Monte Carlo probabilistic sensitivity analysis for patient level simulation models

Anthony O'Hagan, Matt Stevenson and Jason Madan
University of Sheffield

August 8, 2005

Abstract

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 and Bayesian statistics. Methods are presented to estimate the mean and variance of the model output, the cost-effectiveness acceptability curve and value of information calculations. The methods are simple to apply and will typically reduce the computational demand by a factor of at least 20. Three examples are presented.

Keywords: Analysis of variance; Bayesian statistics; cost-effectiveness; cost effectiveness acceptability curve; economic evaluation; economic model; individual-level simulation; micro-simulation; Monte Carlo; patient-level model; osteoporosis; probabilistic sensitivity analysis; rheumatoid arthritis; value of information.

1 Introduction

1.1 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 [1][2]. 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) [3][4][5]. 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 outputs characterises the output uncertainty, and to obtain accurate PSA we typically need 1,000 or more model runs.

Although most economic modelling has used cohort models, in which the output is the appropriate measure of cost-effectiveness for the entire treated population, there is increasing use of patient-level simulation models (also known as micro-simulation or individual-level simulation models) [6][7][8][9][10][11][12], in which treatment and response pathways for individual patients are simulated, 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. 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 in the order of 25 times less computation.

The remainder of this section defines some basic notation and considers the particular example where model output is incremental net benefit, while Section 2 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. Section 3 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 theory is extended to estimating the cost-effectiveness acceptability curve in Section 4, and to value of information analyses in Section 5. Finally, Section 6 discusses incremental cost-effectiveness ratios, alternatives to Monte Carlo and directions for further research. Some technical details are given in the Appendix.

1.2 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 Section 6.4.

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 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. The noise has zero expectation, because the definition of the 'true' model output is the population mean (i.e. averaged over a large population of patients).

In a Monte Carlo PSA, let $\mathbf{x}_i$ denote the i-th sampled parameter set, and let z_{ij} denote the output value for the j-th 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 $\bar{z}_i = \frac{1}{n}\sum_{j=1}^{n} z_{ij}$, and the mean over all Nn patients in all model runs by $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} \bar{z}_i$.

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 Section 6.5.

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 in the remainder of this section and the next concentrates on methods to estimate μ and σ^2.

Another important quantity in all of these methods is the variability between patients in a given run. 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 define

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

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

1.3 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 [13], and treatment 2 is more cost-effective than treatment 1 if $y > 0$. One role of PSA is then to quantify the uncertainty in whether treatment 2 is more cost-effective. 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 it should be to use treatment 2 if $\mu > 0$ [14]. 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.

Our analysis in Section 4 deals explicitly with this case, and with estimating the cost-effectiveness acceptability curve [15] that plots the probability that $y(\mathbf{X})$ is positive as a function of λ. However, net benefit also provides a helpful illustration for the more general theory in Sections 2 and 3.

2 Standard Monte Carlo PSA

2.1 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 $\bar{y} = \frac{1}{N}\sum_{i=1}^N y_i$ and $s^2 = \frac{1}{N-1}\sum_{i=1}^N (y_i - \bar{y})^2$. 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

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

The subscript S here indicates that these are the standard Monte Carlo estimates. The mean and variance of $\hat{\mu}_S$ follows from simple algebra, using the facts that $E(\bar{z}_i) = \mu$ and $\mathrm{var}(\bar{z}_i) = \sigma^2 + \bar{\tau}^2/n$. We find

$$E(\hat{\mu}_S) \ = \ \mu \ , \tag{3}$$

$$\mathrm{var}(\hat{\mu}_S) \ = \ \frac{\sigma^2}{N} + \frac{\bar{\tau}^2}{Nn} \ . \tag{4}$$

Therefore $\hat{\mu}_S$ is an unbiased estimator of μ, and its variance decreases with N in the usual way. Assuming large n, the Central Limit Theorem in statistics ensures that the $\bar{z}_i$s are approximately normally distributed, and hence $\hat{\sigma}^2_S$ has a chi-squared distribution. Its mean and variance are

$$\begin{aligned} E(\hat{\sigma}^2_S) &= \sigma^2 + \bar{\tau}^2/n \ , \\ \mathrm{var}(\hat{\sigma}^2_S) &= \frac{2}{N-1}(\sigma^2 + \bar{\tau}^2/n)^2 \ . \end{aligned}$$

Therefore the standard Monte Carlo estimator $\hat{\sigma}^2_S$ is *biased*. Its bias is $\bar{\tau}^2/n$, which is always positive, so on average it over-estimates σ^2. The main reason for using a large n is to make this bias small.

