STATISTICS IN MEDICINE
*Statist. Med.* 2003; **22**:3687–3709 (DOI: 10.1002/sim.1586)

# Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling

David J Spiegelhalter[1,†] and Nicola G Best[2,*,‡,§]

[1] *MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, U.K.*
[2] *Department of Epidemiology and Public Health, Imperial College Faculty of Medicine, St. Mary's Campus, Norfolk Place, London W2 1PG, U.K.*

## SUMMARY

Increasingly complex models are being used to evaluate the cost-effectiveness of medical interventions. We describe the multiple sources of uncertainty that are relevant to such models, and their relation to either probabilistic or deterministic sensitivity analysis. A Bayesian approach appears natural in this context. We explore how sensitivity analysis to patient heterogeneity and parameter uncertainty can be simultaneously investigated, and illustrate the necessary computation when expected costs and benefits can be calculated in closed form, such as in discrete-time discrete-state Markov models. Information about parameters can either be expressed as a prior distribution, or derived as a posterior distribution given a generalized synthesis of available data in which multiple sources of evidence can be differentially weighted according to their assumed quality. The resulting joint posterior distributions on costs and benefits can then provide inferences on incremental cost-effectiveness, best presented as posterior distributions over net-benefit and cost-effectiveness acceptability curves. These ideas are illustrated with a detailed running example concerning the cost-effectiveness of hip prostheses in different age–sex subgroups. All computations are carried out using freely available software for conducting Markov chain Monte Carlo analysis. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: deterministic sensitivity analysis; evidence synthesis; generalized meta-analysis; Markov chain Monte Carlo simulation; probabilistic sensitivity analysis; WinBUGS

## 1. INTRODUCTION

It is being increasingly recognized that rational health-care policy can use cost-effectiveness analysis to inform decisions. It is also clear that multiple sources of uncertainty should be

[*] Correspondence to: N. G. Best, Department of Epidemiology and Public Health, Imperial College Faculty of Medicine, St. Mary's Campus, Norfolk Place, London W2 1PG, U.K.
[†] E-mail: david.spiegelhalter@mrc-bsu.cam.ac.uk
[‡] E-mail: n.best@imperial.ac.uk
[§] Partially supported by MRC Career Establishment Grant (G9803841).

Contract/grant sponsor: MRC Career Establishment; contract/grant number: G9803841

Received November 2001
Accepted December 2002

acknowledged, and in this paper we bring together four diverse but converging themes into a common framework:

1. complex cost-effectiveness models, in particular discrete-state discrete-time Markov models, which are being increasingly used to make predictions of the consequences of a particular intervention;
2. probabilistic sensitivity analysis in cost-effectiveness, in which distributions are put over uncertain parameters;
3. Bayesian approaches to cost-effectiveness, in particular using Markov chain Monte Carlo (MCMC) methods, to incorporate evidence from a single source (e.g. data arising from a clinical trial) with appropriate propagation of parameter uncertainty;
4. the synthesis of evidence from multiple sources in a form of generalized meta-analysis. There will usually be insufficient randomized evidence to fully inform a model that takes into account long-term consequences of an intervention. A generalized synthesis would allow the use of evidence from studies of different designs, possibly including the controversial practice of combining randomized and non-randomized evidence.

The combined literature on these topics is becoming large and only selected references will be provided. Of particular note, however, is the review by Briggs [1] which introduces many of the issues in this paper in a non-technical style. We also note the special issue on Bayesian methods of the *International Journal of Health Technology Assessment in Health Care* which features many relevant articles [2].

The structure of the paper is as follows. Section 2 describes a general framework for describing and handling uncertainty in complex cost-effectiveness models, closely related to the categorization suggested by the U.S. Panel on Cost-Effectiveness [3], while Section 3 introduces the problem of making predictions using discrete-time cost-effectiveness models allowing for heterogeneous populations. Closed-form and simulation solutions are described, and illustrated in Section 4 with a reworking of an example concerning hip replacements. Probabilistic sensitivity analysis is described in Section 5 and illustrated with our running example, emphasizing the computations necessary to produce a decomposition of total variance into components attributable to heterogeneity and parameter uncertainty. Section 6 discusses the integration of evidence from multiple sources, including randomized, single cohort and registry data. In synthesizing this evidence, we include the option of formally downweighting potentially biased studies; the degree of downweighting is a judgement that should be subject to sensitivity analysis. The Bayesian probability model then results in a posterior distribution for the unknown parameters. This distribution then feeds into the probabilistic sensitivity analysis that underlies the incremental cost-effectiveness analysis described in Section 7. Finally, in Section 8 we draw some conclusions concerning the potential for Bayesian approaches in this context.

Each stage of this process is illustrated using the hip replacement example, and all computations are carried out using the freely available software WinBUGS [4, 5]. We hope that the provision of code (available on `www.mrc-bsu.cam.ac.uk/bugs/examples`) will help practitioners to explore the potential of these methods.

## 2. LEVELS OF UNCERTAINTY AND THEIR ROLE IN SENSITIVITY ANALYSIS

Approaches to uncertainty in cost-effectiveness analysis have been extensively reviewed by Briggs and Gray [6], who emphasize the distinction between conducting 'deterministic' sensitivity analysis in which inputs to the model are systematically varied within a reasonable range, and 'probabilistic' sensitivity analysis in which the relative plausibility of unknown parameters is taken into account.

We now relate these different approaches to sensitivity analysis to different sources of uncertainty, relating our structure to the taxonomy described by Briggs [1] and the U.S. Panel on Cost-Effectiveness [3].

1. *Chance variability*: This is the unavoidable *within-individual* predictive uncertainty concerning their specific outcomes or, equivalently, random variability in outcomes between *homogeneous* individuals. We are usually not interested in this 'first-order' uncertainty [1] since our focus is on the *expected* outcomes in homogeneous populations, but we shall illustrate its calculation in Sections 3.2 and 4.3.
2. *Heterogeneity*: This source concerns *between-individual* variability in their expected outcomes, either due to (a) identifiable subgroups of individuals with characteristics such as age, sex and other covariates, or (b) unmeasurable differences (latent variables). These are termed 'patient characteristics' by Briggs [1]. We shall generally want to use deterministic sensitivity analysis to see how expected outcomes vary between identifiable subgroups, possibly followed by probabilistic averaging over population subgroups according to their incidence.
3. *Parameter uncertainty*: This concerns *within-model* uncertainty as to the appropriate values for parameters. Parameters can be separated into

   (a) *States-of-the-world*, which could, in theory, be measured precisely if sufficient evidence were available, for example risks, disease incidences and so on: these have also been termed 'parameters that could be sampled' [1]. These can have distributions placed on them, corresponding to the 'second-order' uncertainty used in risk analysis [7], and so be subject to probabilistic sensitivity analysis.

   (b) *Assumptions*, which are quantitative judgements placed in the model that can only be made precise through consensus agreement, for example discount rates. These can be considered as one source of 'methodological uncertainty' [1], and sensitivity to assumptions can only be carried out deterministically by re-running analyses under different scenarios.

   The appropriate category for a quantity is not always clear. For example, whether values placed on quality-of-life scales are states-of-the-world or assumptions is a controversial point, and costs might also be placed in either category. If the costs are based on explicit data, then we may be able to judge the error associated with the mean costs: note that for both quality-of-life measures and costs it is the uncertainty about the mean value that is of interest, not the variation in the patient population, which one might expect to be considerable.

4. *Ignorance*: This describes our basic lack of knowledge concerning the appropriate qualitative structure of the model, for example, the dependence of the hazard rates on

> background factors and history. This is also a component of 'methodological uncertainty' [1]. Sensitivity analysis takes the form of running through alternative models (deterministic), although there is an argument that model structure can itself be considered as an unknown state-of-the-world and be subject to probabilistic sensitivity analysis [8].

In this paper, we shall primarily be concerned with probabilistic sensitivity analysis, although we will also illustrate deterministic sensitivity analysis with respect to parameter assumptions.

## 3. COST AND EFFECTIVENESS MODELLING ALLOWING FOR CHANCE VARIATION AND PATIENT HETEROGENEITY

### *3.1. Discrete-time discrete-state Markov models*

Discrete-time discrete-state Markov models comprise a common framework for predicting costs and benefits over time. These models assume that in each 'cycle' an individual is in one of a finite set of states, and that the chance of entering a new state at the end of the cycle does not depend on what path the individual took to their current state (although it may depend on the cycle and other developing risk factors). There are obviously many extensions to this reasonably flexible framework [9, 10].

We shall first formally describe the generic structure of the model for a single homogeneous set of patients with common parameters. Assume a discrete-time model comprising $T$ cycles labelled $t=1,\ldots,T$. Assume that within each cycle $t$ a patient remains in one of $K$ states, and that all transitions occur at the start of each cycle. The probability distribution at the start of the first cycle $t=1$ is represented by the row vector $\boldsymbol{\pi}_1$, and we assume a transition matrix $\boldsymbol{\Lambda}_t$ whose $i$, $j$th element is the probability of moving from state $i$ to state $j$ between cycle $t-1$ and $t$; thus the probability, for example, of being in state $j$ during the second cycle is $\sum_i \pi_{1i}\Lambda_{2,ij}$. Hence, the marginal probability distribution $\boldsymbol{\pi}_t$ during cycle $t>1$ obeys the recursive relationship

$$\boldsymbol{\pi}_t = \boldsymbol{\pi}_{t-1}\boldsymbol{\Lambda}_t \tag{1}$$

Suppose the cost, at current prices, of spending a cycle in state $k$ is $c_k$, $k=1,\ldots,K$, and there is a fixed entry cost $c_0$. It is standard practice in economic evaluations to discount costs that occur in future years, at rate $100\delta_c$ per cent (say) per cycle. Then the total cost acquired by each patient in the population is expected to be

$$E[C] = c_0 + \sum_{t=1}^{T} \frac{\boldsymbol{\pi}_t \mathbf{c}'}{(1+\delta_c)^{t-1}} \tag{2}$$

Similarly if the benefits of being in each state are given by a row vector $\mathbf{b}$, discounted at rate $100\delta_b$ per cent per cycle, the total expected benefit for each patient is

$$E[B] = \sum_{t=1}^{T} \frac{\boldsymbol{\pi}_t \mathbf{b}'}{(1+\delta_b)^{t-1}} \tag{3}$$

We note that different types of benefit may be reported, for example both life expectancy and quality-adjusted life-years.

