f can be expressed as a deterministic function of a set of "fundamental" parameters $\theta = \theta_1, \ldots, \theta_N$. For example, f might be the response rate in a new population made up of subgroups about which we do have some evidence. More generally, we might assume we have available a set of K studies in which we have observed data $y_1, \ldots, y_K$ which depend on parameters $\psi_1, \ldots, \psi_K$, where each ψ_k is itself a function of the fundamental parameters θ. This structure is represented graphically in Figure 6. This situation sounds very complex but in fact is rather common, when we have many studies, each of which informs part of a jigsaw, and which need to be put together to answer the question of interest. An example is provided in Section 3.4.

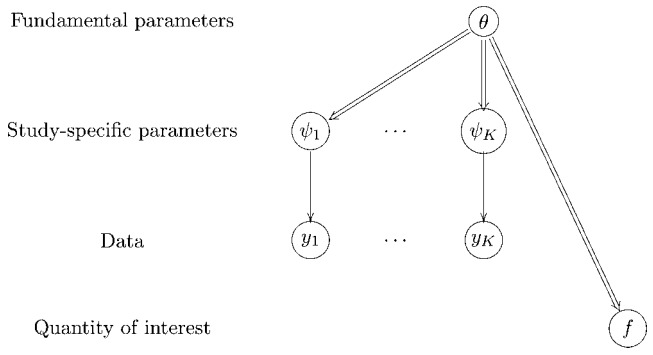

FIG. 6. *Data y_k in each of K studies depends on parameters ψ_k, which are known functions of fundamental parameters θ. We are interested in some other function f of θ, and so need to propagate evidence from the y_k's.*

3.4 Example 2: Cost-Effectiveness of Alternative Strategies for Pre-Natal HIV Testing

This example, derived from Ades and Cliffe (2002), follows the "integrated" approach of Section 3.1, simultaneously conducting a complex evidence synthesis and propagating the results through a cost-effectiveness model.

Intervention. Ades and Cliffe (2002) examine alternative strategies for screening for HIV in pre-natal clinics: *universal* screening of all women, or *targeted* screening of current injecting drug users (IDU) or women born in sub-Saharan Africa (SSA).

Aim of study. To determine the optimal policy taking into account the costs and benefits; however, Ades and Cliffe (2002) point out that the formulation is not wholly realistic as the decision to screen universally throughout England has now been taken, and in any case a strategy of targeted testing may not be politically acceptable.

Study design. Synthesis of multiple sources of evidence to estimate parameters of the epidemiological model shown in Figure 7; however, direct evidence is only available for a limited number of the fundamental parameters.

Outcome measure. SSA and IDU women will be screened under both universal and targeted strategies, and hence the only difference between the strategies comprise the additional tests and additional cases detected in the non-SSA, non-IDU group. Additional tests per 10,000 women comprise those on non-SSA, non-IDU women who are not already diagnosed, and so the rate is given by $10{,}000(1-a-b)(1-eh)$. The rate of new HIV cases detected is $10{,}000(1-a-b)e(1-h)$.

Statistical model and evidence from study. Table 2 summarizes the data sources available; full details and references are provided by Ades and Cliffe (2002), who also describe their efforts to select sources which are as "independent" as possible.

The crucial aspect is that there is no direct evidence concerning the vital parameters e and h for the low-risk group, and hence their values must be inferred indirectly from other studies. For this reason the parameter w is introduced which is not part of the epidemiological model: under the assumption that the low-risk group has the same prevalence of subtype B as SSA women, and that all IDU women are subtype B, allows use of data source 12 on non-SSA women.

Prior distributions. Uniform priors for all proportions are adopted.

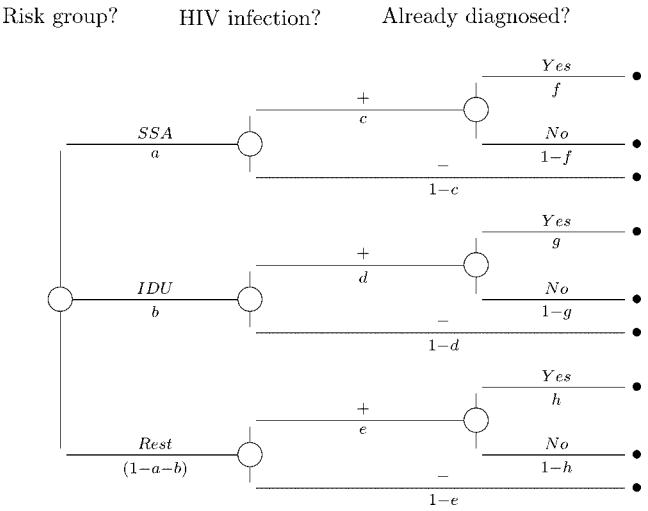

FIG. 7. *Probability tree showing how the proportions of women in different risk groups can be constructed.*

Computation and software. Markov chain Monte Carlo (MCMC) methods were implemented using WINBUGS.

