HEALTH ECONOMICS
Health Econ. **16**: 1009–1023 (2007)
Published online 18 December 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/hec.1199

MONTE CARLO PROBABILISTIC SENSITIVITY ANALYSIS FOR PATIENT LEVEL SIMULATION MODELS: EFFICIENT ESTIMATION OF MEAN AND VARIANCE USING ANOVA

ANTHONY O'HAGAN[a,*], MATT STEVENSON[b] and JASON MADAN[b]

[a] *Department of Probability and Statistics, University of Sheffield, Sheffield S3 7RH, UK*
[b] *School of Health and Related Research, University of Sheffield, Sheffield S1 4DA, UK*

SUMMARY

Probabilistic sensitivity analysis (PSA) is required to account for uncertainty in cost-effectiveness calculations arising from health economic models. The simplest way to perform PSA in practice is by Monte Carlo methods, which involves running the model many times using randomly sampled values of the model inputs. However, this can be impractical when the economic model takes appreciable amounts of time to run. This situation arises, in particular, for patient-level simulation models (also known as micro-simulation or individual-level simulation models), where a single run of the model simulates the health care of many thousands of individual patients. The large number of patients required in each run to achieve accurate estimation of cost-effectiveness means that only a relatively small number of runs is possible. For this reason, it is often said that PSA is not practical for patient-level models.

We develop a way to reduce the computational burden of Monte Carlo PSA for patient-level models, based on the algebra of analysis of variance. Methods are presented to estimate the mean and variance of the model output, with formulae for determining optimal sample sizes. The methods are simple to apply and will typically reduce the computational demand very substantially. Copyright © 2006 John Wiley & Sons, Ltd.

Received 8 August 2005; Revised 11 September 2006; Accepted 25 October 2006

KEY WORDS: analysis of variance; cost-effectiveness; economic evaluation; economic model; individual-level simulation; micro-simulation; Monte Carlo; patient-level model; probabilistic sensitivity analysis

INTRODUCTION

Background

Probabilistic sensitivity analysis (PSA) is increasingly demanded by health care regulators and reimbursement agencies when assessing the cost-effectiveness of technologies based on economic modelling, in order to recognise the uncertainties in parameters and data (National Institute for Clinical Excellence, 2004; Claxton *et al*., 2005). The economic evaluation of competing technologies is generally conducted with the aid of an economic model that synthesises knowledge about a variety of inputs derived from available information sources. PSA entails specifying a joint probability distribution to characterise uncertainty in the model's inputs and propagating that uncertainty through the model to derive probability distributions for its outputs (such as population mean costs or incremental net benefit); see Doubilet *et al*. (1986), Briggs *et al*. (2002), O'Hagan *et al*. (2005). The usual way to propagate the uncertainty is the Monte Carlo method, whereby random values of the model input parameters are simulated and the model is run for each simulated parameter set. The resulting sample of

*Correspondence to: Department of Probability and Statistics, University of Sheffield, Sheffield S3 7RH, UK. E-mail: a.ohagan@sheffield.ac.uk

outputs characterises the output uncertainty, and to obtain accurate PSA we typically need 1000 or more model runs.

Although most economic modelling has used cohort models, there is increasing use of patient-level simulation models (also known as micro-simulation or individual-level simulation models), as exemplified by Szeto and Devlin (1996), Paltiel *et al.* (1998), Chilcott *et al.* (2001), Davies *et al.* (2002), Brennan *et al.* (2004), Barton *et al.* (2004) and Stevenson *et al.* (2005a). The output of a cohort model is the appropriate measure of cost-effectiveness for the entire treated population. In contrast, a patient-level model simulates treatment and response pathways for individual patients, and the outputs are mean costs, effectiveness or cost-effectiveness measures for a sample of individuals. It is often said that we cannot do PSA by Monte Carlo for a patient-level model because the time required to run it for each set of sampled input parameter values means that it is not practical to perform the large number of runs needed for Monte Carlo PSA; see, for example, Stevenson *et al.* (2004), Claxton *et al.* (2005). The lengthy computation time is due to the need to simulate a very large number of patients in order for the simulated sample to give an accurate value for the population cost-effectiveness measure for each input parameter set. The thrust of this article is that there is another way, the analysis of variance (ANOVA) approach, that is simple to use and requires very much less computation.

The remainder of this section defines some basic notation and considers the particular example where model output is incremental net benefit, while the next section, 'Sample size criteria,' introduces the criteria to be used for determining sample sizes. The following section, 'Standard Monte Carlo PSA,' presents the standard Monte Carlo approach to PSA for patient-level models, including analysis of the number of patients required per run and the number of runs required to achieve any desired accuracy in the main PSA analyses. The section 'One way ANOVA' develops the ANOVA theory for more efficient simulation, based on using a smaller number of patients in each run. Estimators for the mean and variance of the model output are derived, with formulae for the optimal number of patients per run and the number of runs required to achieve desired accuracy. The methods are illustrated on a model for osteoporosis in the 'Example' section. The final section discusses incremental cost-effectiveness ratios, alternatives to Monte Carlo and directions for further research. Some technical details are given in the Appendix.

Notation

We suppose that the model simulates independent patients. That is, the patients and their pathways do not interact. Some discussion of the case of non-independent patients can be found in the section 'Non-independent patients'.

Let $\mathbf{x}$ denote the vector of model input parameters, whose uncertainty we wish to account for in the PSA. Let $y(\mathbf{x})$ denote the 'true' model output for input vector $\mathbf{x}$. In cost-effectiveness analysis, $y(\mathbf{x})$ may be the population mean cost, effectiveness or net benefit (but see the discussion about the incremental cost-effectiveness ratio in the section 'ICER'). In a patient-level model, however, we never actually observe $y(\mathbf{x})$. Instead, the model produces for each simulated patient a value z that is $y(\mathbf{x})$ plus noise. Since $y(\mathbf{x})$ is the population mean (i.e. averaged over a large population of patients), the noise has zero expectation.

In a Monte Carlo PSA, let $\mathbf{x}_i$ denote the ith sampled parameter set, and let z_{ij} denote the output value for the jth individual patient in the model run using inputs $\mathbf{x}_i$. The subscript i ranges from 1 to N, the number of parameter sets sampled in the PSA, i.e. the number of model runs. The subscript j runs from 1 to n, the number of patients simulated in each model run. We denote the mean output for run i by $\bar{z}_i = (1/n)\sum_{j=1}^{n} z_{ij}$, and the mean over all Nn patients in all model runs by $\bar{z} = (1/N)\sum_{i=1}^{N} \bar{z}_i$.