Suppose there are $S$ discrete subgroups labelled by $s$. The model described above can clearly be extended to allow, say, for different transition matrices within subgroups by extending the notation to $\boldsymbol{\Lambda}_{st}$.

### 3.2. Making predictions in cost-effectiveness models

Let $\theta$ represent state-of-the-world parameters in a cost-effectiveness model, say $\boldsymbol{\pi}_1$ and $\boldsymbol{\Lambda}_t$, and let $X$ be an unknown generic outcome of interest, whether a cost or a benefit, taking on a value $x$. Suppose, for a specified value of $\theta$ and subgroup $s$, we can specify a predictive distribution $p(x|s,\theta)$, the *chance variability* between future patients (Section 2). Our primary interest is in $E(X|s,\theta) = \int xp(x|s,\theta)\,\mathrm{d}x = m_{s\theta}$, the expected outcome in this homogeneous population. There are two means of determining expected costs and benefits:

1. *Closed form*: For the discrete-time, discrete-state Markov model described above the expectations $m_{s\theta}$ are available in closed form, given by Equation (2) for costs and by Equation (3) for benefits.
2. *Simulation*: If we are using a more complex model, such as a continuous time formulation, then it may be necessary to simulate from $p(x|s,\theta)$ and use the sample mean of the simulations as an estimate of $m_{s\theta}$. This does have the advantage of additionally giving the whole distribution $p(x|s,\theta)$, and in particular $\mathrm{Var}(X|s,\theta) = v_{s\theta}$ among the population. This 'first-order simulation' approach is illustrated by Briggs [1] and has been exploited in the context of evaluating screening interventions using the term 'micro-simulation' [11].

   For example, if we wished to explore this approach for the model described in Section 3.1, then we could simulate a sequence of indicator arrays (representing the state of the $n$th simulated patient at time $t$) as multinomial variables with order 1: i.e.

$$\begin{aligned} y_1^{(n)} &\sim \text{multinomial}(\boldsymbol{\pi}_1, 1) \\ y_t^{(n)} &\sim \text{multinomial}(y_{(t-1)}^{(n)}\boldsymbol{\Lambda}_t, 1), \quad t = 2,\ldots,T \end{aligned} \tag{4}$$

   Substituting $y_t^{(n)}$ for $\pi_t$ into Equations (2) and (3) will give the total cost $C^{(n)}$ and benefit $B^{(n)}$ for the $n$th simulated patient, and averaging over patients (i.e. over iterations $n = 1,\ldots,N$) gives Monte Carlo estimates of the required expectations $m_{s\theta}^{[\mathrm{C}]}$ and $m_{s\theta}^{[\mathrm{B}]}$. We may also calculate the variances of $C^{(n)}$ and $B^{(n)}$ across iterations to obtain Monte Carlo estimates of $v_{s\theta}^{[\mathrm{C}]}$ and $v_{s\theta}^{[\mathrm{B}]}$—i.e. the variability of each outcome due to chance.

## 4. ILLUSTRATIVE EXAMPLE: COST-EFFECTIVENESS ANALYSIS OF HIP REPLACEMENT PROSTHESES

Our running example will concern the choice of prosthesis in total hip replacement: this is a very common orthopaedic procedure with a substantial potential benefit in terms of pain relief and improved physical function. However, there is a wide range of products available and being used, with limited evidence of their relative effectiveness, particularly in terms of their revision rates for different subpopulations. Since prostheses vary considerably in cost, the National Institute of Clinical Evidence for England and Wales (NICE) has issued guidance

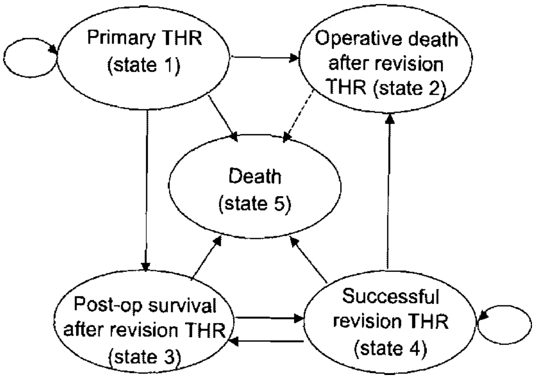

Figure 1. Markov model for outcomes following primary total hip replacement.

as to the cost-effectiveness of different types of prostheses [12], making substantial use of a previous analysis presented by Fitzpatrick and colleagues [13].

Here, we use a model for outcome following hip replacement based on that of Fitzpatrick [13] to illustrate a structured approach to the various sources of uncertainty, and use evidence on relative revision rates for different prostheses quoted in the NICE appraisal [12] to carry out an incremental cost-effectiveness analysis. However, our purpose is in developing statistical methodology, and so our results should not be taken as contributing in any way to guidance as to an appropriate prosthesis: we refer to other publications for a detailed discussion of both the clinical and economic issues [13–15].

### 4.1. Statistical model

Our model for predicting costs and benefits following hip replacement is a discrete-time, discrete-state Markov model. The first cycle ($t = 1$) is assumed to start immediately following the primary total hip replacement (THR) operation; patients may either die at operation or post-operatively, in which case they enter state 5 (death), otherwise they remain in state 1. In subsequent cycles, surviving patients remain in state 1 until they either die from other causes (progress to state 5) or their hip replacement fails and they require a revision THR operation. Patients undergoing a revision operation enter one of two states depending on whether they die post-operation (state 2) or survive (state 3). Surviving patients progress to state 4 (successful revision THR) in the following cycle, unless they die from other causes (progress to state 5). Patients in state 4 remain there until they either die from other causes (state 5) or require another revision THR operation, in which case they progress back to states 2 or 3 as before. We also assume a transition from states 2 to 5 in the cycle following operative death after a revision THR. This is slightly artificial but is necessary to avoid multiple counting of revision costs (see Equation (2)) if patients were to remain in state 2. Figure 1 illustrates the various states and possible transitions between states.

Transitions between states are defined over a time frame (cycle length) of 1 year. The vector of state probabilities in cycle $t = 1$ is $\pi_1 = (1 - \lambda_{\text{op}}, 0, 0, 0, \lambda_{\text{op}})$. We then consider a further 59 cycles of the model, chosen to ensure that patients in the youngest age group at $t = 1$ should have died by the end of the full 60 cycles (years). The transition probability

Table I. Age- and sex-specific mortality rates, and age and sex distribution of patients receiving primary THR in the U.K.

| Age (yr) | Mortality rate | | % of THR recipients | |
|---|---|---|---|---|
| | Men | Women | Men (%) | Women (%) |
| <45 | 0.0017 | 0.0011 | 2 | 2 |
| 45–54 | 0.0044 | 0.0028 | 3 | 4 |
| 55–64 | 0.0138 | 0.0081 | 7 | 10 |
| 65–74 | 0.0379 | 0.0220 | 13 | 22 |
| 75–84 | 0.0912 | 0.0578 | 10 | 26 |
| >84 | 0.1958 | 0.1503 | 0 | 1 |

matrix for $t=2,\ldots,60$ is given below, where $\mathbf{\Lambda}_{t,jk}$ is the probability of being in state $j$ in year $t-1$ and moving to state $k$ at the start of year $t$, $\lambda_{\text{op}}$ is the operative mortality rate, $\gamma_t$ is the hazard for revision in year $t$, $\lambda_t$ is the mortality rate $t$ years after primary operation, and $\rho$ is the re-revision rate which is assumed constant.

$$\mathbf{\Lambda}=\begin{bmatrix} 1-\gamma_t-\lambda_t & \lambda_{\text{op}}\gamma_t & (1-\lambda_{\text{op}})\gamma_t & 0 & \lambda_t \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1-\lambda_t & \lambda_t \\ 0 & \rho\lambda_{\text{op}} & \rho(1-\lambda_{\text{op}}) & 1-\rho-\lambda_t & \lambda_t \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

### 4.2. *Parameters of the model*

We follow Fitzpatrick [13] in adopting the widely used Charnley prosthesis as a baseline analysis, assumed to have a constant post-operative mortality rate, $\lambda_{\text{op}}=0.01$, and a constant re-revision rate, $\rho=0.04$. We also assume a linearly increasing revision hazard $\gamma_t=h(t-1)$ (i.e. no replacements in first year), but unlike Fitzpatrick [13], we allow the annual increment $h$ to depend on the age and sex of the patient. On the basis of revision rates for the Charnley prostheses in the Swedish hip replacement register [16], and assuming higher revision rates for men and younger people [17], we take $h=0.0022$ for men $<65$ years, $h=0.0017$ for women $<65$ years, $h=0.0016$ for men $\geqslant 65$ years and $h=0.0012$ for women $\geqslant 65$ years. Finally, we assume that patients surviving THR operations have the same mortality rates as the general population, and use the national age- and sex-specific rates published by the U.K. Office for National Statistics [12] and reproduced in Table I. Also shown in Table I is the age–sex distribution of patients receiving primary THR using the Charnley prosthesis in the U.K. [12]. These data will be used in analyses allowing for patient heterogeneity (Section 5.1).