Sensitivity analyses. In this section we focus on the consistency of data sources rather than the usual sensitivity analysis to model assumptions. We have synthesized all available data but the results may be misleading if we have included data that does not fit our assumed model. A simple way of assessing possible conflict is to compare the observed proportion in the 12 sources with that fitted by the model, and it is apparent from Table 2 that the observation for source 4 is only just included in the 95% interval, while the data for source 12 lies wholly outside its estimated interval. This is only a crude method, since a source may strongly influence its estimate, so a better procedure is to leave each source out in turn, reestimate the model and then predict the data we would expect in a source that size. This predictive distribution, easily obtained using MCMC methods, is then compared to the observed data and a "cross-validatory" P-value calculated.

Removing data-source 4 from the analysis leads to the cross-validatory P-values shown in the final column of Table 2. The small P-value for source 4 shows its lack of consistency with the remaining data, whereas the predictions for the remaining data seem quite reasonable. Removing source 4 from the analysis leads to an estimate of 8,810 (8,717–8,872) for additional tests per 10,000, and 2.73 (1.31–4.12) for additional cases, so the removal of this divergent source does not in fact have much influence on the conclusions.

TABLE 2

Available data from relevant studies, generally only allowing direct estimation of functions of fundamental parameters of interest. Also provided are estimates and intervals based on full data, and the cross-validatory P-values based on excluding data source 4

Data items and sources	Parameter being estimated	Data	Observed proportion	Estimate	95% interval	P-value (excl. 4)
1 Proportion born in sub-Saharan Africa (SSA), 1999	a	11,044/104,577	0.106	0.106	0.104–0.108	0.47
2 Proportion IDU last 5 years	b	12/882	0.0137	0.0088	0.0047–0.149	0.46
3 HIV prevalence, women born in SSA, 1997–1998	c	252/15,428	0.0163	0.0172	0.0155–0.0189	0.27
4 HIV prevalence in female IDU's, 1997–1999	d	10/473	0.0211	0.0120	0.0062–0.0219	0.004
5 HIV prevalence, women not born in SSA, 1997–1998	$\frac{db+e(1-a-b)}{(1-a)}$	74/136,139	0.000544	0.000594	0.000478–0.000729	0.35
6 Overall HIV seroprevalence in pregnant women, 1999	$ca+db+e(1-a-b)$	254/102,287	0.00248	0.00235	0.00217–0.00254	0.21
7 Diagnosed HIV in SSA women as a proportion of all diagnosed HIV, 1999	$\frac{fca}{fca+gdb+he(1-a-b)}$	43/60	0.717	0.691	0.580–0.788	0.50
8 Diagnosed HIV in IDU's as a proportion of non-SSA diagnosed HIV, 1999	$\frac{gdb}{gdb+he(1-a-b)}$	4/17	0.235	0.298	0.167–0.473	0.40
9 Overall proportion HIV diagnosed	$\frac{fca+gdb+he(1-a-b)}{ca+db+e(1-a-b)}$	87/254	0.343	0.350	0.296–0.408	0.47
10 Proportion of infected IDU's diagnosed, 1999	g	12/15	0.800	0.747	0.517–0.913	0.44
11 Proportion of serotype B in infected women from SSA, 1997–1998	w	14/118	0.119	0.111	0.065–0.171	0.43
12 Proportion of serotype B in infected women not from SSA, 1997–1998	$\frac{db+we(1-a-b)}{db+e(1-a-b)}$	5/31	0.161	0.285	0.201–0.392	0.23
Additional tests per 10,000	$10{,}000(1-a-b)(1-eh)$			8856	8,789–8,898	
Additional HIV cases detected	$10{,}000(1-a-b)e(1-h)$			2.49	1.09–3.87	

Costs and utilities. Ades and Cliffe (2002) specify the cost per test as $T = £3$, and the net benefit K per maternal diagnosis is judged to be around £50,000 with a range of £12,000 to £60,000. In this instance there is explicit monetary net benefit from maternal diagnosis and so it may be reasonable to take K as an unknown parameter, and Ades and Cliffe (2002) perform a probabilistic sensitivity analysis by giving K a somewhat complex prior distribution. In contrast, we prefer to continue to treat K as a willingness-to-pay for each unit of benefit, and therefore we conduct a deterministic sensitivity analysis in which K is varied up to £60,000.

The pre-natal population in London is $N = 105{,}000$, and hence the annual incremental net benefit (INB) of implementing full rather than targeted screening is

$$\text{INB} = N(1-a-b)\big(Ke(1-h) - T(1-eh)\big).$$

We would also like to know, for fixed K, the probability $Q(K) = P(\text{INB} > 0|\text{data})$: when plotted as a function of K this is known as the cost-effectiveness acceptability curve (CEAC); see, for example, Briggs (2000) and O'Hagan, Stevens and Montmartin (2000, 2001) for detailed discussion of these quantities from a Bayesian perspective.