We have assumed for clarity that the same number of patients will be simulated in each run. This is the usual situation, although the theory can be generalised to the case of unequal numbers; see the section 'Unbalanced sampling and heterogeneity of patient-level variance'.

DOI: 10.1002/hec

The purpose of PSA is to derive relevant properties of the probability distribution of $y(\mathbf{X})$. Notice that $\mathbf{X}$ here is a capital letter, denoting that it is a random variable. The distribution of $y(\mathbf{X})$ is the distribution that would be obtained if we were able to compute $y_i = y(\mathbf{x}_i)$ for a very large sample of parameter sets $\mathbf{x}_i$. The two most important aspects of that distribution are its mean,

$$\mu = E(y(\mathbf{X}))$$

and its variance,

$$\sigma^2 = \mathrm{var}(y(\mathbf{X}))$$

Their interpretations are that μ is the best estimate of the output y allowing for uncertainty in the model inputs, while σ^2 describes the uncertainty around that estimate due to input uncertainty. Our analysis concentrates on methods to estimate μ and σ^2.

The variance σ^2 is often referred to as due to *second-order* uncertainty, i.e. uncertainty in the parameters of the economic model. However, in the context of patient-level modelling, so-called *first-order* uncertainty is also important. This arises from the variability between patients in the population. Generally, we let $\tau^2(\mathbf{x})$ be the patient-level variance for simulations of patients with parameters $\mathbf{x}$, and let

$$\bar{\tau}^2 = E(\tau^2(\mathbf{X}))$$

be the mean value of $\tau^2(\mathbf{x})$ averaged with respect to the uncertainty in $\mathbf{X}$. In general, the larger the patient-level variability the more patients we will need to sample in each run. We let k denote the ratio of these two variances,

$$k = \bar{\tau}^2/\sigma^2$$

so that $\bar{\tau}^2 = k\sigma^2$.

Incremental net benefit

Although the individual patient output z might be any measure of cost, effectiveness or cost-effectiveness, it will be helpful to keep in mind as an example the case where the model is comparing two treatments and z is the incremental net benefit for treatment 2 over treatment 1 for this patient. This is defined as

$$z = \lambda \times \Delta e - \Delta c \tag{1}$$

where Δe is this patient's increment in effectiveness, Δc is the patient's increment in costs and λ is the willingness to pay coefficient, expressing the monetary value to the health care provider of one unit increase in effectiveness. Then y is the population mean incremental net benefit (Stinnett and Mullahy, 1998), and treatment 2 is cost-effective relative to treatment 1 if $y > 0$. One role of PSA is then to quantify the uncertainty in whether $y > 0$. The mean μ is the best estimate of the population mean incremental net benefit $y(\mathbf{X})$, and if a decision is required to use one treatment or the other then Claxton (1999) points out that it should be to use treatment 2 if $\mu > 0$. The variance σ^2 describes uncertainty in this decision. For instance, if μ is positive but σ is not small relative to μ, then there is an appreciable risk that the decision to use treatment 2 will be found to be wrong because $y(\mathbf{X})$ is really negative. Conversely, if the absolute value of μ is large relative to σ (for instance, 3σ or more) then there is very low decision uncertainty.

SAMPLE SIZE CRITERIA

We will develop theory here both for the standard Monte Carlo estimators of μ and σ^2 and for a new estimator of σ^2 that corrects for bias in the standard estimator. We will also derive formulae for

DOI: 10.1002/hec

choosing sample sizes – both the number n of patients sampled in each model run (i.e. for each simulated parameter set) and the number N of model runs. In this section, we will denote estimators of μ and σ^2 by $\hat{\mu}$ and $\hat{\sigma}^2$; in the next two sections subscripts will be used to designate particular estimators.

Sample sizes will be chosen sufficiently large to achieve the following three conditions.

- *Accuracy of* $\hat{\mu}$: The primary focus of the cost-effectiveness analysis is μ, which is the best estimate of the cost-effectiveness output y in the light of input uncertainty. Our first criterion is that we wish to estimate μ with standard deviation less than or equal to d, so that a 95% interval has width no more than $\pm 1.96d$. However, in the context where the model output is incremental net benefit, as discussed in the previous section, interest will focus on the magnitude of μ relative to σ. Then it is appropriate to set d to some small multiple of σ, so that the uncertainty in the estimate of μ does not cloud the assessment of whether its absolute value is large enough relative to σ to imply low decision uncertainty. We therefore set $d = c_1\sigma$, requiring that
$$\mathrm{var}(\hat{\mu}) \leq c_1^2\sigma^2 \tag{2}$$
- *Accuracy of* $\hat{\sigma}^2$: Although μ is a key component of the cost-effectiveness analysis, the primary objective of PSA is to identify the amount of uncertainty in the model output, which is measured by σ^2. It is usual to require accuracy of variance estimates to be expressed in terms of the coefficient of variation. We therefore require that the coefficient of variation be less than or equal to c_2. This can be expressed as
$$\mathrm{var}(\hat{\sigma}^2) \leq c_2^2 E(\hat{\sigma}^2)^2 \tag{3}$$
- *Bias of* $\hat{\sigma}^2$: Finally, in the case of the standard Monte Carlo estimates, we will require the bias in estimating σ^2 to be small. A natural objective is to make the bias in $\hat{\sigma}^2$ small compared with the width of a confidence interval for σ^2. Requiring the ratio of the bias to the width of the 95% confidence interval to be less than or equal to c_3 gives the condition
$$|E(\hat{\sigma}^2) - \sigma^2| \leq c_3 \times 1.96\sqrt{\mathrm{var}(\hat{\sigma}^2)} \tag{4}$$

We therefore have three conditions to be met, defined by the three criteria c_1, c_2 and c_3. These can in principle be set to any values, according to the needs of any particular application, and we will give formulae specifying n and N in terms of these. However, we suggest that in practice the following might be reasonable simplifications.

First, when the interest is in incremental net benefit we would generally wish to have both μ and σ estimated to comparable precision. With coefficient of variation for estimating σ^2 set to c_2, the precision in σ will be of the order of $c_2/2$, so setting $c_2 = 2c_1$ may be appropriate in this case.

Second, so that the bias in the standard Monte Carlo estimator of σ^2 should be negligible relative to the uncertainty in the estimator, we suggest setting c_3 to about 0.05.

Note that the resulting choices for n and N will depend on the ratio k of the underlying variances σ^2 and τ^2 to which the data is subject, and it is therefore necessary to obtain initial estimates or guesses in order to apply the formulae.