### 4.3. *Costs and benefits*

Estimates of the costs of a primary and revision THR operation using the Charnley prosthesis were obtained from Fitzpatrick [13]. For a typical patient, primary THR costs are $C_{\text{p}}=£4052$, and revision THR costs are $C_{\text{R}}=£5290$, and we use an annual discount rate for costs occurring in future years of $\delta_c=6$ per cent per annum.

Health-related quality of life (HRQL) is measured by quality-adjusted life-years (QALYs) based on the degree of severity of pain patients would be likely to experience in different states of the model. Based on results from a Canadian study [18], Fitzpatrick [13] assigns values $v_1 = 1$, $v_2 = 0.69$, $v_3 = 0.38$ and $v_4 = 0.19$ for the HRQL of patients experiencing no, mild, moderate or severe pain, respectively. Then they assume that after a successful THR operation, 80 per cent of patients experience no pain and 20 per cent experience mild pain. For patients whose hip replacements fail, they assume that 15 per cent experience severe pain and 85 per cent experience moderate pain in the year preceding the year of the revision operation, with a 50–50 split between those experiencing moderate pain and severe pain in the year of operation. We, therefore, calculate QALYs for each state in our Markov model as follows:

$$
\begin{aligned}
\text{QALY}_1 &= 0.8v_1 + 0.2v_2 = 0.938\\
\text{QALY}_2 &= 0 + 1.06 \times (0.85v_3 + 0.14v_4 - 0.8v_1 - 0.2v_2) = -0.622\\
\text{QALY}_3 &= (v_3 + v_4)/2 + 1.06 \times (0.85v_3 + 0.15v_4 - 0.8v_1 - 0.2v_2) = -0.337\\
\text{QALY}_4 &= 0.8v_1 + 0.2v_2 = 0.938\\
\text{QALY}_5 &= 0
\end{aligned}
$$

We note the somewhat anomalous negative values for states 2 and 3, which represent a subtraction of quality from the preceding year for patients requiring a revision operation. As for costs, we discount QALYs (and also life expectancy) in future years at a rate of $\delta_b = 6$ per cent per annum: we note that a different discount rate for benefits and costs may be a more reasonable assumption [1] and we investigate sensitivity to this in Section 7.2.

The top section of Table II gives the expected costs and benefits for each subgroup, calculated both in closed form and via Monte Carlo simulation. Monte Carlo estimates of the chance variability are expressed by the sampling standard deviations of these costs and benefits. The simulation-based estimates of the expectations agree well with the exact results within each subgroup, and the standard deviations show substantial variability between the outcomes attained by individual patients. While the expected costs are reasonably constant across subgroups, there is clear heterogeneity in expected benefits.

From now on we will calculate all expectations $m_{s\theta}$ in closed form, and so ignore the 'first-order' chance variability.

## 5. PROBABILISTIC SENSITIVITY ANALYSIS

In Section 2 we identified two sources of uncertainty to which probabilistic sensitivity analysis might be applied: population heterogeneity and parameter uncertainty. We shall consider each in turn and then their simultaneous analysis.

### *5.1. Sensitivity to patient heterogeneity for fixed parameters $\theta$*

Suppose we desired an overall measure of cost-effectiveness across an entire population, but with a summary of the variability due to patient heterogeneity. We are willing at this stage

Table II. Expected costs and benefits of THR for patient subgroups with fixed parameters calculated both exactly and using Monte Carlo simulation, and (bottom three rows) overall, allowing for subgroup heterogeneity with fixed parameters.

| | Costs (£) | | | Life expectancy (yr) | | | QALYs | | |
|---|---|---|---|---|---|---|---|---|---|
| | Exact | Monte Carlo | | Exact | Monte Carlo | | Exact | Monte Carlo | |
| Subgroup | $m_{s\theta}^{[C]}$ | $m_{s\theta}^{[C]}$ | $\sqrt{v_{s\theta}^{[C]}}$ | $m_{s\theta}^{[L]}$ | $m_{s\theta}^{[L]}$ | $\sqrt{v_{s\theta}^{[L]}}$ | $m_{s\theta}^{[Q]}$ | $m_{s\theta}^{[Q]}$ | $\sqrt{v_{s\theta g}^{[Q]}}$ |
| *Men* | | | | | | | | | |
| 35–44 yr | 5781 | 5793 | 1892 | 14.5 | 14.5 | 2.9 | 13.2 | 13.2 | 2.6 |
| 45–54 yr | 5417 | 5435 | 1889 | 12.7 | 12.7 | 3.4 | 11.6 | 11.6 | 3.0 |
| 55–64 yr | 4989 | 4974 | 1659 | 10.3 | 10.3 | 3.7 | 9.5 | 9.5 | 3.3 |
| 65–74 yr | 4466 | 4454 | 1211 | 7.7 | 7.7 | 3.5 | 7.2 | 7.2 | 3.2 |
| 75–84 yr | 4263 | 4250 | 905 | 5.4 | 5.4 | 3.0 | 5.0 | 5.1 | 2.8 |
| >84 yr | 4193 | 4203 | 806 | 4.1 | 4.1 | 3.0 | 3.8 | 3.9 | 2.7 |
| *Women* | | | | | | | | | |
| 35–44 yr | 5626 | 5641 | 1835 | 15.1 | 15.2 | 2.6 | 13.8 | 13.8 | 2.4 |
| 45–54 yr | 5350 | 5346 | 1765 | 13.7 | 13.7 | 3.0 | 12.5 | 12.6 | 2.8 |
| 55–64 yr | 5002 | 5020 | 1666 | 11.6 | 11.6 | 3.6 | 10.7 | 10.7 | 3.2 |
| 65–74 yr | 4487 | 4484 | 1242 | 9.1 | 9.0 | 3.7 | 8.4 | 8.4 | 3.4 |
| 75–84 yr | 4282 | 4277 | 955 | 6.5 | 6.4 | 3.3 | 6.0 | 6.0 | 3.1 |
| >84 yr | 4212 | 4209 | 812 | 5.0 | 5.0 | 3.4 | 4.6 | 4.6 | 3.2 |
| | Exact | Monte Carlo | | Exact | Monte Carlo | | Exact | Monte Carlo | |
| *Overall* | | | | | | | | | |
| Mean, $m_\theta$ | 4603 | 4600 | | 8.7 | 8.7 | | 8.0 | 8.0 | |
| SD, $\sqrt{v_\theta}$ | 403 | 409 | | 2.6 | 2.6 | | 2.3 | 2.3 | |
| CV $= \sqrt{v_\theta}/m_\theta$ | 0.09 | 0.09 | | 0.30 | 0.30 | | 0.29 | 0.29 | |

*Note*: The third column for each outcome gives the sampling standard deviation in that outcome, estimated using Monte Carlo integration.

to consider the parameters $\theta$ of the model to be known. Allowing $s$ to have a distribution $p(s|\theta)$ provides

$$E_{s|\theta}[m_{s\theta}] = m_\theta$$
$$\text{Var}_{s|\theta}[m_{s\theta}] = v_\theta$$

where $E_{s|\theta}$ represents an expectation with respect to the distribution of $s$ for fixed $\theta$. Thus, $m_\theta$ and $v_\theta$ are the mean and variance of the expected outcomes over the subpopulations, for fixed $\theta$. If $m_{s\theta}$ is available in closed form for each of a finite set of subgroups, then $m_\theta$ and $v_\theta$ may be obtained by direct calculation.

*Example*: For the hip prosthesis example, the subpopulations $S$ comprise the age–sex groups as shown in Table I. Calculation of $m_\theta$ and $v_\theta$ can then be performed by computing the weighted mean and variance of either the closed form or the Monte Carlo expectations for each subgroup, where the weights are given by the subgroup percentages in Table I. The final three rows of Table II give, respectively, $m_\theta$, $\sqrt{v_\theta}$ and the coefficient of variation $\sqrt{v_\theta}/m_\theta$ for each of the three outcomes of interest; these summarize the expected costs and benefits of THR and their variability across patient subgroups. As informally noted before, these measures emphasize the reasonable consistency of costs but the substantial variability of benefits across population subgroups.

### *5.2. Sensitivity to uncertain parameters for a fixed patient subgroup s*

We now consider a contrasting situation in which we are concerned with individual subgroups but wish to summarize the consequences of parameter uncertainty. Allowing $\theta$ to have distribution $p(\theta|s)$ within each subgroup $s$ provides

$$E_{\theta|s}[m_{s\theta}] = m_s$$

$$\mathrm{Var}_{\theta|s}[m_{s\theta}] = v_s$$

where $E_{\theta|s}$ represents an expectation with respect to the distribution of $\theta$ for a given subgroup $s$. Thus $m_s$ and $v_s$ are the mean and variance of the expected outcomes for specific sub-populations allowing for uncertainty in $\theta$.

Assuming $m_{s\theta}$ is available in closed form, $m_s$ and $v_s$ can be estimated by simulating values of $\theta$ from $p(\theta|s)$, evaluating $m_{s\theta}$ and taking the sample mean and variance over $\theta$. This is a natural application of Monte Carlo methods to deal with 'second-order uncertainty' in homogeneous populations, which has become a standard tool in risk analysis. It is implementable as macros for Excel, either from commercial software such as @RISK [19] and Crystal Ball [20], or self-written. Here, however, we use the freely available WinBUGS software [4] in order to facilitate extensions to include evidence synthesis (Section 6).