We can also conduct a "value of information" analysis (Claxton, Lacey and Walker, 2000). For some unknown quantity θ, the "value of perfect information" VPI(θ) is defined as the amount we would gain by knowing θ exactly: VPI(θ) is 0 when INB(θ) > 0, and $-$INB(θ) when INB(θ) < 0, and hence can be expressed as

$$\mathrm{VPI}(\theta) = \max\bigl(-\mathrm{INB}(\theta), 0\bigr).$$

Hence the "expected value of perfect information" EVPI is

$$\mathrm{EVPI} = \mathrm{E}\bigl[\max\bigl(-\mathrm{INB}(\theta), 0\bigr)|\mathrm{data}\bigr]. \tag{1}$$

This may be calculated in two ways: first using MCMC methods, and second by assuming a normal approximation to the posterior distribution of INB(K) and using a closed form identity. Taking a 10-year horizon and discounting at 6% per year gives a multiplier of 7.8 (not discounting the first year) to the annual figure.

Bayesian interpretation. Following the previous findings the analysis is conducted without data-source 4. Figure 8(a) shows the normal approximations to the posterior distributions of INB for different values of K. The expected INB and 95% limits are shown in Figure 8(b) for K up to £60,000, indicating that the policy of universal testing is preferred on balance provided that the benefit K from a maternal diagnosis is greater than around £10,000: K is certainly judged to exceed this value. The cost-effectiveness acceptability curve in Figure 8(c) points a high probability of universal testing being cost-effective for reasonable values of K. Figure 8(d) shows the EVPI (± 2 Monte Carlo errors) calculated using 100,000 MCMC iterations and also using the approximation to the distribution of INB, which provides an adequate approximation. The EVPI

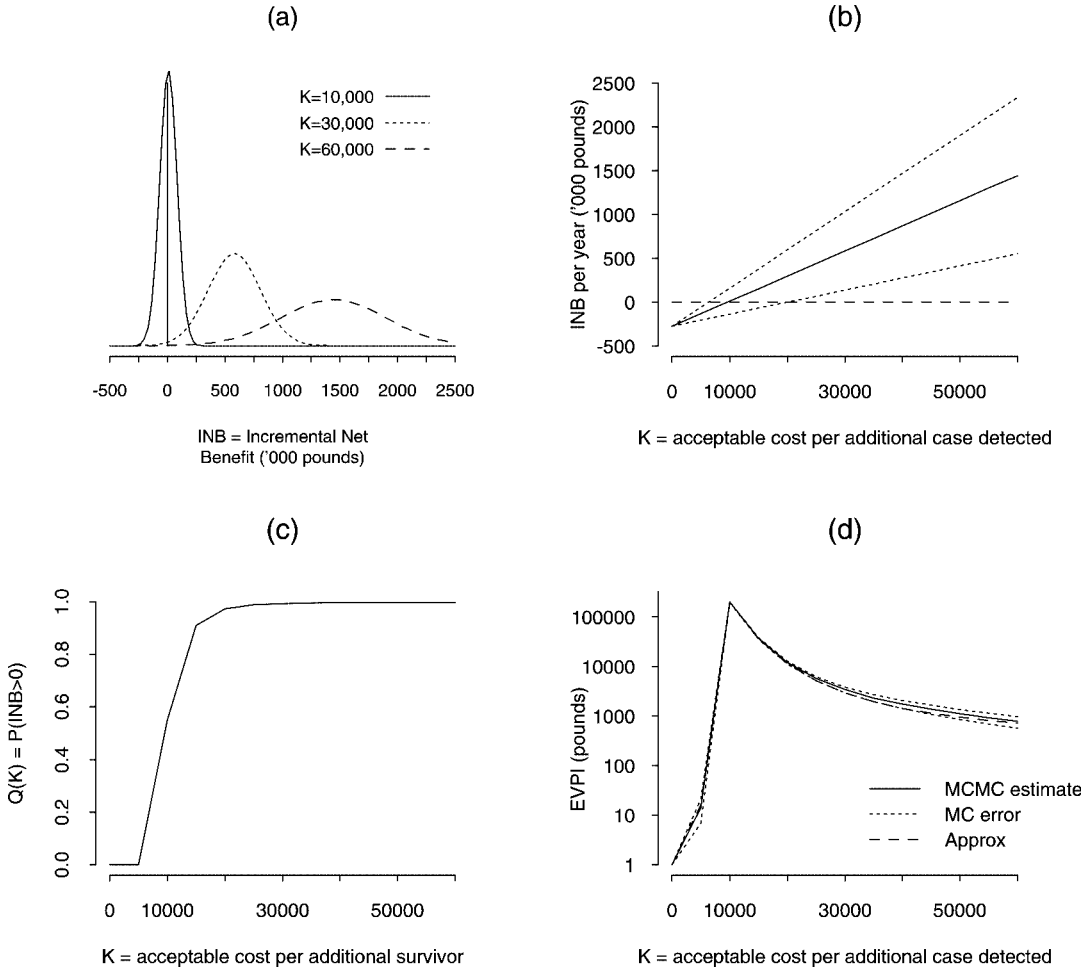

FIG. 8. (a) *and* (b) *incremental net benefits*; (c) *cost-effectiveness acceptability curve*; *and* (d) *expected value of perfect information for universal versus targeted prenatal testing for HIV. Note the EVPI is maximized at the threshold value of K at which the optimal decision changes.*

is substantial for low values of K, but for values around £50,000 the EVPI is negligible. Hence there appears to be little purpose in further research to determine the parameters more accurately.