STANDARD MONTE CARLO PSA

Standard MC estimators

In conventional economic models without patient-level simulation, we observe $y_i = y(\mathbf{x}_i)$ in run i, and the Monte Carlo estimators of μ and σ^2 are, respectively, $\bar{y} = (1/N)\sum_{i=1}^{N} y_i$ and $s^2 = \{1/(N-1)\}$ $\sum_{i=1^{N}(y_i-\bar{y})^2}$. These estimators are unbiased. The standard approach to using Monte Carlo with patient-level models is to make n large enough so that each $\bar{z}_i$ is deemed to be a sufficiently accurate computation of y_i, and then to apply the usual estimators. Hence, we have

$$\hat{\mu}_S = \bar{z}, \quad \hat{\sigma}_S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{z}_i - \bar{z})^2 \tag{5}$$

The subscript S here indicates that these are the standard Monte Carlo estimates. The following results are derived in Appendix A.

$$E(\hat{\mu}_S) = \mu \tag{6}$$

$$\mathrm{var}(\hat{\mu}_S) = \frac{\sigma^2}{N} + \frac{\bar{\tau}^2}{Nn} \tag{7}$$

$$E(\hat{\sigma}_S^2) = \sigma^2 + \bar{\tau}^2/n \tag{8}$$

$$\mathrm{var}(\hat{\sigma}_S^2) = \frac{2}{N-1}(\sigma^2 + \bar{\tau}^2/n)^2 \tag{9}$$

Equation (6) shows that $\hat{\mu}_S$ is an unbiased estimator of μ, and its variance (7) decreases with N in the usual way. However, Equation (8) implies that the standard Monte Carlo estimator $\hat{\sigma}_S^2$ is *biased*. Its bias is $\bar{\tau}^2/n$, which is always positive, so on average it over-estimates σ^2. The bias arises because variability in the $\bar{z}_i$s inflates their variance, over and above the variability in the true means $y(\mathbf{x}_i)$ that is represented by σ^2. The main reason for using a large n is to make this bias small.

Note that the final equation, the variance of $\hat{\sigma}_S^2$, applies for large n, but the other results are exact.

Sample sizes for standard estimators

We now identify values of n and N that would be required to satisfy conditions (2)–(4) for the standard Monte Carlo estimators $\hat{\mu}_S$ or $\hat{\sigma}_S^2$. Although these estimators are widely used in PSA of economic models, we do not believe that these explicit sample size calculations have been presented before in this context.

From (7), condition (2) becomes

$$\frac{\sigma^2}{N} + \frac{\bar{\tau}^2}{Nn} \le c_1^2\sigma^2$$

from which we have $N \ge (\sigma^2 + \bar{\tau}^2/n)/(c_1^2\sigma^2)$, and therefore

$$N \ge (1 + k/n)/c_1^2 \tag{10}$$

Next, from (8) and (9), condition (3) becomes

$$\frac{2}{N-1}(\sigma^2 + \bar{\tau}^2/n)^2 \le c_2^2(\sigma^2 + \bar{\tau}^2/n)^2$$

and therefore

$$N \ge 1 + 2/c_2^2 \tag{11}$$

If we adopt the simplification that $c_2 = 2c_1$, as suggested in the section 'Sample size criteria' then by comparing (10) and (11) we can see that the former is the more stringent requirement. This suggests that in general the number of runs in standard Monte Carlo should normally be chosen to satisfy the requirement (10) for accurate estimation of the mean μ.

Next, condition (4) leads to the choice of n, via (8). In Appendix A, it is shown that this approximates to

$$n \ge 0.36k\sqrt{N}/c_3 \tag{12}$$

If we let $c_3 = 0.05$, as suggested in the section 'Sample size criteria', then we have

$$n \geq 7.2k\sqrt{N} \tag{13}$$

Finally, we note that the total number of patients to be sampled is Nn. Adopting the above recommendations, and noting that (10) gives a value for N that is certainly greater than $1/c_1^2$, we see that the total number of patients will be greater than $7.2k/c_1^3$.

ONE WAY ANOVA

Using fewer patients per run

If the only objective of the PSA were to be estimating μ, then the following argument shows that the approach of using a large number of patients in each run would be far from optimal. The derivation of the mean and variance of $\hat{\mu}_S$ in Equations (6) and (7) does not depend on using a large n, and in particular we see that $\hat{\mu}_S$ is unbiased for any n. Now suppose that the number of patients that we can run in total is fixed, say $Nn = M$. To estimate μ as accurately as possible we should try to minimise var($\hat{\mu}_S$), which from (7) is equal to $\sigma^2/N + \bar{\tau}^2/M$. Minimising this variance for fixed M means making N as large as possible. Therefore the most efficient way is to make $n = 1$, i.e. to sample just one patient per parameter set. We then get var$(\hat{\mu}_S) = \sigma^2/M + \bar{\tau}^2/M$.

The problem with only sampling one patient per parameter set is that we cannot separate σ^2 from $\bar{\tau}^2$, and so we cannot estimate σ^2. In practice, PSA is performed not only to estimate μ but also to estimate output uncertainty, as described in particular by σ^2. However, we now consider how by accepting a smaller number of patients per run, and by correcting the resulting bias in the estimate of σ^2, we can reduce the overall computational load to perform PSA on patient-level models.

Estimate of σ^2 and its variance

The one-way analysis of variance in frequentist statistical theory allows us to estimate σ^2 and $\bar{\tau}^2$ separately. Define the usual within-groups and between-groups sums of squares

$$S_w = \sum_{i=1}^{N} \sum_{j=1}^{n} (z_{ij} - \bar{z}_i)^2, \quad S_b = n \sum_{i=1}^{N} (\bar{z}_i - \bar{z})^2$$

so that in particular the standard Monte Carlo estimator of σ^2 is $\hat{\sigma}_S^2 = S_b/\{(N-1)n\}$. Then we find

$$E(S_w) = N(n-1)\bar{\tau}^2, \quad E(S_b) = (N-1)n\sigma^2 + (N-1)\bar{\tau}^2$$

So, provided $n > 1$, an unbiased estimator of σ^2 is

$$\hat{\sigma}_A^2 = \frac{1}{n}\left(\frac{S_b}{N-1} - \frac{S_w}{N(n-1)}\right) \tag{14}$$

which is $\hat{\sigma}_S^2$ minus a estimate of the bias.

The fact that we can produce a simple unbiased estimator of σ^2 without simulating huge numbers of patients for each run is a valuable result. However, we need also to ask how good this estimator is.