2.2 Sample sizes for standard estimators

We now identify values of n and N that would be required to obtain any desired accuracy in $\hat{\mu}_S$ or $\hat{\sigma}^2_S$. 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. As usual, the sample sizes depend on the unknown values of the variances, in this case σ^2 and $\bar{\tau}^2$, and it is therefore necessary to obtain initial estimates or guesses in order to apply the formulae.

The primary focus of the cost-effectiveness analysis is μ, the best estimate of the cost-effectiveness output y in the light of input uncertainty. Suppose that we wish to estimate μ with standard deviation d, so that a 95% interval has width $\pm 1.96d$. Then we would need

$$N \geq \frac{\sigma^2 + \bar{\tau}^2/n}{d^2} = \frac{1 + k/n}{d^2}\ \sigma^2 \ . \qquad (5)$$

If n has been chosen large enough to make $\bar{\tau}^2/n$ very small compared with σ^2, then this is approximately σ^2/d^2, which is the sample size required in conventional cohort models.

In the context where the model output is incremental net benefit, as discussed in Section 1.3, 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. For instance, if we set $d = c_1\sigma$ then (5) becomes

$$N \geq (1 + k/n)/c_1^2 \ . \qquad (6)$$

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, which is

$$CV(\hat{\sigma}^2_S) = \frac{\sqrt{\mathrm{var}(\hat{\sigma}^2_S)}}{E(\hat{\sigma}^2_S)} = \sqrt{\frac{2}{N-1}} \ .$$

So suppose that we wish to achieve a coefficient of variation less than or equal to c_2. For instance, setting $c_2 = 0.05$ means that we require to estimate σ^2 with a standard deviation no more than 5% of σ^2 itself, and hence an approximate 95% confidence interval of ±10%. Then the required number of runs is

$$N \geq 1 + 2/c_2^2 \ . \tag{7}$$

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. Then by comparing (6) and (7) we can see the former is the more stringent requirement: the number of runs in standard Monte Carlo should normally be chosen to satisfy the requirement (6) for accurate estimation of the mean μ.

A natural objective in choosing n would be to make the bias in $\hat{\sigma}_S^2$ small compared with the width of a confidence interval for σ^2. We therefore suggest that in general n should be made large enough so that the bias is only 10% of $c_2\sigma^2$. Then, remembering that $k = \bar{\tau}^2/\sigma^2$, this implies

$$n \geq 10k/c_2 \ . \tag{8}$$

In the case where the output is incremental net benefit, we can combine this with the preceding suggestion that $c_2 = 2c_1$ and apply (6) to obtain $N \geq (1 + 0.2c_1)/c_1^2$, so that the total number of patients to be sampled, Nn, is at least $5k/c_1^3$.

2.3 Example 1: osteoporosis

To illustrate the sample size calculations, we consider a large model developed at Sheffield for assessing the cost-effectiveness of many treatments for osteoporosis [12]. 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. The relative risk of fracture by using alendronate was estimated (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 [16]. 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.

Initial estimates of variances were $\bar{\tau}^2 = 2.38 \times 10^9$ and $\sigma^2 = 205072$, and hence $k = 2.38 \times 10^9/205072 = 11606$. The derivation of these initial estimates is described in Section 3.6, since they rely on the ANOVA methods developed in Section 3. On the basis of these estimates, equation (8) suggests using $n = 116060/c_2$, and even setting $c_2 = 0.2$ implies more than half a million patients per run. On this basis, each run of the model would have required approximately forty-two hours of computing time on a fast PC, making any serious PSA infeasible. In fact, from (6) and using the corresponding $c_1 = 0.1$ we would require $N = 1.02/0.1^2 = 102$ runs, and a total computing time of almost six months. It is to address the infeasibility of PSA for many patient-level models using the standard Monte Carlo method that the theory in the following section is developed.