Application of these techniques to our example is described in Sections 5.3 and 5.4 and is shown in Table III.

### *5.3. Joint sensitivity to uncertain parameters and heterogeneity*

When we wish to simultaneously investigate sensitivity to both heterogeneity and parameter uncertainty, then we need to consider the joint distribution $p(s, \theta)$ which provides summary statistics

$$\begin{aligned} E_{s\theta}[m_{s\theta}] &= m \\ \mathrm{Var}_{s\theta}[m_{s\theta}] &= v \end{aligned} \tag{5}$$

that quantify the expectation and variance of the outcome over patient subgroups and plausible parameter values.

The overall summaries $m$ and $v$ can be obtained in two ways, corresponding to expressing $p(s, \theta)$ as $p(s|\theta)p(\theta)$ or as $p(\theta|s)p(s)$.

1. We might condition on the parameters and average over the subgroups with respect to $p(s|\theta)$ (as in Section 5.1) followed by simulating from the uncertain $p(\theta)$. This is perhaps a natural order when the behaviour of individual subgroups is considered unimportant.

Table III. Expectation and standard deviation of costs and benefits of THR for patient subgroups allowing for parameter uncertainty, and (bottom two panels) overall, allowing for parameter uncertainty and patient heterogeneity.

| | Costs (£) | | Life expectancy (yr) | | QALYs | |
|---|---|---|---|---|---|---|
| Subgroup | $m_s^{[C]}$ | $\sqrt{v_s^{[C]}}$ | $m_s^{[L]}$ | $\sqrt{v_s^{[L]}}$ | $m_s^{[Q]}$ | $\sqrt{v_s^{[Q]}}$ |
| *Men* | | | | | | |
| 35–44 yr | 5787 | 231 | 14.5 | 0.0052 | 13.2 | 0.060 |
| 45–54 yr | 5425 | 202 | 12.7 | 0.0037 | 11.6 | 0.052 |
| 55–64 yr | 4997 | 152 | 10.3 | 0.0021 | 9.5 | 0.038 |
| 65–74 yr | 4472 | 75 | 7.7 | 0.0008 | 7.2 | 0.019 |
| 75–84 yr | 4266 | 40 | 5.4 | 0.0003 | 5.0 | 0.010 |
| >84 yr | 4196 | 27 | 4.1 | 0.0002 | 3.8 | 0.007 |
| *Women* | | | | | | |
| 35–44 yr | 5636 | 218 | 15.1 | 0.0052 | 13.8 | 0.057 |
| 45–54 yr | 5359 | 194 | 13.7 | 0.0038 | 12.5 | 0.050 |
| 55–64 yr | 5010 | 154 | 11.7 | 0.0024 | 10.7 | 0.039 |
| 65–74 yr | 4493 | 79 | 9.1 | 0.0010 | 8.4 | 0.020 |
| 75–84 yr | 4285 | 44 | 6.5 | 0.0004 | 6.0 | 0.011 |
| >84 yr | 4215 | 31 | 5.0 | 0.0003 | 4.6 | 0.008 |
| *Overall* (using $p(s,\theta) = p(s\mid\theta)p(\theta)$) | | | | | | |
| Overall expectation $m$ | | 4609 | | 8.7 | | 8.0 |
| Var. due to uncertainty, $v_{\text{P1}} = \text{Var}_\theta[m_\theta]$ | | 1013 | | 0.0000002 | | 0.00006 |
| Var. due to heterogeneity, $v_{\text{H1}} = E_\theta[v_\theta]$ | | 174 400 | | 6.7 | | 5.5 |
| Total variance $v_1 = v_{\text{P1}} + v_{\text{H1}}$ | | 175 413 | | 6.7000002 | | 5.50006 |
| Percentage variance due to heterogeneity | | 99.4% | | 99.9% | | 99.9% |
| *Overall* (using $p(s,\theta) = p(\theta\mid s)p(s)$) | | | | | | |
| Overall expectation $m$ | | 4609 | | 8.7 | | 8.0 |
| Var. due to uncertainty, $v_{\text{P2}} = E_s[v_s]$ | | 11 473 | | 0.000003 | | 0.0007 |
| Var. due to heterogeneity, $v_{\text{H2}} = \text{Var}_s[m_s]$ | | 163 953 | | 6.7 | | 5.5 |
| Total variance $v_2 = v_{\text{P2}} + v_{\text{H2}}$ | | 175 426 | | 6.700003 | | 5.5007 |
| Percentage variance due to heterogeneity | | 93.5% | | 99.9% | | 99.9% |

Then, we obtain from standard identities

$$\begin{aligned} m &= E_\theta[E_{s|\theta}[m_{s\theta}]] = E_\theta[m_\theta] \\ v &= E_\theta[\text{Var}_{s|\theta}[m_{s\theta}]] + \text{Var}_\theta[E_{s|\theta}[m_{s\theta}]] = E_\theta[v_\theta] + \text{Var}_\theta[m_\theta] = v_{\text{H1}} + v_{\text{P1}} \end{aligned} \tag{6}$$

The latter can be considered as a decomposition of the total variance $v$ in expected outcome into two components corresponding to patient heterogeneity ($v_{\text{H1}}$) and parameter uncertainty ($v_{\text{P1}}$), respectively. Since we are assuming $m_\theta$ and $v_\theta$ can be obtained in closed form, the decomposition can be obtained using Monte Carlo estimates of the required quantities.

2. When individual subgroups are of more importance, it is natural to first condition on the subgroups and simulate parameters from $p(\theta|s)$ (as in Section 5.2) followed by averaging

with respect to $p(s)$. Then, we obtain

$$\begin{aligned} m &= E_s[E_{\theta|s}[m_{\theta|s}]] = E_s[m_s] \\ v &= E_s[\mathrm{Var}_{\theta|s}[m_{\theta|s}]] + \mathrm{Var}_s[E_{\theta|s}[m_{\theta|s}]] = E_s[v_s] + \mathrm{Var}_s[m_s] = v_{\mathrm{P2}} + v_{\mathrm{H2}} \end{aligned} \tag{7}$$

This final decomposition of the total variance $v$ in expected outcome into components corresponding to parameter uncertainty ($v_{\mathrm{P2}}$) and heterogeneity ($v_{\mathrm{H2}}$) is illustrated in Section 5.4. Making our standard assumptions, $m_s$ and $v_s$ may be obtained from Monte Carlo estimates, while $v_{\mathrm{P2}}$ and $v_{\mathrm{H2}}$ are calculated directly from the $m_s$ and $v_s$ using the discrete prior $p(s)$. Thus, the percentage of variability due to the two sources can be calculated.

The second approach would appear to be most commonly relevant, although we illustrate both approaches in our example below.

### 5.4. Example: sensitivity to heterogeneity and parameter uncertainty

One relevant state-of-the-world parameter in our model for prognosis following THR is the revision hazard $h$. It may be reasonable to assume uncertainty of ±50 per cent about our assumed revision hazards (which we now denote $h_0$) for each age and sex group. This gives an approximate 95 per cent interval of ($h_0t/1.5$, $h_0t \times 1.5$) for the revision hazard, which we represent as a normal distribution for the log hazard

$$\log h \sim N(\log h_0, 0.2^2) \tag{8}$$

The top part of Table III gives the expectation $m_s$ and standard deviation $\sqrt{v_s}$ of the costs and benefits for each subgroup, allowing for uncertainty in the revision hazard (Section 5.2). The bottom two panels of this table give the overall expectation and variance of each outcome across subgroups and the hazard distribution, evaluated using, respectively, Equations (6) and (7), and taking both $p(s)$ and $p(s|\theta)$ equal to the age–sex distribution provided in Table I.

It is clear that even considerable uncertainty about revision hazard rates has little influence on life expectancy or QALYs, but does lead to substantial sensitivity on costs. When combined with the influence of heterogeneity, parameter uncertainty is only responsible for 6.5 per cent or less of the total variance for costs and less than 0.01 per cent of the total variance for benefits.

### 5.5. If closed-form expectations are not available

Although not relevant to our example, it is important to realize that closed-form expectations may not be available in more complex models and a micro-simulation approach may be necessary (Section 3.2) in which individual patient outcomes are simulated. We briefly discuss the necessary computations, assuming a single subgroup (no heterogeneity).

A time-consuming nested simulation procedure [21] is required. A value $\theta^j$ for $\theta$ is simulated from $p(\theta)$, followed by simulation of $M$ (where $M$ is large) values of the outcome $X_1^j, \ldots, X_M^j$ conditional on $\theta^j$. The sample mean $\bar{X}_M^j$ and variance $V_M^j$ are stored. Over many simulations of $\theta$, monitoring any $X_i$ will provide the overall expectation $m$ and variance $v$ for a single individual, although the variance combines both parameter uncertainty and chance variability and will generally be of little interest. Monitoring $\bar{X}_M$ and $V_M$ will, however, allow estimation of the components of the overall variability, since $\mathrm{Var}_\theta[\bar{X}_M] \approx v_\mathrm{P}$ will estimate variability due to parameter uncertainty, while $E_\theta[V_M]$ gives that due to chance variability. However, this technique will be laborious, particularly when heterogeneity is present. See Reference [11] for an application.

## 6. INTEGRATING EVIDENCE WITH THE MODEL

### *6.1. Generalized meta-analysis of evidence*

Up until now we have assumed that any available evidence (e.g. on the revision hazard) can be summarized as a prior distribution whose influence is assessed by propagating uncertainty through the model using 'forward' Monte Carlo methods. This two-stage process can be integrated into a single analysis in which the posterior distribution arising from a data analysis feeds directly into the cost-effectiveness without an intermediate summary step. This corresponds to a full Bayesian probability model and requires Markov chain Monte Carlo rather than simply Monte Carlo techniques, since in effect the evidence from the data has to be propagated 'backwards' in order to give the uncertainty on the parameters, and then 'forwards' through the cost-effectiveness model. A schematic representation is shown in Figure 2.