One immediate problem with $\hat{\sigma}_A^2$ is that it can be negative. Factors that increase this risk are

- when $\bar{\tau}^2$ is large relative to σ^2, and
- when n is small.

The first of these will often arise in patient-level simulation models, where variability between patients is much larger than the variability induced by uncertainty over model inputs. The second means that taking very few patients per run may not be wise.

DOI: 10.1002/hec

We can approximate the sampling variance of $\hat{\sigma}_A^2$ by supposing that S_w and S_b have independent chi-square sampling distributions with degrees of freedom $N(n-1)$ and $N-1$. This assumption is correct if the distributions of $y(\mathbf{X})$ and of the patient-level variability are normal, and if $\tau^2(\mathbf{x}) = \bar{\tau}^2$ for all $\mathbf{x}$; otherwise it may still be a reasonable approximation, although variability in the $\tau^2(\mathbf{x})$ values will certainly increase the variance of $\hat{\sigma}_A^2$.

Under the assumed independent chi-square distributions, the variance of $\hat{\sigma}_A^2$ becomes

$$\mathrm{var}(\hat{\sigma}_A^2) = 2\left(\frac{(\sigma^2 + \bar{\tau}^2/n)^2}{N-1} + \frac{\bar{\tau}^4}{Nn^2(n-1)}\right) \tag{15}$$

Optimal allocation of N and n

The new method will work for any choices of n and N. We will wish to choose these so as to satisfy conditions (2) and (3) for the estimators $\hat{\mu}_S$ and $\hat{\sigma}_A^2$. However, condition (4) no longer applies because $\hat{\sigma}_A^2$ is unbiased, and this gives us extra flexibility. Note that the total sampling effort is represented by $M = Nn$, the total number of patients to be sampled. It is possible to optimally choose the balance between n and N in order to minimise the total sampling effort required to achieve any desired accuracy in the estimators. The results in this section are obtained as follows. First we identify the number n of patients to be sampled in each run in order to minimise (15) for fixed M. Then we find the minimal M to satisfy condition (3). These two steps give optimal values of N and n, and we find that they also satisfy (2) when, as suggested in the section 'Sample size criteria', we set $c_2 = 2c_1$. Full details of these derivations are given in Appendix A, and we report here the key results.

First, the optimal allocation of n for given total sampling effort M is

$$n = \frac{M(1+k)+k}{M+2k} \tag{16}$$

Then with this value of n the total sampling effort required to satisfy (3) is

$$M = \frac{1}{2c_2^2}\left(c_2^2 + 2 + 8k + \sqrt{c_2^4 + 4c_2^2 + 16c_2^2k + 4 + 32k + 64k^2 + 32c_2^2k^2}\right) \tag{17}$$

These two values determine $N = M/n$. (Both N and n should be rounded up to integer values.)

For most practical purposes, we can use the following simple approximations to the above formulae:

$$M = 8k/c_2^2 \tag{18}$$

$$n = 1 + k \tag{19}$$

These approximations will be sufficiently accurate whenever k is at least 25 and c_2 is less than or equal to 0.2.

Although this theory has been developed under an assumption of normality and heteroscedasticity, we suggest that $n = 1 + k$ and $N = 8/c_2^2$ are likely to be good choices generally. Note also that the optimal n should minimise the risk of obtaining a negative estimate of σ^2 (since it is the coefficient of variation of the estimator that is actually being minimised).

Summary of the ANOVA method

We can summarise all the above results in the following simple steps. Note that for steps 1 and 2 we need to have a prior estimate of $k = \bar{\tau}^2/\sigma^2$, which is discussed in the section 'Implementation' below:

1. Given a desired sampling precision c_2 for estimating σ^2, choose M using Equation (17) or its simple form (18).
2. Now choose n using (16) or its simple form (19), and set $N = M/n$.

3. Carry out the Monte Carlo sampling with these choices of N and n (rounded up to integer values).
4. Estimate μ by $\hat{\mu}_S = \bar{z}$. Estimate σ^2 by $\hat{\sigma}_A^2$, using (14).
5. The variances of these estimators are given by (7) and (15), respectively. These can be estimated by substituting into them the estimate $S_w/\{N(n-1)\}$ for $\bar{\tau}^2$, $\hat{\sigma}_A^2$ for σ^2, and the ratio of these for k.

If in step 1 the required overall sampling effort M is impractically large, the method can still be followed through by using whatever M can realistically be resourced. With the prior estimate of k, we can estimate that this M will achieve the approximate coefficient of variation $c_2 = \sqrt{8k/M}$.

Efficiency gain over standard Monte Carlo

We found in the section 'Sample sizes for standard estimators' that the appropriate values for N and n using the standard Monte Carlo approach would yield a total sampling load of $M = Nn = 7.2k/c_1^3$, at least in the case where the model output is incremental net benefit and the simplifications of the section 'Sample size criteria' are adopted. The above analysis yields a value of $M = 8k/c_2^2$ with the ANOVA method. Under the suggested relationship $c_2 = 2c_1$, the latter becomes $2k/c_1^2$. Therefore the gain in efficiency is shown by a typical reduction in overall sampling by a factor of $3.6/c_1$. For acceptable precision in estimating μ, in practice, we will usually require c_1 to be 0.1 or less, implying an efficiency gain of 36 times or more.

The particular values suggested here may not be appropriate in any particular application, but it is simple to apply the formulae developed here to identify values of n and N for any criteria values c_1, c_2 and c_3, both for standard Monte Carlo estimation and for the ANOVA approach. The resulting efficiency gain with the ANOVA approach may be more or less than the suggested factor of 36, but is always likely to be substantial. We suggest that the fact that the ANOVA method requires much less overall computing effort will make it a feasible way to perform PSA in many models for which the standard Monte Carlo approach is impractical.

Implementation

The theory of optimal allocation requires that we know the ratio $k = \bar{\tau}^2/\sigma^2$, which of course in practice will be unknown. It is necessary first to obtain a prior estimate of k, which in itself may be difficult for a large patient-level simulation model. In practice, it is natural to obtain estimates from a preliminary PSA using a small number of runs. Even a modest preliminary PSA will produce a good estimator of $\bar{\tau}^2$ in the form of $S_w/\{N(n-1)\}$. The key to implementing the method is to obtain a realistic initial estimate of σ^2.