3 One way ANOVA

3.1 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 (3) and (4) 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 $\mathrm{var}(\hat{\mu}_S)$, which from (4) 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 to make $n = 1$, i.e. to sample just one patient per parameter set. We then get $\mathrm{var}(\hat{\mu}_S) = \sigma^2/M + \bar{\tau}^2/M$.

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.

3.2 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

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

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

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

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

$$\hat{\sigma}^2_A = \frac{1}{n}\left(\frac{S_b}{N-1} - \frac{S_w}{N(n-1)}\right) \ , \tag{9}$$

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.

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

$$\mathrm{var}(\hat{\sigma}_A^2) = 2\left(\frac{(\sigma^2 + \bar{\tau}^2/n)^2}{N-1} + \frac{\bar{\tau}^4}{Nn^2(n-1)}\right) . \tag{10}$$

3.3 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 obtain suitably small variances (4) and (10) for the estimators of μ and σ^2. However, having the freedom now to choose both n and N 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 (10) for fixed M. Then we find the minimal M to achieve the required accuracy for estimating σ^2. These two steps give optimal values of N and n, and we find that they also give the desired accuracy for estimating μ. Full details of these derivations are given in the Appendix, and we report here the key results.

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

$$n = \frac{M(1+k)+k}{M+2k} . \tag{11}$$

Suppose again that we wish to achieve a coefficient of variation for estimating σ^2 less than or equal to c_2, so that we require $\mathrm{var}(\hat{\sigma}_A^2) \leq c_2^2\sigma^4$. Then the required total sampling effort is

$$M = \frac{1}{2c_2^2}\left(c_2^2 + 2 + 8k + \sqrt{c_2^4 + 4c_2^2 + 16c_2^2k + 4 + 32k + 64k^2 + 32c_2^2k^2}\right) \tag{12}$$

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.

$$M = 8k/c_2^2\ , \tag{13}$$

$$n = 1 + k\ . \tag{14}$$

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).

3.4 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 Sections 3.6 and 3.7 below.

1. Given a desired sampling precision c_2 for estimating σ^2, choose M using equation (12) or its simple form (13).

2. Now choose n using (11) or its simple form (14), 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 (9).

5. The variances of these estimators are given by (4) and (10), 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 $c_2 = \sqrt{8k/M}$.

3.5 Efficiency gain over standard Monte Carlo

We found in Section 2.2 that the appropriate values for N and n using the standard Monte Carlo approach would yield a total sampling load of $M = Nn = 5k/c_1^3$, at least in the case where the model output is incremental net benefit. 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 $2.5/c_1$, and since we will usually require c_1 to be 0.1 or less this implies an efficiency gain of 25 times or more.

We suggest that the fact that the ANOVA method requires of the order of 25 times 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.

3.6 Example 1: osteoporosis (continued)

Continuing the analysis of the osteoporosis model in Section 2.3, we will now apply the theory for the ANOVA method. 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.

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 choice of 15,000 patients was based on the fact that the standard error of the mean is the square root of $2.4\times10^9/15000$, i.e. 400, which is small enough relative to the observed mean of 1308 to be confident that the true mean incremental net benefit $y(\mathbf{x}_0)$ is positive. It is then 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 $3 \times 3 \times 3$ factorial design with each fracture probability input set at three levels — its mean value and its mean value plus or minus one standard deviation. 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 2.38×10^9 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 219429. Subtracting the estimated bias of $2.38 \times 10^9/15000 = 158667$ (in effect, applying equation (9)) gives an initial estimate of 60762 for the underlying variance across these 27 runs. In order to convert this to an estimate of σ^2, note that the variance of the three values used for each input in the factorial design is actually two-thirds of the variance describing the uncertainty in that input. We therefore estimate σ^2 by

$60762 \times 1.5^3 = 205072$. The correction factor here is based on the model output being approximately linear in its inputs. This is a very crude estimate, being based only on 27 runs and on approximate linearity (and we were lucky it did not come out negative), but suggests a value for k of $2.38 \times 10^9/205072 = 11606$. 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 minutes 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 ±25%. Our main analysis is therefore based on $N = 500$ runs using fracture probability inputs randomly sampled from their uncertainty distributions, and $n = 10000$ patients per run. From these data, we found $\bar{z} = 879.2$, $S_w = 1.1935658 \times 10^{16}$ and $S_b = 230495731$. 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 $\hat{\sigma}^2_A = 223178$.