4. CONCLUSIONS

4.1 Current Status of Bayesian Methods

As mentioned in Section 1, there has been a major growth in Bayesian publications but these mainly comprise applications of complex modelling. A notable exception is the use of Bayesian models in cost-effectiveness analysis, in which informative prior distributions may be based on a mixture of evidence synthesis and judgement (O'Hagan and Luce, 2003).

When considering health-care evaluations one cannot ignore the regulatory framework which controls the release onto the market of both new pharmaceuticals and medical devices. Given the need to exercise strict control, it is hardly surprising that institutions such as the U.S. Food and Drug Administration adopt a fairly conservative line in statistical innovations and retain a strong interest in frequentist properties of any statistical method. Nevertheless, 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" (International Conference on Harmonisation E9 Expert Working Group, 1999). 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.

A recent example has proved, however, that it is possible to obtain regulatory approval for a large and complex adaptive trial that uses a Bayesian monitoring procedure. Berry et al. (2002) describe the design of a phase II/III dose-finding study in acute stroke, in which 15 different doses were to be given at random at the start of randomization, with steady adaptation to the range of doses around the ED95, that is, the minimum dose that provides 95% of the maximum efficacy. The original decision-theoretic stopping criterion was replaced by one based on posterior tail-areas being less than a certain value: a frequentist assessment of the size and power of the study was based on pretrial simulations and approved by the FDA. The trial was closely monitored, with the statistician of the data monitoring committee (myself) receiving weekly summaries of the posterior distributions in order to check whether the critical boundaries had been crossed. The DMC recommended stopping when the "futility" boundary was crossed after recruiting over 900 patients; the trial stopped immediately and subsequently reported an essentially "flat" dose–response curve (Krams et al., 2003).

The greatest enthusiasm for Bayesian methods appears to be in U.S. FDA Center for Devices and Radiological Health (CDRH) (Campbell, 1999). Devices differ from pharmaceuticals in having better understood physical mechanisms, which means that effectiveness is generally robust to small changes. Since devices tend to develop in incremental steps, a large body of relevant evidence often exists 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.

4.2 The Role of Decision Theory

The debate about the appropriate role of formal decision theory in health-care evaluation continues. Claims for a strong role of decision theory include the following:

- In the context of clinical trials, Lindley (1994) categorically states that

  > clinical trials are not there for inference but to make decisions,

  while Berry (1994) 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 and Simon (1978) 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.

- Within a pharmaceutical company it is natural to try to maximize profitability, and this naturally leads to the use of utilities.
- Within a health-policy setting, decision theory and economic argument clearly state that maximized expected utility is the sole criteria for choosing between two options. Therefore measures of "significance," posterior tail areas of incremental net

benefit and so on are all irrelevant (Claxton and Posnett, 1996). Claxton, Lacey and Walker (2000) point 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.

Uncertainty is only taken into account through evaluating the benefit of further experimentation, as measured by a value of information analysis.

- To maximize the health return from the limited resources available from a health budget, health-care purchasers should use rational resource allocation procedure. Otherwise the resulting decisions could be considered as irrational, inefficient and unethical.
- Overall, a decision-theoretic framework provides a formal basis for designing trials, assessing whether to approve an intervention for use, deciding whether an intervention is cost-effective and commissioning further research.

Claims against the use of decision theory include the following:

- 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 (1985) in the discussion of Bather (1985), 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 the subsequent dissemination of a treatment.

- The idea of a null hypothesis (the *status quo*), which lies behind the use of "statistical significance" or posterior tail-areas, is fundamentally different from an alternative hypothesis (a novel intervention). The consequences and costs of the former are generally established, whereas the impact of the latter must contain a substantial amount of judgement. Often, therefore, a choice between two treatments is not a choice between two equal contenders to be decided solely on the balance of net benefit—some convincing evidence is required before changing policy.
- A change in policy carries with it many hidden penalties: for example, it may be difficult to reverse if later found to be erroneous, and it may hinder the development of other, better, innovations. It would be difficult to explicitly model these phenomena with any plausibility.
- Value of information analysis is strongly dependent on having the "correct" model, which is never known and generally cannot be empirically checked. Sensitivity analysis can only compensate to some extent for this basic ignorance.

Whitehead (1997, page 208) 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 may be modelled as decision problems, but that phase III trials cannot.