O'Hagan and colleagues [22–24] have illustrated this technique for evidence from a single trial and a simple cost-effectiveness model, while Fryback and colleagues [25] provide a further example of a posterior distribution being used as a direct input to probabilistic sensitivity analysis. The potential advantages of this integrated approach over the two-stage process are discussed in Section 6.3.

The common situation in which evidence is available from a variety of sources demands a more challenging statistical analysis. If the evidence comprises a set of similar trials, then

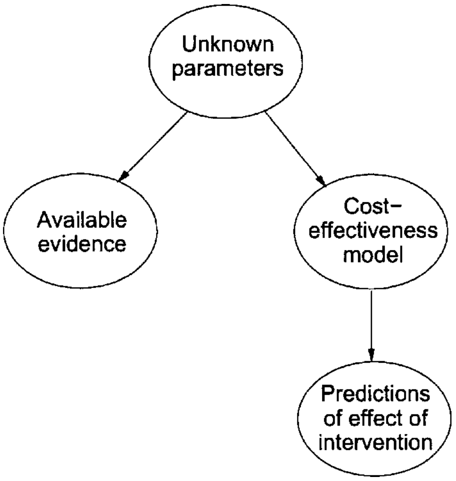

Figure 2. Schematic graph showing the dependence of both available evidence and future predictions on unknown parameters. After taking into account the available evidence, initial prior opinions on the parameters are revised by Bayes theorem to posterior distributions, the effects of which are propagated 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.

a standard Bayesian random-effects meta-analysis may be sufficient [26, 27]. In more complex situations there will be multiple studies with relevance to the quantities in question but which may suffer from a range of potential inadequacies, such as being based on different populations, having non-randomized control groups, outcomes measured on different scales, and so on. Formal combination of such disparate sources is fraught with methodological problems but has been studied under a range of labels. *Cross-design synthesis* [28] is a general term for pooling evidence from different study designs, while the *Confidence Profile Method* of Eddy and colleagues [29] explicitly models a range of internal and external biases that a study may exhibit. It is natural to extend Bayesian random-effects modelling to allow variance components corresponding to different study designs (i.e. assuming study types are 'exchangeable') resulting in hierarchical models with a study-type 'level': examples include pooling randomized and non-randomized studies on breast cancer screening [30], and pooling open- and closed-trial designs [31, 32].

There are clearly a number of issues in carrying out such potentially controversial modelling, such as when to judge studies or study-types 'exchangeable', how to put appropriate prior distributions on variance components, and how to carry out sensitivity analyses.

We shall consider as an illustration a somewhat simple formulation of such a model. Suppose we have a set of studies that are each intending to estimate a single parameter $\delta$ but, due to differences in populations studied and so on, any particular study (if carried out meticulously) would in fact be estimating a biased parameter $\delta + \delta_h$. Here $\delta_h$ is the 'external bias', and a standard random effects formulation might then assume $\delta_h \sim N(0, \sigma_h^2)$ (note that the mean would not necessarily be 0 if we suspected systematic bias in one direction). However, suppose that due to quality limitations there is additional 'internal bias' in the study, so that the true parameter being estimated is $\delta + \delta_h + \delta_b$. Then we might assume $\delta_b \sim N(0, \sigma_b^2)$ if we did not suspect the internal bias would favour one or the other treatment. Overall, we are left with a random effects model in which, for study $i$, the data is estimating a parameter

$$\begin{aligned}\delta_i &\sim N(\delta, \sigma_h^2 + \sigma_{bi}^2)\\ &\sim N(\delta, \sigma_h^2/q_i)\end{aligned}$$

where $q_i = \sigma_h^2/(\sigma_{bi}^2 + \sigma_h^2)$ can be considered the 'quality weight' for each study, being the proportion of between-study variability unrelated to internal biasing factors. Thus, a high-quality randomized trial might have $q = 1$, while a non-randomized study may be downweighted by assigning $q = 0.1$.

Estimates or prior distributions of the between-study variance $\sigma_h^2$ and the quality weights $q_i$ might be obtained from a possible combination of empirical random-effects analyses of RCTs of this intervention, historical 'similar' case studies, and judgement. Of course, sensitivity analysis to a range of assumptions about the quality weights can be carried out, as illustrated in the following example.

### 6.2. Example: evidence synthesis for comparison of revision rates

In order to illustrate the trade-off between increased costs and benefits, we shall compare the cost-effectiveness of the Charnley prosthesis with a hypothetical alternative cemented prosthe-

Table IV. Summary of evidence on revision hazards for Charnley and Stanmore prostheses: hazard ratios $<1$ are in favour of Stanmore.

| | Charnley | | Stanmore | | Estimated | |
|---|---|---|---|---|---|---|
| Source | Number of patients | Revision rate | Number of patients | Revision rate | hazard ratio (HR) | (95% int.) |
| | | | | | *Fixed-effects model* | |
| Registry | 28 525 | 5.9% | 865 | 3.2% | 0.55 | (0.37–0.77) |
| RCT | 200 | 3.5% | 213 | 4.0% | 1.34 | (0.45–3.46) |
| Case series | 208 | 16.0% | 982 | 7.0% | 0.44 | (0.28–0.66) |
| | | | | | *Common-effect model* | |
| | | | | | 0.52 | (0.39–0.67) |
| Quality weights [registry, RCT, case series] | | | | | *Random-effects model* | |
| | | | | [1, 1, 1] | 0.54 | (0.37–0.78) |
| | | | | [0.5, 1, 0.2] | 0.61 | (0.36–0.98) |
| | | | | [0.1, 1, 0.05] | 0.82 | (0.36–1.67) |

ses costing an extra £350 but with some evidence for lower revision rates. We assume that all other costs (operating staff/theatre costs, length of hospital stay, X-rays, etc.) are the same for both prosthesis types, and that the same method of QALY assessment is applicable for both types of prosthesis.

For illustration, we assume that the revision hazard for our hypothetical alternative is similar to that for the Stanmore prostheses (a popular alternative to the Charnley in practice). Evidence on the relative revision hazards for the two prostheses is limited. The report by NICE on the cost-effectiveness of different prostheses for THR [12] cites three sources providing direct comparisons between Charnley and Stanmore revision rates:

1. The Swedish Hip Registry [17] provides non-randomized data submitted from all hospitals in Sweden from 1979, with record linkage to further procedures and death. Nine-year follow-up results are used for around 30 000 Charnley and Stanmore prostheses.
2. A U.K. Randomized Controlled Trial (RCT) [33] randomized around 400 patients to Charnley or Stanmore and reported a mean follow-up of 6.5 years.
3. A Case Series [34] of around 1200 patients in a single hospital with a mean follow-up of 8 years.

The available evidence from these three sources on revision hazards for Charnley and Stanmore prostheses is summarized in Table IV.

We assume the following model for pooling evidence on the revision hazard ratio for Stanmore versus Charnley prostheses. Let $N_{ik}$ and $r_{ik}$ denote the total number of patients receiving prosthesis $i$ ($1=$ Charnley, $2=$ Stanmore) in study $k$, and the number requiring a revision operation, respectively. We assume $r_{ik}$ is binomially distributed with proportion $p_{ik}$, and $H_{ik}$ is the cumulative hazard up to the mean follow-up, so that $\log(-\log(1-p_{ik}))=\log H_{ik}$. Assuming a proportional hazards model, with hazard ratio $\mathrm{HR}_k$ for Stanmore versus Charnley prostheses, leads to the following likelihood:

$$r_{ik} \sim \text{Binomial}(p_{ik}, N_{ik}), \quad i = 1,2$$

$$\log(-\log(1 - p_{1k})) = \log H_{1k}$$

$$\log(-\log(1 - p_{2k})) = \log H_{2k} = \log H_{1k} + \log \text{HR}_k$$

Placing uniform prior distributions over $\log H_{1k}$ and $\text{HR}_k$ gives the 'fixed-effects' estimates of the hazard ratio for each source shown in the first three rows of Table IV, revealing reasonable concordance between the non-randomized studies but with the randomized trial showing some evidence against the Stanmore. Forcing a common hazard ratio leads to the registry overwhelming the other sources (row 4 of Table IV).

The random-effects analysis with quality weights described in Section 6.1 leads to the model

$$\log \text{HR}_k \sim N\left(\log \overline{\text{HR}}, \frac{\sigma_h^2}{q_k}\right)$$

where $\overline{\text{HR}}$ is the overall estimate of the revision hazard ratio pooled across studies.

Three studies do not provide sufficient evidence to accurately estimate the between-study standard deviation $\sigma_h$, and so substantial prior judgement is necessary. We would expect considerable heterogeneity in revision rates between studies, even if they are internally unbiased, and so assume that $\sigma_h$ has a normal distribution with mean 0.2 and standard deviation 0.05, corresponding to expecting ±50 per cent variability in true hazard ratios between studies, with 95 per cent uncertainty limits of 20–80 per cent variability. The results of a random-effects analysis with all quality weights assumed to be 1 are shown in row 5 of Table IV, again showing the domination of the registry data.