The resulting estimate of k is the ratio of these last two estimates, 10695, 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 (10), we obtain an estimated standard deviation for $\hat{\sigma}^2_A$ of 29244, so an approximate 95% interval for σ^2 is $223178 \pm 58488 = [164690, 281666]$. As expected, the interval has range approximately ±25%. The corresponding estimate and 95% interval for σ become 472.4 and $[405.8, 530.7]$. The interval has range approximately ±13%. These various estimates and intervals are the primary results of the PSA.

To confirm the efficiency of the ANOVA method, suppose that we had chosen to apply the standard Monte Carlo method with the same target coefficient of variation of $c_2 = 0.126$ for estimating σ^2. Then following the analysis in Section 2.3 we should have used $n = 10k/c_2 = 850,000$ patients per run. A sample of size $N = 125$ would now suffice to estimate σ^2 with coefficient of variation 0.126. However, to estimate μ with a comparable standard deviation of 0.063σ would have required $N = 256$ runs. Even if such huge numbers of patients could be handled in each run, the total number of patients simulated would have been over two hundred million and would have taken almost two years of solid computation. Our actual analysis used 500 runs of 10,000 patients each, or 5 million patients in all, which represents a forty-fold saving in effort (agreeing with the formula $2.5/c_1 = 2.5/0.063 = 39.7$).

3.7 Implementation

The key to implementing the method is to obtain an initial estimate of σ^2. The method used for the osteoporosis model can be adapted for use more generally. First note that the model input values for the initial set of runs were not chosen randomly. In order to obtain a useful estimate of σ^2 it is important to use model input sets that are well separated. In fact the choice of levels for the three factors (i.e. model inputs) in that experiment was probably not good, in that they were not sufficiently well spread out and it was necessary to scale up the resulting estimate of σ^2. Instead, we suggest setting the three levels of each input to its mean and the mean plus and minus 1.5 standard deviations. Then instead of multiplying the resulting σ^2 estimate by 1.5^d, where d is the number of inputs, the appropriate factor is $(\frac{2}{3})^d$.

The number of initial runs in the osteoporosis example was 27, and we suggest that about this size of preliminary sample should suffice to obtain a first estimate of σ^2, using quite large numbers of patients per run. With more than $d = 3$ parameters, it will not be possible then to use all the 3^d combinations of parameter values for a full factorial experiment. There is some theory of fractional factorial experiments in the statistics literature, which would certainly yield good designs. However, in practice it may be adequate simply to use a random selection of 20 to 30 combinations (sampling without replacement) from those 3^d.

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 $d = 4$ 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.

4 The probability of cost-effectiveness

As discussed in Section 1.3, a common objective of PSA is to estimate the probability that treatment 2 is more cost-effective than treatment 1. If the model output y is the incremental net benefit of treatment 2 with respect to treatment 1, then treatment 2 is more cost-effective if y is positive. Because of uncertainty about model inputs, there is uncertainty about cost-effectiveness, and it is therefore of interest to ask for the *probability* that y is positive, i.e. $P(y(\mathbf{X}) > 0)$.

4.1 Two approaches

The simplest way to estimate this probability is to use just the estimated mean $\hat{\mu}_S = \bar{z}$ and the estimated variance $\hat{\sigma}_A^2$ of the uncertainty distribution. If we assume that this distribution is approximately normal, then we can estimate $P = P(y(\mathbf{X}) > 0)$ by

$$\hat{P}_N = \Phi(\bar{z}/\hat{\sigma}_A) \ ,$$

where Φ denotes the standard normal distribution function. We will refer to $\hat{P}_N$ as the *normal-distribution estimate.* For instance, in the example of Section 3.6 we found $\bar{z} = 879.2$ and $\hat{\sigma}_A = \sqrt{223178} = 472.4$. So $\bar{z}$ is $879.2/472.4 = 1.86$ standard deviations above zero, and the probability that alendronate is cost-effective relative to no treatment is estimated to be $\hat{P}_N = \Phi(1.86) = 0.969$.