First we suggest that the model input values for the initial set of runs should not be chosen randomly. In order to obtain a useful estimate of σ^2 it is important that the input parameter sets used should give broad coverage of the input space. Whereas Monte Carlo sampling will do this if we have large numbers of input parameter sets, in a small preliminary PSA it makes sense to deliberately choose model input sets that are well separated. The lack of random sampling invalidates results about unbiasedness and variances, but this is not important at this preliminary stage. We suggest using a small number N of runs, of the order of 25–40, with input sets chosen to be spread broadly across the space of possible inputs, and a moderate to large value of n.

Letting u denote the number of uncertain input parameters for the model, the space of possible input sets is u-dimensional. We might consider a design for the input sets based on a grid of points in this space. If we use p values for each input, then a grid of all possible combinations would have $N = p^u$ parameter sets. This can be enormous, but if u is small we can get manageable designs. For instance, if $u = 2$ we could have $p = 5$ values for each input ($N = 25$). For $u = 3$, $p = 3$ ($N = 27$), and for $u = 5$, $p = 2$ ($N = 32$) we also obtain reasonable choices. However, u will often be much larger than this, and a complete grid (a full factorial design) would be impractical, even with $p = 2$.

Fractional factorial designs with $p = 2$ or 3 have been developed in the experimental design literature, and would certainly yield good designs; see for example Box and Draper (1987). However, in practice it may be adequate simply to use a random selection of 20–30 combinations (sampling without replacement) from the 3^u points on a grid with $p = 3$ values for each input.

It may help to constrain the choice so that each level of each factor is used the same number of times. This could be achieved by the following procedure (based on Latin Hypercube sampling). Select 3 sample points by arranging the three levels of each factor in a random order. For instance, with $u = 4$ this might yield the orders (L, H, M), (L, M, H), (H, M, L), (M, H, L) for the four parameters, where L, M and H, respectively, denote the low, mean and high levels of a factor. Then this would give the three sample points (L, L, H, M), (H, M, M, H) and (M, H, L, L), i.e. the first point has inputs 1 and 2 at their low levels, input 3 at its high level and input 4 at its mean level. Repeating this process to generate more sets of three points (and rejecting any set that produces a point that has already been chosen) will yield sample designs with the desired balance.

Having chosen a suitable design of N input sets and carried out the model runs, we can use the ANOVA method to estimate both $\bar{\tau}^2$ and σ^2. However, some correction is needed when estimating σ^2, to allow for the spread of the points. In Monte Carlo sampling, the input parameter values are sampled from their appropriate distributions, and the resulting points naturally have the right spread. If we choose points in the preliminary PSA that are more or less spread than this, the estimate of σ^2 will tend to be too large or too small.

Let v_j be the variance of the p values chosen for input j, expressed as a proportion of the correct variance for that input according to its PSA distribution. For instance, if $p = 3$ and we choose values for this input that are equal to its mean (value M) and its mean plus and minus s standard deviations (values L and H), then the variance of these three points is $v_j = 2s^2/3$ times the underlying variance for this parameter. If we now suppose that the output of the economic model is roughly linear in its inputs, then the ANOVA estimate of σ^2 should be divided by the product of the v_js in order to correct for the spread of the design points.

In the light of more experience with the ANOVA method, it will no doubt be possible to improve on the above rather tentative suggestions.

EXAMPLE

To illustrate the sample size calculations, we consider a large model developed at Sheffield for assessing the cost-effectiveness of many treatments for osteoporosis (Stevenson *et al.*, 2005a). For this example, we chose to compare alendronate, a bisphosphonate costing £301 per annum, with no treatment. The patient population was defined to be women without a prior clinical fracture and a T-Score of −2.5SD. Stevenson *et al.* (2005b) estimate the relative risk of fracture by using alendronate (with 95% uncertainty interval) to be 0.46 (0.23–0.91) at the hip, 0.53 (0.42–0.67) at the vertebrae and 0.48 (0.31–0.75) at the wrist. Other inputs to the model were the costs and disutilities associated with fracture, which for the purposes of this analysis were fixed at their central estimates. Our output measure was the incremental net benefit (INB) at a willingness to pay threshold of £30 000 per QALY. We wish to conduct PSA to assess uncertainty in the INB due to uncertainty in the three relative risk parameters.

An initial run of the osteoporosis model was made with relative risk inputs set at their mean values (which we will denote by $\mathbf{x}_0$) and with 15 000 patients. This yielded a mean INB of 1308.2 and a patient-level variance of about 2.4×10^9. The decision to stop sampling at 15 000 patients was based on the fact that the standard error of the mean is the square root of $2.4 \times 10^9/15\,000$, i.e. 400, which is small

DOI: 10.1002/hec

enough relative to the observed mean of 1308 to be confident that the true mean incremental net benefit $y(\mathbf{x}_0)$ is positive.

A deterministic cost-effectiveness analysis of this kind does not take account of uncertainty in the input parameters. It is therefore necessary to perform a PSA for the usual two reasons: first, to estimate μ, recognising that because of non-linearity this will generally be different from $y(\mathbf{x}_0)$; second, to assess the uncertainty in the estimate of μ, as measured by σ^2.

A further 26 runs of the model were performed, also with 15 000 patients per run. Together with the initial baseline run, the 27 runs comprised a 3^3 factorial design as discussed in the section 'Implementation'. Each fracture probability input was set at three levels – its mean value and its mean value plus or minus one standard deviation – so that in the notation of the section 'Implementation' we had $s = 1$ for each input and hence $v_j = \frac{2}{3}$. This design was intended to provide initial indications of sensitivity to each input, but also serves to give a rough estimate of σ^2. It was found that the patient-level variance was $S_w/\{N(n-1)\} = 2.38 \times 10^9$ averaged over all of the runs (and apparently constant across runs), and so this is an initial estimate of $\bar{\tau}^2$. The variance between the means of these 27 runs was found to be $S_b/\{n(N-1)\} = 219\,429$. Subtracting the estimated bias of $2.38 \times 10^9/15\,000 = 158\,667$ (in effect, applying Equation (14)) gives an initial estimate of 60 762 for the underlying variance across these 27 runs. In order to convert this to an estimate of σ^2, we divide by $(2/3)^3$, as discussed in the section 'Implementation'. We therefore estimate σ^2 by $60\,762 \times 1.5^3 = 205\,072$. This in turn suggests a value for k of $2.38 \times 10^9/205\,072 = 11\,606$. On this basis it was decided to perform the main PSA using 10 000 patients per run.