Our knowledge of the potential biases of registries and case series suggest downweighting the non-randomized evidence. As a baseline assumption for the quality weights, we take $q_k$ equal to 0.5, 1.0 and 0.2, respectively, for the registry, RCT and case series studies. This corresponds to assuming that 'bias' in the registry and case series studies leads to a 2-fold or 5-fold increase in the revision rate variance, respectively, over and above the between-study variability expected for RCTs. Row 6 of Table IV shows that the hazard ratio is still estimated in favour of the Stanmore but that the 95 per cent interval now only just excludes 1. As a further sensitivity analysis, we take $q_k$ equal to 0.1, 1.0 and 0.05, respectively, which leads to an equivocal result with substantial uncertainty (final row of Table IV).

### *6.3. Comparison of integrated Bayesian and two-stage approach*

The 'integrated' approach to evidence synthesis and cost-effectiveness analysis simultaneously derives the joint posterior distribution of all unknown parameters from a Bayesian probability model, and propagates the effects of the resulting uncertainty through the predictive model underlying the cost-effectiveness analysis. In contrast, the 'two-stage' approach would first carry out the evidence synthesis, summarizing the joint posterior distribution parametrically, and then in a separate analysis use this as a prior distribution in a probabilistic sensitivity analysis in the cost-effectiveness model.

Advantages of the integrated approach include the following. 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 [35], 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.

The 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 in a combination of standard statistical and spreadsheet programs.

## 7. INCREMENTAL COST-EFFECTIVENESS

### 7.1. Theory

Suppose we have cost-effectiveness models for two interventions. For fixed parameter values, let the expected outcomes for intervention $i = 1,2$ decompose into expected costs and benefits $m_{\theta i} = (m_{\theta i}^{[C]}, m_{\theta i}^{[B]})$. Then the incremental expected costs and benefits of intervention 2 over intervention 1 are $\mathrm{IC}_\theta = m_{\theta 2}^{[C]} - m_{\theta 1}^{[C]}$ and $\mathrm{IB}_\theta = m_{\theta 2}^{[B]} - m_{\theta 1}^{[B]}$. Many authors [36–39, 22, 25] have argued that statements of cost-effectiveness should be based on the joint distribution of $\mathrm{IC}_\theta$ and $\mathrm{IB}_\theta$ with respect to $p(\theta)$, the joint distribution for all uncertain parameters in the models. In particular, a plot of the joint distribution of $\mathrm{IC}_\theta$ and $\mathrm{IB}_\theta$ can be particularly informative.

A traditional summary of the comparison between two treatments is the incremental cost-effectiveness ratio (ICER) $\mathrm{IC}_\theta/\mathrm{IB}_\theta$ which, for fixed $\theta$, is the expected additional cost per unit additional benefit. However, when uncertainty about $\theta$ is acknowledged, inference on the ICER is hampered by the possibility that $\mathrm{IB}_\theta = 0$, and hence the ICER is infinite. A solution is to use the concepts of 'net benefit' and 'cost-effectiveness acceptability curves'.

For example, suppose $K$ is a given threshold cost per unit benefit, in that the health-care provider is willing to pay up to $K$ for an additional unit of benefit. Then, for fixed parameters $\theta$, the net benefit from the new intervention is

$$\beta_\theta(K) = K\,\mathrm{IB}_\theta - \mathrm{IC}_\theta$$

The distribution of $\beta_\theta(K)$ for fixed $K$ provides a variety of summary measures [23]. For example, $E_\theta[\beta_\theta(K)]$ is the expected net benefit, and Claxton [40] argues that intervention 2 should be chosen if this expectation is positive, without regard to 'statistical significance'. Perhaps a more flexible approach is to calculate $Q(K) = p_\theta[\beta_\theta(K) > 0]$, and plot this against $K$ to produce a cost-effectiveness acceptability curve (CEAC). Further discussion and examples of these concepts have been provided by others [36–39, 22, 25].

### 7.2. Example: Comparative cost-effectiveness analysis of two different hip prostheses

We now compare expected costs and benefits by running the Markov model for each of the Charnley and Stanmore prostheses, with appropriate allowance for uncertainty and heterogeneity. As before, we assume the distribution given in Equation (8) for the Charnley prosthesis hazard (now denoted $h_1$); for the hypothetical alternative prosthesis we estimate the revision hazard as $h_2 = h_1 \times \overline{\mathrm{HR}}$, where the hazard ratio $\overline{\mathrm{HR}}$ is estimated simultaneously with the

Table V. Summary of results of comparative analysis of cost-effectiveness for a hypothetical alternative versus the Charnley prostheses, using quality weights of [0.5, 1, 0.2] for weighting the registry, RCT and case study evidence, respectively.

| | $\mathrm{IC}_\theta$ (£) | | $\mathrm{IQ}_\theta$ (QALYs) | | ICER | | |
|---|---|---|---|---|---|---|---|
| Subgroup | Mean | SD | Mean | SD | Median | $Q(6000)$ | $Q(10\,000)$ |
| *Men* | | | | | | | |
| 35–44 yr | −90 | 256 | 0.136 | 0.063 | −846 | 0.92 | 0.94 |
| 45–54 yr | −28 | 216 | 0.113 | 0.053 | −457 | 0.91 | 0.93 |
| 55–64 yr | 71 | 156 | 0.081 | 0.038 | 581 | 0.87 | 0.92 |
| 65–74 yr | 216 | 75 | 0.038 | 0.018 | 5190 | 0.55 | 0.77 |
| 75–84 yr | 279 | 40 | 0.020 | 0.009 | 13 220 | 0.04 | 0.26 |
| > 84 yr | 303 | 26 | 0.013 | 0.006 | 21 830 | 0.00 | 0.02 |
| *Women* | | | | | | | |
| 35–44 yr | −63 | 238 | 0.127 | 0.059 | −691 | 0.91 | 0.94 |
| 45–54 yr | −14 | 206 | 0.109 | 0.051 | −349 | 0.90 | 0.93 |
| 55–64 yr | 66 | 161 | 0.083 | 0.039 | 537 | 0.87 | 0.92 |
| 65–74 yr | 209 | 79 | 0.040 | 0.019 | 4710 | 0.60 | 0.80 |
| 75–84 yr | 274 | 43 | 0.021 | 0.010 | 12 030 | 0.07 | 0.34 |
| > 84 yr | 297 | 28 | 0.015 | 0.007 | 18 790 | 0.00 | 0.06 |
| Overall | 183 | 90 | 0.048 | 0.022 | 3246 | 0.73 | 0.85 |

Markov model using the model for evidence synthesis based on comparison of Charnley and Stanmore revision rates described in Section 6.2.

Table V summarizes the expectation and variability due to parameter uncertainty of the incremental costs ($\mathrm{IC}_\theta = m_{\theta 2}^{[C]} - m_{\theta 1}^{[C]}$) and quality of life benefits ($\mathrm{IQ}_\theta = m_{\theta 2}^{[Q]} - m_{\theta 1}^{[Q]}$) of using the alternative prosthesis rather than the Charnley both for specific patient subgroups and also averaged over all patients. Note that similar summaries are possible for life expectancy. Table V also gives the median of the distribution of the incremental cost-effectiveness ratio ($\mathrm{ICER} = \mathrm{IC}_\theta/\mathrm{IQ}_\theta$): note the preceding discussion on the difficulty of giving interval estimates for this quantity when $\mathrm{IQ}_\theta = 0$ is a plausible value. The additional benefit from the alternative prostheses clearly decreases with increasing age, while the expected cost changes from favouring the alternative to favouring Charnley with increasing age. This leads to a negative ICER for younger ages.

The full joint distribution of $\mathrm{IC}_\theta$ and $\mathrm{IQ}_\theta$ is shown in Figure 3 for each subgroup and averaged over all patients. Values within the bottom right quadrant indicate both lower cost and greater benefit arising from the alternative prosthesis, and hence a strictly dominating intervention. The diagonal dashed line in each plot indicates the pairs of values of $\mathrm{IC}_\theta$ and $\mathrm{IQ}_\theta$ yielding a zero expected net benefit for Stanmore if the health-care provider is willing to pay up to $K = £6000$ for each additional QALY of benefit (i.e. $\beta_\theta(6000) = 0$); the proportion of points in the joint distribution below this line represents the probability of cost-effectiveness for the alternative prosthesis for $K = £6000$ (i.e. $Q(6000) = p_\theta[\beta_\theta(6000) > 0]$). Likewise, the diagonal dotted line represents pairs of values ($\mathrm{IC}_\theta, \mathrm{IQ}_\theta$) yielding a zero expected net benefit for $K = £10\,000$ (i.e. $\beta_\theta(10\,000) = 0$), and the proportion of points below the dotted line correspond to $Q(10\,000)$, the probability of cost-effectiveness for the alternative prosthesis at £10 000 per QALY. The cost-effectiveness probabilities $Q(6000)$ and $Q(10\,000)$ for each