The discussion in Section 2.1 has revealed the complexity of the context in which health-care evaluation takes place, and clearly a simplistic decision-theoretic approach is inappropriate. Nevertheless in the context of any real decision that must be made, it would seem beneficial to have at least a qualitative expression of the potential gains and losses, and from there to move toward a full quantitative analysis.

4.3 Increasing the Appropriate Use of Bayesian Methods

We conclude by some brief personal opinions about innovations that may lead to wider and improved use of Bayesian methods.

First, we need a set of good worked *examples* based on realistic situations and that set good standards in specification and analysis. Second, we need a structure for *reporting* Bayesian analyses that permits rapid critical appraisal: as Berry (2002) says:

> There is as much Bayesian junk as there is frequentist junk. Actually, there's probably more of the former because, to the uninitiated, the Bayesian approach appears to provide a free lunch.

Third, following the running theme of this paper, there is a need to understand and *integrate* with the current methodology and software used in studies. Finally, it should be acknowledged that Bayesian

methods do *not provide a panacea*. Problems should be clearly highlighted and it should be acknowledged that sampling properties of systems may be important in some contexts. The general statistical community, who are not stupid, have justifiably found somewhat tiresome the tone of hectoring self-righteousness that has often come from the Bayesian lobby. Fortunately that period seems to be coming to a close, and with luck the time has come for the appropriate use of Bayesian thinking to be pragmatically established.

REFERENCES

ADES, A. E. and CLIFFE, S. (2002). Markov chain Monte Carlo estimation of a multiparameter decision model: Consistency of evidence and the accurate assessment of uncertainty. *Medical Decision Making* **22** 359–371.

BATHER, J. A. (1985). On the allocation of treatments in sequential medical trials (with discussion). *Internat. Statist. Rev.* **53** 1–13, 25–36.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300.

BENSON, K. and HARTZ, A. (2000). A comparison of observational studies and randomized controlled trials. *New England J. Medicine* **342** 1878–1886.

BERGER, J. O. and BERRY, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist* **76** 159–165.

BERRY, D. A. (1994). Discussion of "Bayesian approaches to randomized trials," by D. J. Spiegelhalter, L. S. Freedman and M. K. B. Parmar. *J. Roy. Statist. Soc. Ser. A* **157** 399.

BERRY, D. A. (2002). Adaptive clinical trials and Bayesian statistics (with discussion). In *Pharmaceutical Report—American Statistical Association* **9** 1–11. Amer. Statist. Assoc., Alexandria, VA.

BERRY, D. A. (2004). Bayesian statistics and the efficiency and ethics of clinical trials. *Statist. Sci.* **19** 175–187.

BERRY, D. A., MÜLLER, P., GRIEVE, A., SMITH, M., PARKE, T., BLAZEK, R., MITCHARD, N. and KRAMS, M. (2002). Adaptive Bayesian designs for dose-ranging drug trials. In *Case Studies in Bayesian Statistics*. *Lecture Notes in Statist.* **162** 99–181. Springer, New York.

BERRY, D. A. and STANGL, D. K. (1996a). Bayesian methods in health-related research. In *Bayesian Biostatistics* (D. A. Berry and D. K. Stangl, eds.) 3–66. Dekker, New York.

BERRY, D. A. and STANGL, D. K., eds. (1996b). *Bayesian Biostatistics*. Dekker, New York.

BRIGGS, A. (2000). Handling uncertainty in cost-effectiveness models. *Pharmacoeconomics* **17** 479–500.

BRITTON, A., MCKEE, M., BLACK, N., MCPHERSON, K., SANDERSON, C. and BAIN, C. (1998). Choosing between randomised and non-randomised studies: A systematic review. *Health Technology Assessment* **2** 1–124.

BROOKS, S. P. (1998). Markov chain Monte Carlo method and its application. *The Statistician* **47** 69–100.

BROPHY, J. and JOSEPH, L. (2000). A Bayesian analysis of random mega-trials for the choice of thrombyotic agents in acute myocardial infarction. In *Meta-Analysis in Medicine and Health Policy* (D. K. Stangl and D. A. Berry, eds.) 83–104. Dekker, New York.

BURTON, P. R. (1994). Helping doctors to draw appropriate inferences from the analysis of medical studies. *Statistics in Medicine* **13** 1699–1713.

CAMPBELL, G. (1999). A regulatory perspective for Bayesian clinical trials. Food and Drug Administration, Washington.

CASELLA, G. and GEORGE, E. (1992). Explaining the Gibbs sampler. *Amer. Statist.* **46** 167–174.

CHALONER, K. (1996). Elicitation of prior distributions. In *Bayesian Biostatistics* (D. A. Berry and D. K. Stangl, eds.) 141–156. Dekker, New York.

CHALONER, K. and RHAME, F. (2001). Quantifying and documenting prior beliefs in clinical trials. *Statistics in Medicine* **20** 581–600.

CHESSA, A. G., DEKKER, R., VAN VLIET, B., STEYERBERG, E. W. and HABBEMA, J. D. F. (1999). Correlations in uncertainty analysis for medical decision making: An application to heart-valve replacement. *Medical Decision Making* **19** 276–286.