Each run of 10 000 patients takes about 50 min on a fast PC, so the PSA will still be highly computer intensive. We had resources to make 500 model runs. If 10 000 patients per run were indeed optimal, this would enable us to estimate σ^2 with a coefficient of variation $c_2 = \sqrt{8/500} = 0.126$, so σ^2 will be estimated to within about $\pm 25\%$. Our main analysis is therefore based on $N = 500$ runs using fracture probability inputs randomly sampled from their uncertainty distributions, and $n = 10\,000$ patients per run. From these data, we found $\bar{z} = 879.2$, $S_w = 1.1935658 \times 10^{16}$ and $S_b = 230\,495\,731$. Thus, the estimate of μ is 879.2, the estimate of $\bar{\tau}^2$ is $S_w/(500 \times 999) = 2.387 \times 10^9$ and we obtain $\hat{\sigma}_A^2 = 223\,178$.

The resulting estimate of k is the ratio of these last two estimates, 10 695, so the optimal number of patients per run would be approximately 10 700, which is fortuitously close to the original estimate of 11 600 and to the 10 000 we actually used.

It is appropriate now to ask how accurate the estimates of μ and σ^2 are, and how much sampling has been saved by using the ANOVA method. The estimate of μ has variance $(\sigma^2 + \bar{\tau}^2/n)/N$, which is estimated by $S_b/\{N(N-1)\} = 923.8$, corresponding to a standard error of 30.4. So a 95% interval for μ is approximately $879.2 \pm 60.8 = [818.4, 940.0]$. Using Equation (15), we obtain an estimated standard deviation for $\hat{\sigma}_A^2$ of 29 244, so an approximate 95% interval for σ^2 is $223\,178 \pm 58\,488 = [164\,690, 281\,666]$. As expected, the interval has range approximately $\pm 25\%$. The corresponding estimate and 95% interval for σ become 472.4 and [405.8, 530.7]. The interval has range approximately $\pm 13\%$. These various estimates and intervals are the primary results of the PSA.

To confirm the efficiency of the ANOVA method, consider how much sampling would have been required with the standard Monte Carlo method to achieve results of comparable accuracy. First, to obtain the same accuracy in estimating μ, we set $c_1^2 = 923.8/223\,178 = 0.00414$, and hence we would need $N > 1/c_1^2$, or at least $N = 242$ runs of the model. Then to meet the bias condition (4), and using the suggested setting of $c_3 = 0.05$, formula (13) applies and we would need $n = 7.2k\sqrt{N} = 7.2 \times 10\,695 \times \sqrt{242}$, or 1.2 million patients per run. Even if such huge numbers of patients could be handled in each run, the total number of patients simulated would have been nearly three hundred million and would have taken more than two years of solid computation. Our actual analysis used 500 runs of 10 000 patients each, or 5 million patients in all, which represents almost a 60-fold saving in effort (agreeing with the formula in the section 'Efficiency gain over standard Monte Carlo' of $3.6/c_1 = 3.6/\sqrt{0.00414} = 56$).

DOI: 10.1002/hec

DISCUSSION

Principal conclusions

We have presented methods to calculate the key PSA outputs using a patient-level simulation model for cost-effectiveness analysis. These will make it possible to carry out PSA for many such models by the familiar and simple Monte Carlo approach, where hitherto the computational demand was thought to be prohibitive. An important feature of this work has been the derivation of explicit sample size formulae, both for the standard Monte Carlo method and for the new ANOVA method.

The remainder of this section discusses a variety of issues, including some further extensions and alternative computation methods.

ICER

Until relatively recently, cost-effectiveness analysis was almost exclusively based on the incremental cost-effectiveness ratio (ICER), with treatment 2 being deemed more cost-effective than treatment 1 if the ICER was less than λ. There are two reasons why we do not develop a PSA analysis based on the ICER here. First, on a fundamental level, the claim that treatment 2 is more cost-effective if the ICER is less than λ only works if treatment 2 is more effective than treatment 1. Otherwise, the inequality must be reversed; see for example O'Hagan *et al.* (2000). The definition based on incremental net benefit is much cleaner. It also leads to much simpler techniques for accounting for uncertainty, which is our second reason for not analysing the ICER here. In the case of patient-level simulation models, the estimated ICER for a given run is the sample mean incremental cost divided by the sample mean incremental effectiveness. This is not an average of patient-level ratios, and therefore all of the above theory is inapplicable.

Gaussian process emulation

When the economic model is so computer-intensive that even the methods presented here are impractical, there is an even more efficient methodology based on Gaussian process emulation; see O'Hagan *et al.* (1999) for the range of application of these methods. This is a mathematically more advanced technique, and in the absence of user-friendly software is not accessible to most practising health economists. Estimates of μ and σ^2 can be calculated using the methods of Oakley and O'Hagan (2004) and Stevenson *et al.* (2004). A similar approach has been proposed by Cronin *et al.* (1998) based on the more restrictive idea of fitting a response surface to the model output instead of a Gaussian process.

Non-independent patients

The theory has been developed on the assumption that the sampled values z_{ij} for patients $j = 1, 2, \ldots, n$ are independent. Whilst this is true for many patient-level simulation models, it is possible for the value obtained for one patient to depend on those obtained for earlier patients. This will arise, for example, when the simulation takes account of limited availability of resources, so that the outcome for one patient may depend on the utilisation of resources by previous patients in the simulation, as for example in Ratcliffe *et al.* (2001). If patient outcomes are not independent, then a more appropriate modelling approach is by discrete event simulation; see Davies *et al.* (2003), Barton *et al.* (2004) or Brennan *et al.* (2006). The essence of the assumption that patients are simulated independently is that $\text{var}(\bar{z}_i) = \sigma^2 + \bar{\tau}^2/n$. The same formula may be expected to apply to models with interacting patients when the model is in equilibrium, because the variance of the sample mean will still decline proportionally to $1/n$. However, the interpretation of $\bar{\tau}^2$ will change, and it will need to be estimated differently. This is a topic for future research.

DOI: 10.1002/hec

Unbalanced sampling and heterogeneity of patient-level variance

We have assumed that the number of patients sampled in each run is the same, but there are at least two reasons for considering generalising this to the case when n_i patients are sampled in run i. First, when $\tau_i^2 = \tau^2(\mathbf{x}_i)$ varies substantially with the sampled input vector $\mathbf{x}_i$, it should be better to sample more patients in runs where the patient-level variation is found to be larger. Notice that in the ANOVA theory the optimal n effectively implies making $\bar{\tau}^2/n$ equal to σ^2. If there is substantial heterogeneity of patient-level variances, then we conjecture that it would be more efficient to choose n_is to make τ_i^2/n_i equal to σ^2 for each i.