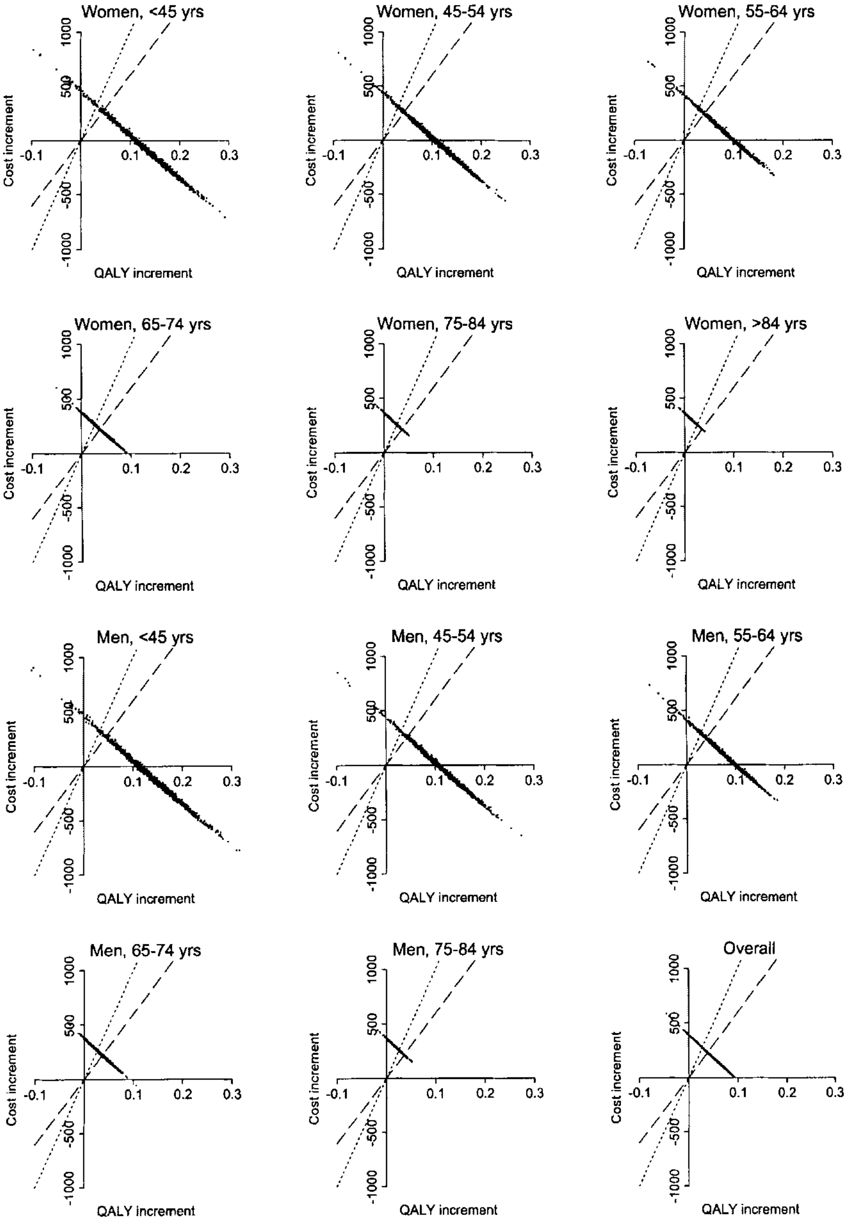

Figure 3. Incremental HRQL benefits ($IQ_\theta$) versus incremental costs ($IC_\theta$) for the alternative versus Charnley prostheses, for patient subgroups and overall (note that results for men aged $>84$ years are not shown due to space limitations and because the distribution of patients over age–sex groups shown in Table I indicates that there are no THR recipients in this subgroup). Diagonal lines indicate zero expected net benefit ($\beta_\theta(K)=0$) for the alternative prosthesis for $K=$ £6000 (-----) and $K=$ £10 000 (..........) per QALY.

subgroup and averaged over all subgroups are summarized in the final two columns of Table V. We see that, for $K = £10\,000$, the probability that the alternative prosthesis is the cost-effective option is around 80 per cent or more for both men and women under 75 years, but declines rapidly thereafter. For $K = £6000$ the threshold is around 65 years.

Figure 4 shows $Q(K)$, the probability of cost-effectiveness for the alternative prosthesis if the health-care provider is willing to pay up to $£K$ for each additional QALY of benefit, plotted against $K$ for each subgroup and averaged over all subgroups. The solid line is based on the results of the analysis reported above; the other curves in each plot indicate the sensitivity of the cost-effectiveness probabilities to various model assumptions. Specifically, we have re-run the model using downweighted evidence from the non-RCT studies on the revision hazard ratio for Stanmore versus Charnley. This was achieved by using quality weights $q_s$ equal to 0.1, 1.0 and 0.05, respectively, for the registry, RCT and case series studies. Sensitivity to the assumption that benefits are to be discounted by $\delta_b = 6$ per cent per annum was also examined, by re-running the models using a reduced health discount rate of 1.5 per cent per annum.

The results indicate that cost-effectiveness depends strongly on age (and to a lesser extent on sex), which suggests that economic evaluations should be made separately for the different subgroups. However, there is considerable sensitivity to the choice of quality weights used in the evidence synthesis, with further downweighting of the non-randomized evidence leading to consistently lower cost-effectiveness probabilities for the alternative prosthesis in all age and sex groups: the probability of cost-effectiveness does not rise above 75 per cent for any value of $K$ considered. This is to be expected, since the RCT provided less favourable evidence of reduced revision rates for the alternative prosthesis than did the non-randomized studies. Sensitivity to the health discount rate is not particularly strong in general, but is more apparent for older age groups.

It is of interest to compare the two-stage approach, which separates the data analysis and evidence synthesis from the cost-effectiveness analysis, to the integrated approach described above (Section 6.3). We applied the two-stage approach using the three data sources (registry, RCT and case series) to estimate the revision hazard for Charnley (rather than using the values for $h$ derived in Section 4.2) as well as the hazard ratio. Independent normal distributions were then assumed for the log hazard for Charnley and for the log hazard ratio. The results were virtually identical to the integrated analysis—the posterior standard deviations are about 1–2 per cent smaller under the two-stage approach and the CEA curves were very similar. The correlation between the log hazard for Charnley and the hazard ratio is also quite small (about $-0.15$) in the model, which would explain the similar results from the two approaches.

## 8. CONCLUSIONS

In this paper, we have attempted to explore a range of concerns that arise in cost-effectiveness modelling, but acknowledge that there are a number of issues that we have passed over. In particular, we have not explored the sensitivity of the conclusions to 'ignorance' about the structure of the appropriate model as discussed in Section 2: alternative models that could be used in this context include survival-type models with competing risks. It is vital to admit that even a reasonably complex model, such as that investigated in our example, cannot be assumed to be realistic and must be subject to careful criticism [41, 42].

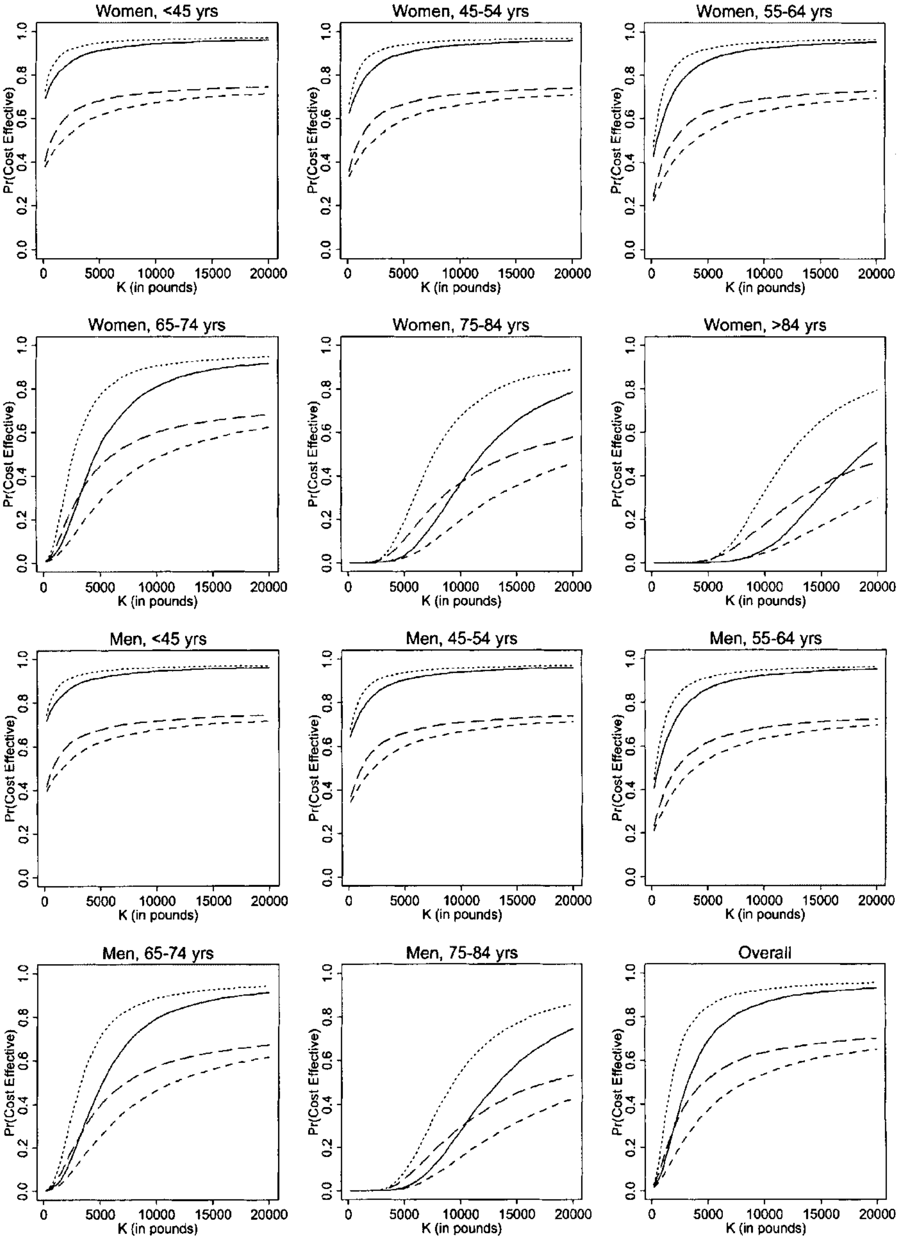

Figure 4. Probability of cost-effectiveness for the alternative prosthesis for each subgroup and overall, versus cost (£$K$) that the health-care provider is willing to pay for each additional QALY of benefit (note that results for men aged >84 years are not shown due to space limitations and because the distribution of patients over age–sex groups shown in Table I indicates that there are no THR recipients in this subgroup). Results are shown for different choices of quality weights $q_s$ for the registry, RCT and case series studies, and different health discount rates $d$, as follows: —— shows $q_s=(0.5,1.0,0.2)$ and $d=6$ per cent; ······ shows $q_s=(0.5,1.0,0.2)$ and $d=1.5$ per cent; ---- shows $q_s=(0.1,1.0,0.05)$ and $d=6$ per cent; — — — shows $q_s=(0.1,1.0,0.05)$ and $d=1.5$ per cent.