CHRISTIANSEN, C. L. and MORRIS, C. N. (1997a). Improving the statistical approach to health care provider profiling. *Annals of Internal Medicine* **127** 764–768.

CHRISTIANSEN, C. L. and MORRIS, C. N. (1997b). Hierarchical Poisson regression modeling. *J. Amer. Statist. Assoc.* **92** 618–632.

CLAXTON, K., LACEY, L. F. and WALKER, S. G. (2000). Selecting treatments: A decision theoretic approach. *J. Roy. Statist. Soc. Ser. A* **163** 211–225.

CLAXTON, K. and POSNETT, J. (1996). An economic approach to clinical trial design and research priority-setting. *Health Economics* **5** 513–524.

CORNFIELD, J. (1966). A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. *J. Amer. Statist. Assoc.* **61** 577–594.

CORNFIELD, J. (1969). The Bayesian outlook and its application. *Biometrics* **25** 617–657.

CORNFIELD, J. (1976). Recent methodological contributions to clinical trials. *American J. Epidemiology* **104** 408–421.

COX, D. R. (1999). Discussion of "Some statistical heresies," by J. K. Lindsey. *The Statistician* **48** 30.

CRONIN, K. A., FREEDMAN, L. S., LIEBERMAN, R., WEISS, H. L., BEENKEN, S. W. and KELLOFF, G. J. (1999). Bayesian monitoring of phase II trials in cancer chemoprevention. *J. Clinical Epidemiology* **52** 705–711.

DANIELS, M. J. (1999). A prior for the variance components in hierarchical models. *Canad. J. Statist.* **27** 567–578.

DECISIONEERING (2000). Crystal ball. Technical report. Available at http://www.decisioneering.com/crystal_ball.

DERSIMONIAN, R. (1996). Meta-analysis in the design and monitoring of clinical trials. *Statistics in Medicine* **15** 1237–1248.

DIGNAM, J. J., BRYANT, J., WIEAND, H. S., FISHER, B. and WOLMARK, N. (1998). 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 Clinical Trials* **19** 575–588.

DIXON, D. O. and SIMON, R. (1991). Bayesian subset analysis. *Biometrics* **47** 871–881.

DOMINICI, F., PARMIGIANI, G., WOLPERT, R. and HASSELBLAD, V. (1999). Meta-analysis of migraine

headache treatments: Combining information from heterogeneous designs. *J. Amer. Statist. Assoc.* **94** 16–28.

DUMOUCHEL, W. and NORMAND, S. (2000). Computer-modeling and graphical strategies for meta-analysis. In *Meta-Analysis in Medicine and Health Policy* (D. K. Stangl and D. A. Berry, eds.) 127–178. Dekker, New York.

EDDY, D. M., HASSELBLAD, V. and SHACHTER, R. (1992). *Meta-Analysis by the Confidence Profile Method*: *The Statistical Synthesis of Evidence*. Academic Press, San Diego.

ETZIONI, R. D. and KADANE, J. B. (1995). Bayesian statistical methods in public health and medicine. *Annual Review of Public Health* **16** 23–41.

FAYERS, P. M., ASHBY, D. and PARMAR, M. K. B. (1997). Tutorial in biostatistics: Bayesian data monitoring in clinical trials. *Statistics in Medicine* **16** 1413–1430.

FAYERS, P. M., CUSCHIERI, A., FIELDING, J., CRAVEN, J., USCINSKA, B. and FREEDMAN, L. S. (2000). Sample size calculation for clinical trials: The impact of clinician beliefs. *British J. Cancer* **82** 213–219.

FLETCHER, A., SPIEGELHALTER, D., STAESSEN, J., THIJS, L. and BULPITT, C. (1993). Implications for trials in progress of publication of positive results. *The Lancet* **342** 653–657.

FRYBACK, D. G., STOUT, N. K. and ROSENBERG, M. A. (2001). An elementary introduction to Bayesian computing using WINBUGS. *International J. Technology Assessment in Health Care* **17** 98–113.

GILBERT, J. P., MCPEEK, B. and MOSTELLER, F. (1977). Statistics and ethics in surgery and anesthesia. *Science* **198** 684–689.

GILKS, W. R., RICHARDSON, S. and SPIEGELHALTER, D. J., eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, New York.

GOLDSTEIN, H. and SPIEGELHALTER, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance (with discussion). *J. Roy. Statist. Soc. Ser. A* **159** 385–443.

GOULD, A. L. (1991). Using prior findings to augment active-controlled trials and trials with small placebo groups. *Drug Information J.* **25** 369–380.

GRAY, R. J. (1994). A Bayesian analysis of institutional effects in a multicenter cancer clinical trial. *Biometrics* **50** 244–253.

GREENHOUSE, J. B. and WASSERMAN, L. (1995). Robust Bayesian methods for monitoring clinical trials. *Statistics in Medicine* **14** 1379–1391.