Another situation where unequal n_is will naturally arise is when an initial estimate of k is found to be inaccurate. We have suggested setting n using estimates of σ^2 and $\bar{\tau}^2$ based on a small-scale initial sample. It would be sensible to check this value by re-estimating σ^2 and $\bar{\tau}^2$ part way through the main sampling exercise. If it then seems that a different value of n should be used the subsequent sampling can use the new value. This will lead to a combined sample using two (or more, if further checks are applied) different values of n.

Some of the theory developed here for equal n_is may be readily generalised to unequal values, but again this is a topic for further research.

Note that heterogeneity of the τ_i^2 will affect the validity of the chi-square approximation in the section 'Estimate of σ^2 and its variance' and in particular Equation (15). The chi-square approximation also relies on approximate normality of the patient-level values, and if they have markedly greater or lower kurtosis than the normal distribution then (15) is likely to be a poor approximation.

More than two treatments

Where we have referred to comparing treatments, we have developed methods that apply only for two treatments. It is increasingly common to compare more than two treatments in an economic evaluation. Instead of assuming that the output is incremental net benefit for treatment 2 versus treatment 1, we could handle many treatments by considering as outputs the net benefits for each treatment separately, and by generalising the theory to handle multivariate outputs. This is another topic for future research.

APPENDIX A: FURTHER DETAILS

Means and variances of standard MC estimators

We prove here formulae (6)–(9) for the means and variances of the standard MC estimators $\hat{\mu}_S = \bar{z} = (1/N)\sum_{i-1}^{N} \bar{z}_i$ and $\hat{\sigma}_S^2 = \{1/(N-1)\}\sum_{i=1}^{N}(\bar{z}_i - \bar{z})^2$. Note that $\hat{\mu}_S$ is the sample mean of the $\bar{z}_i$s. Its mean and variance therefore follow from standard results for the mean and variance of a sample mean, using the facts that $E(\bar{z}_i) = \mu$ and $\text{var}(\bar{z}_i) = \sigma^2 + \bar{\tau}^2/n$.

Similarly, $\hat{\sigma}_S^2$ is the sample variance of the $\bar{z}_i$s, and so is an unbiased estimator of the population variance. However, this population variance is that of the $\bar{z}_i$s, i.e. $\sigma^2 + \bar{\tau}^2/n$, which proves (8). Assuming large n, the Central Limit Theorem in statistics ensures that the $\bar{z}_i$s are approximately normally distributed, and hence that $C = (N-1)\hat{\sigma}_S^2/(\sigma^2 + \bar{\tau}^2/n)$ has a chi-squared distribution with $N-1$ degrees of freedom. The variance of a chi-squared random variable is twice the degrees of freedom, hence

$$\begin{aligned}\text{var}(\hat{\sigma}_S^2) &= \text{var}\{(\sigma^2 + \bar{\tau}^2/n)C/(N-1)\}\\ &= \left(\frac{\sigma^2 + \bar{\tau}^2/n}{N-1}\right)^2 \text{var}(C)\\ &= \frac{2}{N-1}(\sigma^2 + \bar{\tau}^2/n)^2\end{aligned}$$

Although this has been proved on the assumption of large n, in practice standard MC estimation is invariably done with n (the number of patients in each run) more than large enough for this approximation to be very accurate.

Choice of n for standard MC estimators

Using (9), condition (4) becomes

$$(\sigma^2 + \bar{\tau}^2/n) - \sigma^2 \le 1.96 c_3 \sqrt{\frac{2}{N-1}(\sigma^2 + \bar{\tau}^2/n)^2}$$

This can be rewritten as

$$\bar{\tau}^2/n \le 1.96 c_3 \sqrt{2/(N-1)}(\sigma^2 + \bar{\tau}^2/n)$$

$$\therefore\ n \ge \frac{1 - 1.96 c_3 \sqrt{2/(N-1)}}{1.96 c_3 \sqrt{2/(N-1)}} k$$

In practice, N will be large and the numerator of the fraction will be approximately 1, leading to

$$n \ge \frac{k}{1.96 c_3 \sqrt{2/(N-1)}} = \frac{1}{1.96 c_3 \sqrt{2}} k\sqrt{N-1} = 0.36 k\sqrt{N-1}/c_3$$

Equation (12) results from the further approximation of $\sqrt{N-1}$ by $\sqrt{N}$.

Optimal sample sizes for ANOVA estimators

First, using the definitions of $k = \bar{\tau}^2/\sigma^2$ and $M = Nn$ we can rewrite (15) as

$$2\sigma^4\left(\frac{(n+k)^2}{n(M-n)} + \frac{k^2}{nM(n-1)}\right) = 2\sigma^4\left(\frac{(M+k)^2}{M(M-n)} + \frac{k^2}{M(n-1)} - 1\right)$$

To minimise this with respect to n for fixed M, we differentiate with respect to n to give

$$2\sigma^4\left(\frac{(M+k)^2}{M(M-n)^2} - \frac{k^2}{M(n-1)^2}\right) \tag{A1}$$

We then equate this to zero, giving

$$(n-1)^2(M+k)^2 - k^2(M-n)^2 = 0$$

This is a quadratic equation in n which has two solutions, $n = (M + k + Mk)/(M + 2k)$ and $n = -(-M - k + Mk)/M$. It is straightforward to confirm that the first of these, Equation (16), is the required solution, by differentiating (A1) again with respect to n and checking that it is negative at $n = (M + k + Mk)/(M + 2k)$.

With this choice of n, (15) reduces to

$$\text{var}(\hat{\sigma}_A^2) = 2\sigma^4 \frac{M + 4k(M+k)}{M(M-1)}$$

Condition (3) now requires us to choose M so that $\text{var}(\hat{\sigma}_A^2) \le c_2^2\sigma^4$. Setting $\text{var}(\hat{\sigma}_A^2) = c_2^2\sigma^4$ yields another quadratic equation, this time in M:

$$M(M-1)c_2^2 - M - 4k(M+k) = 0$$

The left-hand side is negative at $M = 0$ and becomes positive for sufficiently large M, so there is a single positive solution, which we find to be Equation (17).