As attempts are made towards evidence-based health policy in both clinical and public-health contexts, models will inevitably become more complex and, while the methods described in this paper may appear complicated, we feel that techniques such as these may well become commonplace in the future. If decisions made with the help of such analyses are to be truly accountable, it is important that the models and methods are transparent, easily updatable, and can be run by many parties in order to check sensitivity. Models implemented in spreadsheet programs have some of these characteristics, although personal experience suggests that such programs are very clumsy in handling multi-dimensional arrays, and their expressions of complex formulae are quite opaque. Thus, the supposed transparency of popular spreadsheet programs may be somewhat illusory, and we feel that user-friendly Bayesian simulation programs could contribute substantially to the field.

The hip replacement data and WinBUGS code to fit each of the models discussed here are available from `www.mrc-bsu.cam.ac.uk/bugs/examples`.

## ACKNOWLEDGEMENTS

We are very grateful to Andy Briggs and Mark Sculpher for providing their original model and for helpful discussions, and to Andrew Thomas for rapidly adapting the WinBUGS program to be able to handle complex logical dependencies.

## REFERENCES

1. Briggs AH. Handling uncertainty in cost-effectiveness models. *Pharmacoeconomics* 2000; **17**:479–500.
2. Luce BR, Shih YCT, Claxton K. Bayesian approaches to technology assessment and decision making. *International Journal of Technology Assessment in Health Care* 2001; **17**:1–5.
3. Manning WG, Fryback FG, Weinstein M. Reflecting uncertainty in cost-effectiveness analysis. In *Cost Effectiveness in Health and Medicine*, MR G, JR S, MC W, LB R (eds). Oxford University Press: New York, 1996; 247–275.
4. Spiegelhalter DJ, Thomas A, Best NG, Lunn D. *WinBUGS Version 1.4 User Manual*. MRC Biostatistics Unit, Cambridge, available from www.mrc-bsu.cam.ac.uk/bugs 2002.
5. Fryback DG, Stout NK, Rosenberg MA. An elementary introduction to Bayesian computing using WinBUGS. *International Journal of Technology Assessment in Health Care* 2001; **17**:98–113.
6. Briggs AH, Gray AM. Methods in health service research—handling uncertainty in economic evaluations of healthcare interventions. *British Medical Journal* 1999; **319**:635–638.
7. Burmaster DE, Wilson AM. An introduction to second-order random variables in human health risk assessments. *Human Ecology Risk Assess* 1996; **2**:892–919.
8. Draper D. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B* 1995; **57**:45–97.
9. Briggs A, Sculpher M. Markov models of medical prognosis—commentary. *British Medical Journal* 1997; **314**:354–355.
10. Briggs A, Sculpher M. An introduction to Markov modelling for economic evaluation. *Pharmacoeconomics* 1998; **13**:397–409.
11. Cronin KA, Legler JM, Etzioni RD. Assessing uncertainty in microsimulation modelling with application to cancer screening interventions. *Statistics in Medicine* 1998; **17**(21):2509–2523.
12. NICE Appraisal Group. The effectiveness and cost effectiveness of different prostheses for primary total hip replacement. *Technical Report*, http://www.nice.org.uk 2000.
13. Fitzpatrick R, Shortall E, Sculpher M, Murray D, Morris R, Lodge M, *et al*. Primary total hip replacement surgery: a systematic review of outcomes and modelling cost effectiveness associated with different prostheses. *Health Technology Assessment* 1998; **20**(2):1–64.
14. Baxter K, Bevan G. An economic model to estimate the relative costs over 20 years of different hip prostheses. *Journal of Epidemiology and Community Health* 1999; **53**:542–547.
15. Faulkner A, Kennedy LG, Baxter K, Donovan J, Wilkinson M, Bevan G. Effectiveness of hip prostheses in primary total hip replacement: a critical review of evidence and an economic model. *Health Technology Assessment* 1998; **20**(6):1–133.

16. Malchau H, Herberts P, Ahnfelt L. Prognosis of total hip-replacement in Sweden—follow-up of 92,675 operations performed 1978–1990. *Acta Orthopaedica Scandinavica* 1993; **64**:497–506.
17. Malchau H, Herberts P. Prognosis of total hip replacement: revision and re-revision rate in thr. a revision-risk study of 148,359 primary operations. In *Proceedings of the 65th Annual Meeting of America Academy of Orthopaedic Surgeons*, New Orleans, U.S.A., 1998.
18. Laupacis A, Bourne R, Rorabeck C, Feeny D, Wong C, Tugwell P, *et al.* The effect of elective total hip-replacement on health-related quality-of-life. *Journal of Bone and Joint Surgery, American Volume* 1993; **75A**:1619–1626.
19. Palisade Europe. @RISK 4.0. *Technical Report*, http://www.palisade-europe.com 2001.
20. Decisioneering. Crystal Ball. *Technical Report*, http://www.decisioneering.com/crystal_ball 2000.
21. Halpern EF, Weinstein MC, Hunink MGM, Gazelle GS. Representing both first- and second-order uncertainties by Monte Carlo simulation for groups of patients. *Medical Decision Making* 2000; **20**:314–322.
22. O'Hagan A, Stevens JW, Montmartin J. Inference for the cost-effectiveness acceptability curve and cost-effectiveness ratio. *Pharmacoeconomics* 2000; **17**:339–349.
23. O'Hagan A, Stevens JW. A framework for cost-effectiveness analysis from clinical trial data. *Health Economics* 2001; **10**:303–315.
24. O'Hagan A, Stevens JW, Montmartin J. Bayesian cost-effectiveness analysis from clinical trial data. *Statistics in Medicine* 2001; **20**:733–753.
25. Fryback DG, Chinnis JO, Ulvila JW. Bayesian cost-effectiveness analysis—an example using the gusto trial. *International Journal of Technology Assessment in Health Care* 2001; **17**:83–97.
26. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian graphical modelling applied to random effects meta-analysis. *Statistics in Medicine* 1995; **14**:2685–2699.
27. Sutton A, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-Analysis in Medical Research*. Wiley: Chichester, 2000.
28. 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. *International Journal of Technology Assessment in Health Care* 1993; **9**:440–449.
29. Eddy DM, Hasselblad V, Shachter R. *Meta-Analysis by the Confidence Profile Method*: *the Statistical Synthesis of Evidence*. Academic Press: San Diego, CA, 1992.
30. Prevost TC, Abrams KR, Jones DR. Hierarchical models in generalised synthesis of evidence: an example based on studies of breast cancer screening. *Statistics in Medicine* 2000; **19**:3359–3376.
31. Larose DT, Dey DK. Grouped random effects models for Bayesian meta-analysis. *Statistics in Medicine* 1997; **16**:1817–1829.
32. Dominici F, Parmigiani G, Wolpert R, Hasselblad V. Meta-analysis of migraine headache treatments: combining information from heterogeneous designs. *Journal of American Statistical Association* 1999; **94**:16–28.
33. Marston RA, Cobb AG, Bentley G. Stanmore compared with Charnley total hip replacement—a prospective study of 413 arthroplasties. *Journal of Bone and Joint Surgery British Volume* 1996; **78B**:178–184.
34. Britton AR, Murray DW, Bulstrode CJ, McPherson K, Denham RA. Long-term comparison of Charnley and Stanmore design total hip replacements. *Journal of Bone and Joint Surgery British Volume* 1996; **78B**: 802–808.
35. Chessa AG, Dekker R, van Vliet B, Steyerberg EW, Habbema JDF. Correlations in uncertainty analysis for medical decision making: an application to heart-valve replacement. *Medical Decision Making* 1999; **19**: 276–286.
36. Grieve AP. Issues for statisticians in pharmaco-economic evaluations. *Statistics in Medicine* 1998; **17**: 1715–1723.
37. Heitjan DF, Moskowitz AJ, Whang W. Bayesian estimation of cost-effectiveness ratios from clinical trials. *Health Economics* 1999; **8**:191–201.
38. Sendi PP, Craig BA, Meier G, Pfluger D, Gafni A, Opravil M, *et al.* Cost-effectiveness of azithromycin for preventing Mycobacterium avium complex infection in HIV-positive patients in the era of highly active antiretroviral therapy. *Journal of Antimicrobial and Chemotherapy* 1999; **44**:811–817.
39. Briggs AH. A Bayesian approach to stochastic cost-effectiveness analysis—an illustration and application to blood pressure control in type 2 diabetes. *International Journal of Technology Assessment in Health Care* 2001; **17**:69–82.
40. Claxton K. Bayesian approaches to the value of information: implications for the regulation of new pharmaceutical. *Health Economics* 1999; **8**:269–274.
41. Russell LB. Modelling for cost-effectiveness analysis. *Statistics in Medicine* 1999; **18**:3235–3244.
42. Sculpher M, Fenwick E, Claxton K. Assessing quality in decision analytic cost-effectiveness models—a suggested framework and example of application. *Pharmacoeconomics* 2000; **17**:461–477.