GRIEVE, A. P. (1994). Discussion of "Bayesian approaches to randomized trials," by D. J. Spiegelhalter, L. S. Freedman and M. K. B. Parmar. *J. Roy. Statist. Soc. Ser. A* **157** 387–388.

HANSON, T., BEDRICK, E., JOHNSON, W. and THURMOND, M. (2003). A mixture model for bovine abortion and foetal survival. *Statistics in Medicine* **22** 1725–1739.

HARRELL, F. E. and SHIH, Y. C. T. (2001). Using full probability models to compute probabilities of actual interest to decision makers. *International J. Technology Assessment in Health Care* **17** 17–26.

HASSELBLAD, V. (1998). Meta-analysis of multi-treatment studies. *Medical Decision Making* **18** 37–43.

HEALY, M. J. R. and SIMON, R. (1978). New methodology in clinical trials. *Biometrics* **34** 709–712.

HEITJAN, D. F. (1997). Bayesian interim analysis of phase II cancer clinical trials. *Statistics in Medicine* **16** 1791–1802.

HIGGINS, J. P. and WHITEHEAD, A. (1996). Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* **15** 2733–2749.

IBRAHIM, J. G. and CHEN, M.-H. (2000). Power prior distributions for regression models. *Statist. Sci.* **15** 46–60.

INTERNATIONAL CONFERENCE ON HARMONISATION E9 EXPERT WORKING GROUP (1999). Statistical principles for clinical trials: ICH harmonised tripartite guideline. *Statistics in Medicine* **18** 1905–1942. Available at http://www.ich.org.

IOANNIDIS, J. P. A., HAIDICH, A. B., PAPPA, M., PANTAZIS, N., KOKORI, S. I., TEKTONIDOU, M. G., CONTOPOULOS-IOANNIDIS, D. G. and LAU, J. (2001). Comparison of evidence of treatment effects in randomized and nonrandomized studies. *J. American Medical Association* **286** 821–830.

KADANE, J. B. (1995). Prime time for Bayes. *Controlled Clinical Trials* **16** 313–318.

KASS, R. E. and GREENHOUSE, J. B. (1989). A Bayesian perspective. Comment on "Investigating therapies of potentially great benefit: ECMO," by J. H. Ware. *Statist. Sci.* **4** 310–317.

KRAMS, M., LEES, K., HACKE, W., GRIEVE, A., ORGOGOZO, J. and FORD, G. (2003). Acute stroke therapy by inhibition of neutrophils (ASTIN): An adaptive dose–response study of UK-279,276 in acute ischemic stroke. *Stroke* **34** 2543–2548.

KUNZ, R. and OXMAN, A. D. (1998). The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials. *British Medical J.* **317** 1185–1190.

LAROSE, D. T. and DEY, D. K. (1997). Grouped random effects models for Bayesian meta-analysis. *Statistics in Medicine* **16** 1817–1829.

LAU, J., SCHMID, C. H. and CHALMERS, T. C. (1995). Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. *J. Clinical Epidemiology* **48** 45–57.

LINDLEY, D. V. (1994). Discussion of "Bayesian approaches to randomized trials," by D. J. Spiegelhalter, L. S. Freedman and M. K. B. Parmar. *J. Roy. Statist. Soc. Ser. A* **157** 393.

LINDLEY, D. V. (2000). The philosophy of statistics (with discussion). *The Statistician* **49** 293–337.

MATTHEWS, R. A. J. (2001). Methods for assessing the credibility of clinical trial outcomes. *Drug Information J.* **35** 1469–1478.

NATARAJAN, R. and KASS, R. E. (2000). Reference Bayesian methods for generalized linear mixed models. *J. Amer. Statist. Assoc.* **95** 227–237.

NORMAND, S.-L., GLICKMAN, M. E. and GATSONIS, C. A. (1997). Statistical methods for profiling providers of medical care: Issues and applications. *J. Amer. Statist. Assoc.* **92** 803–814.

O'HAGAN, A. and LUCE, B. (2003). *A Primer on Bayesian Statistics in Health Economics and Outcomes Research.* Centre for Bayesian Statistics in Health Economics, Sheffield, UK.

O'HAGAN, A., STEVENS, J. W. and MONTMARTIN, J. (2000). Inference for the cost-effectiveness acceptability curve and cost-effectiveness ratio. *Pharmacoeconomics* **17** 339–349.

O'HAGAN, A., STEVENS, J. W. and MONTMARTIN, J. (2001). Bayesian cost-effectiveness analysis from clinical trial data. *Statistics in Medicine* **20** 733–753.

O'NEILL, R. T. (1994). Conclusions. 2. *Statistics in Medicine* **13** 1493–1499.

PALISADE EUROPE (2001). @RISK 4.0. Technical report. Available at http://www.palisade-europe.com.

PARMAR, M. K. B., GRIFFITHS, G. O., SPIEGELHALTER, D. J., SOUHAMI, R. L., ALTMAN, D. G. and VAN DER SCHEUREN, E. (2001). Monitoring of large randomised clinical trials—a new approach with Bayesian methods. *The Lancet* **358** 375–381.