Now consider the condition (2), that $\mathrm{var}(\hat{\mu}_S) \leq c_1^2\sigma^2$. Given the chosen value of n and equation $\mathrm{var}(\hat{\mu}_S) = (\sigma^2 n + \bar{\tau}^2)/M$, we now wish to solve the equation

$$c_1^2\sigma^2 = \frac{\sigma^2}{M}\left(\frac{M(1+k)+k}{M+2k} + k\right)$$

which yields another quadratic equation for M:

$$c_1^2 M(M+2k) - M(1+2k) - k(2k+1) = 0$$

whose positive solution is

$$M = \frac{1}{2c_1^2}\left(1 + 2k - 2c_1^2 k + \sqrt{1 + 4k + 4k^2 + 4c_1^4 k^2}\right) \tag{A2}$$

If we suppose that k is large and c_1^2 small, then the square root is approximately $2k$, and (A2) is approximately $2k/c_1^2$. In the case where the output is incremental net benefit, we have argued in the section 'Sample size criteria' that we should set $c_1 = c_2/2$, in which case this equates to (18). Therefore, the same values (19) and (18) that optimise the number of patients per run and achieve the desired accuracy for estimating σ^2 will also achieve the associated accuracy for estimating μ. Notice, for instance, that the 95% intervals for μ and σ in the section 'Example' have approximately the same width.

ACKNOWLEDGEMENTS

This research was stimulated by a Focus Fortnight on patient-level simulation models organised by the Centre for Bayesian Statistics in Health Economics (CHEBS) at Sheffield in October 2004. The authors are grateful for discussions with Ruth Davies and Stephen Chick, and particularly to Alan Brennan for many helpful comments. We also thank the referees for some helpful and perceptive comments that have appreciably improved the presentation of this paper.

CHEBS has received funding from AstraZeneca and Merck. The research reported here received no other external support and the authors have no conflicts of interest. No ethical approvals were required.

REFERENCES

Barton P, Bryan S, Robinson S. 2004. Modelling in the economic evaluation of health care: selecting the appropriate approach. *Journal of Health Service Research Policy* **9**: 110–118.

Barton P, Jobanputra P, Wilson J, Bryan S, Burls A. 2004. The use of modeling to evaluate new drugs for patients with a chronic condition: the case of antibodies against tumour necrosis factor in rheumatoid arthritis. *Health Technology Assessment* **8**: 1–104.

Box GEP, Draper NR. 1987. *Empirical Model Building and Response Surfaces*. Wiley: New York.

Brennan A, Bansback N, Reynolds A, Conway P. 2004. Modelling the cost-effectiveness of etanercept in adults with rheumatoid arthritis in the UK. *Rheumatology* **43**: 62–72.

Brennan A, Chick SE, Davies R. 2006. A taxonomy of model structures for economic evaluation of health technologies. *Health Economics* **15**(12): 1295–1310.

Briggs AH, Goeree R, Blackhouse G, O'Brien BJ. 2002. Probabilistic analysis of cost-effectiveness models: choosing between treatment strategies for gastro-esophageal reflux disease. *Medical Decision Making* **22**: 290–308.

Chilcott JB, Whitby SM, Moore R. 2001. Clinical impact and health economic consequences of posttransplant type 2 diabetes mellitus. *Transplantation Proceedings* **33**(Suppl. 5A): 32S–39S.

Claxton K. 1999. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics* **18**: 341–364.

Claxton K, Sculpher M, McCabe C *et al*. 2005. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. *Health Economics* **14**: 339–347.

DOI: 10.1002/hec

Cronin KA, Legler JM, Etzioni RD. 1998. Assessing uncertainty in microsimulation modelling with application to cancer screening interventions. *Statistics in Medicine* **17**: 2509–2523.

Davies R, Roderick P, Raftery J. 2003. The evaluation of disease prevention and treatment using simulation models. *European Journal of Operational Research* **150**: 53–66.

Davies R, Roderick P, Raftery J, Crabbe D, Patel P, Goddard JR. 2002. A simulation to evaluate screening for helicobacter pylori infection in the prevention of peptic ulcers and gastric cancers. *Health Care Management Science* **5**: 249–258.

Doubilet P, Begg CB, Weinstein MC, Braun P, McNeil BJ. 1986. Probabilistic sensitivity analysis using Monte Carlo simulation. *Medical Decision Making* **6**: 85–92.

National Institute for Clinical Excellence. 2004. *Guide to the Methods of Technology Appraisal*. NICE: London. http://www.nice.org.uk/pdf/TAP_Methods.pdf (accessed July 2005).

Oakley JE, O'Hagan A. 2004. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society B* **66**: 751–769.

O'Hagan A, Kennedy MC, Oakley JE. 1999. Uncertainty analysis and other inference tools for complex computer codes (with Discussion). In *Bayesian Statistics*, vol. 6, Bernardo JM *et al*. (eds). Oxford University Press: Oxford, 503–524.

O'Hagan A, McCabe C, Akehurst RL *et al*. 2005. Incorporation of uncertainty in health economic modelling studies. *PharmacoEconomics* **23**: 529–536.

O'Hagan A, Stevens JW, Montmartin J. 2000. Inference for the cost-effectiveness acceptability curve and cost-effectiveness ratio. *PharmacoEconomics* **17**: 339–349.

Paltiel AD, Scharfstein JA, Seage 3rd GR *et al*. 1998. A Monte Carlo simulation of advanced HIV disease: application to prevention of CMV infection. *Medical Decision Making* **18**(Suppl.): S93–S105.

Ratcliffe J, Young T, Buxton M, Eldabi T, Paul R, Burroughs A. 2001. Simulation modelling approach to evaluating alternative policies for the management of the waiting list for liver transplantation. *Health Care Management Science* **4**: 117–121.

Stevenson MD, Brazier JE, Calvert NW, Lloyd-Jones M, Oakley J, Kanis JA. 2005a. Description of an individual patient methodology for calculating the cost-effectiveness of treatments for osteoporosis in women. *Journal of Operational Research Society* **56**: 214–221.

Stevenson MD, Lloyd Jones M, de Nigris E, Brewer N, Oakley JE. 2005b. A systematic review and economic evaluation of alendronate, etidronate, risedronate, raloxifene and teriparatide for the prevention and treatment of postmenopausal osteoporosis. *Health Technology Assessment* **9**(22): 1–160.

Stevenson MD, Oakley J, Chilcott JB. 2004. Gaussian process modelling in conjunction with individual patient simulation modelling: a case study describing the calculation of cost-effectiveness ratios for the treatment of osteoporosis. *Medical Decision Making* **24**: 89–100.

Stinnett AA, Mullahy J. 1998. Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Medical Decision Making* **18**(Suppl.): S68–S80.

Szeto KL, Devlin NJ. 1996. The cost-effectiveness of mammography screening: evidence from a microsimulation model for New Zealand. *Health Policy* **38**: 101–115.

DOI: 10.1002/hec