When doing PSA with a cohort model, the same approach can be used, in which the standard Monte Carlo estimators, the sample mean and variance of the observed y_is, take the place of $\bar{z}$ and $\hat{\sigma}_A^2$. However, in practice a nonparametric approach is used instead which does not assume that the input uncertainty leads to output uncertainty that has the normal distribution form. The actual sampled y_is may not look like a sample from a normal distribution, for instance having skewness or long tails, and it is hard then to justify a method that assumes normality. Instead, it is usual simply to estimate $P(y(\mathbf{X}) > 0)$ by the proportion of sampled y_is that are positive. This nonparametric estimate avoids the normality assumption and is more responsive to the shape of the sample.

In a patient-level simulation model, if we can make sufficiently large runs to ignore the noise, the proportion of $\bar{z}_i$s that are positive could be used instead. We will refer to this as the standard Monte Carlo estimate, and denote it by $\hat{P}_S$. However, it is easy to see that when we do not have such large n this will be a biased method. Because of sampling variability in the $\bar{z}_i$s, they will yield a sample that is more spread out than the corresponding y_is would be. For instance, in the osteoporosis example of Section 3.6, $\hat{P}_S = 445/500 = 0.89$, which underestimates the true probability of cost-effectiveness. We need to develop a method that takes account of this extra variability.

4.2 A Bayesian estimate

We propose a hybrid method as an alternative to the normal-distribution estimate $\hat{P}_N$, based on estimating the true y_is by a standard Bayesian argument assuming normally distributed values, but then using the nonparametric approach to estimate $P(y(\mathbf{X}) > 0)$. For the first step, we suppose that $\bar{z}_i$ is normally distributed around its mean value of y_i, with variance $\bar{\tau}^2/n$. Because of the Central Limit Theorem (CLT), this will almost always be a reasonable assumption in practice. We also assume that y_i is normally distributed about its mean of μ and with variance σ^2. We cannot appeal to the CLT to justify normality in this case, and it is assumed at this stage essentially for convenience.

To estimate y_i, we can use a Bayesian argument in which the observation is $\bar{z}_i$ and the unknown parameter is y_i. The distribution $N(y_i, \bar{\tau}^2/n)$ for the observation provides the likelihood function, and the prior distribution for y_i is

$N(\mu, \sigma^2)$. Now if μ, σ^2 and $\bar{\tau}^2$ are known, the Bayesian posterior distribution of y_i is normal with mean

$$\hat{y}_i = \frac{n\bar{z}_i/\bar{\tau}^2 + \mu/\sigma^2}{n/\bar{\tau}^2 + 1/\sigma^2} = w\bar{z}_i + (1-w)\mu \ , \tag{15}$$

where

$$w = \frac{n/\bar{\tau}^2}{n/\bar{\tau}^2 + 1/\sigma^2} = \frac{n}{n+k} \ , \tag{16}$$

and variance

$$v = w\bar{\tau}^2/n \ .$$

To use this result, we substitute estimates derived in Section 3. Thus, we use equation (9) for σ^2, $S_w/\{N(n-1)\}$ for $\bar{\tau}^2$ and $\bar{z}$ for μ. Note that this ignores uncertainty in these parameter estimates, and so is not a fully Bayesian solution. In effect, we suppose that the sampling is adequate to estimate these parameters accurately. Whereas this will in practice be true for μ and $\bar{\tau}^2$ it may not hold for σ^2. However, a fully Bayesian analysis would be much more complex, and we prefer the simpler approximation because it is readily understood and implemented.

To estimate $P(y(\mathbf{X}) > 0)$, we do not simply use the proportion of $\hat{y}_i$s that are positive, since $\hat{y}_i$ is only an estimate of y_i. We need to take account also of the variance v. From the Bayesian posterior distribution, the probability that y_i is positive is $\Phi(\hat{y}_i/\sqrt{v})$. Hence we obtain the estimate

$$\hat{P}_H = \frac{1}{N}\sum_{i=1}^{N} \Phi(\hat{y}_i/\sqrt{v}) \ , \tag{17}$$

which we will refer to as the *hybrid estimate*. Of course, this solution is expressed in terms of the unknown parameters μ, σ^2 and $\bar{\tau}^2$, and in practice we need to replace these by estimates. If we substitute the ANOVA estimate (9) for σ^2 and $S_w/\{N(n-1)\}$ for $\bar{\tau}^2$ we find that

$$w = n(N-1)\hat{\sigma}_A^2/S_b = 1 - 1/F \ , \tag{18}$$

where $F = \frac{S_b}{N-1}/\frac{S_w}{N(n-1)}$ is the usual F-statistic in one-way analysis of variance, and

$$v = \hat{\sigma}_A^2/F \ . \tag{19}$$

From (18), and using the estimate $\hat{\mu}_S = \bar{z}$ for μ, we can rewrite (15) as

$$\hat{y}_i = \bar{z}_i - (\bar{z}_i - \bar{z})/F \ . \tag{20}$$

It is then simple to apply (17) using (20) and (19).