PARMAR, M. K. B., SPIEGELHALTER, D. J. and FREEDMAN, L. S. (1994). The CHART trials: Bayesian design and monitoring in practice. *Statistics in Medicine* **13** 1297–1312.

PARMAR, M. K. B., UNGERLEIDER, R. S. and SIMON, R. (1996). Assessing whether to perform a confirmatory randomized clinical trial. *J. National Cancer Institute* **88** 1645–1651.

PETO, R. (1985). Discussion of "On the allocation of treatments in sequential medical trials," by J. Bather. *Internat. Statist. Rev.* **53** 31–34.

POCOCK, S. (1976). The combination of randomized and historical controls in clinical trials. *J. Chronic Diseases* **29** 175–188.

PREVOST, T. C., ABRAMS, K. R. and JONES, D. R. (2000). Hierarchical models in generalized synthesis of evidence: An example based on studies of breast cancer screening. *Statistics in Medicine* **19** 3359–3376.

RACINE, A., GRIEVE, A. P., FLUHLER, H. and SMITH, A. F. M. (1986). Bayesian methods in practice—experiences in the pharmaceutical industry (with discussion). *Appl. Statist.* **35** 93–150.

REEVES, B., MACLEHOSE, R., HARVEY, I., SHELDON, T., RUSSELL, I. and BLACK, A. (2001). A review of observational, quasi-experimental and randomized study designs for the evaluation of the effectiveness of healthcare interventions. In *The Advanced Handbook of Methods in Evidence Based Healthcare* (A. Stevens, K. Abrams, J. Brazier, R. Fitzpatrick and R. Lilford, eds.) 116–135. Sage, London.

SANDERSON, C., MCKEE, M., BRITTON, A., BLACK, N., MCPHERSON, K. and BAIN, C. (2001). Randomized and nonrandomized studies: Threats to internal and external validity. In *The Advanced Handbook of Methods in Evidence Based Healthcare* (A. Stevens, K. Abrams, J. Brazier, R. Fitzpatrick and R. Lilford, eds.) 95–115. Sage, London.

SHAKESPEARE, T. P., GEBSKI, V. J., VENESS, M. J. and SIMES, J. (2001). Improving interpretation of clinical studies by use of confidence levels, clinical significance curves, and risk-benefit contours. *The Lancet* **357** 1349–1353.

SIMON, R. (1994). Some practical aspects of the interim monitoring of clinical trials. *Statistics in Medicine* **13** 1401–1409.

SIMON, R., DIXON, D. O. and FRIEDLIN, B. (1996). Bayesian subset analysis of a clinical trial for the treatment of HIV infections. In *Bayesian Biostatistics* (D. A. Berry and D. K. Stangl, eds.) 555–576. Dekker, New York.

SONG, F., ALTMAN, D., GLENNY, A. and DEEKS, J. J. (2003). Validity of indirect comparison for estimating efficacy of competing interventions: Empirical evidence from published meta-analyses. *British Medical J.* **326** 472–476.

SPIEGELHALTER, D. J. (2001). Bayesian methods for cluster randomized trials with continuous responses. *Statistics in Medicine* **20** 435–452.

SPIEGELHALTER, D. J., ABRAMS, K. R. and MYLES, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health Care Evaluation*. Wiley, New York.

SPIEGELHALTER, D. J. and BEST, N. G. (2003). Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in Medicine* **22** 3687–3708.

SPIEGELHALTER, D. J., FREEDMAN, L. S. and PARMAR, M. K. B. (1994). Bayesian approaches to randomized trials (with discussion). *J. Roy. Statist. Soc. Ser. A* **157** 357–416.

SPIEGELHALTER, D. J., MYLES, J., JONES, D. and ABRAMS, K. (2000). Bayesian methods in health technology assessment: A review. *Health Technology Assessment* **4** 1–130.

STANGL, D. K. and GREENHOUSE, J. B. (1998). Assessing placebo response using Bayesian hierarchical survival models. *Lifetime Data Analysis* **4** 5–28.

SUTTON, A. J., ABRAMS, K. R., JONES, D. R., SHELDON, T. A. and SONG, F. (2000). *Methods for Meta-Analysis in Medical Research*. Wiley, New York.

TSIATIS, A. A. (1981). The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika* **68** 311–315.

TURNER, R., OMAR, R. and THOMPSON, S. (2001). Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine* **20** 453–472.

WHITEHEAD, A. (2002). *Meta-Analysis of Controlled Clinical Trials*. Wiley, New York.

WHITEHEAD, J. (1997). *The Design and Analysis of Sequential Clinical Trials*, 2nd ed. Wiley, New York.

ZUCKER, D. R., SCHMID, C. H., MCINTOSH, M. W., D'AGOSTINO, R. B., SELKER, H. P. and LAU, J. (1997). Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. *J. Clinical Epidemiology* **50** 401–410.