Applying the hybrid estimator to the osteoporosis example yields $\hat{P}_H = 0.965$, which is very close to the normal-distribution estimate of 0.969.

4.3 Example 2: simulated data

In order to test the accuracy of $\hat{P}_N$ and $\hat{P}_H$ when the underlying distribution is non-normal, we conducted a simulation exercise. The true distribution of $y(\mathbf{X})$ is shown in Figure 1, which shows it to be far more peaked in the centre and far more long-tailed than the normal distribution. It also exhibits a moderate degree of skewness. The construction of this distribution is described in the Appendix.

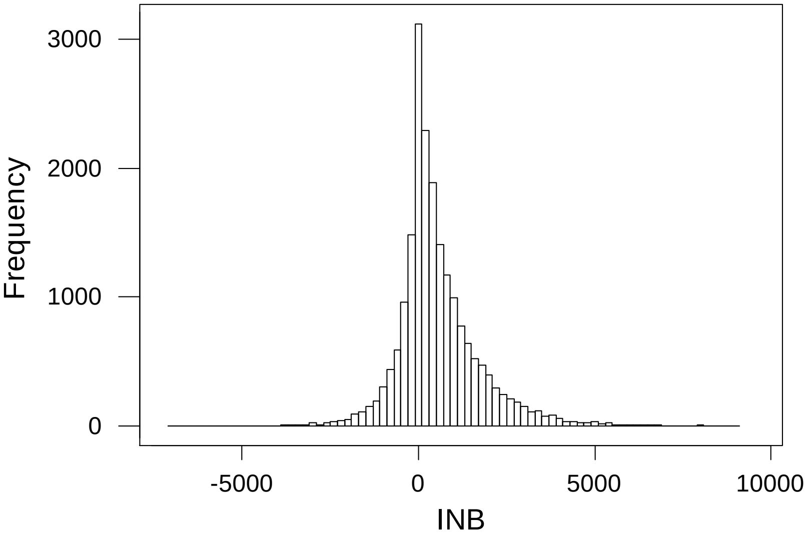

Figure 1. True distribution of incremental net benefit, Example 2.

The true mean μ is 550 and the true variance is $\sigma^2 = 1161^2 = 1.35 \times 10^6$. The true probability that net benefit is positive is $P(y(\mathbf{X}) > 0) = 0.705$. The simulation assigned a patient-level variance of $\bar{\tau}^2 = 4 \times 10^8$, so that the optimal value of n would be $1 + \bar{\tau}^2/\sigma^2 = 297$. We suppose that it is decided to perform PSA with $N = 500$ runs of $n = 300$ patients per run. For each run a true output y_i was sampled from the distribution shown in Figure 1. A sample mean $\bar{z}_i$ was generated by adding a normally distributed error to y_i with zero mean and variance $\bar{\tau}^2/n = 1333333$. The simulation was repeated 10,000 times. In each simulation, $\hat{\sigma}^2_A$, $\hat{P}_N$ and $\hat{P}_H$ were computed, as well as the standard Monte Carlo estimate $\hat{P}_S$ based on the proportion of positive sample means $\bar{z}_i$. The results are shown in Table 1.

	Mean	(std. dev.)
$\hat{\sigma}_A$	1157	(91)
$\hat{P}_N$	0.683	(0.024)
$\hat{P}_H$	0.681	(0.024)
$\hat{P}_S$	0.624	(0.022)

Table 1. Mean and standard deviation for PSA estimates based on 10,000 simulations.

It can be seen that the true PSA standard deviation σ is estimated quite accurately. It is worth noting that although $\hat{\sigma}_A^2$ is an unbiased estimator of σ^2, $\hat{\sigma}_A$ will strictly be a biased estimator of σ. Table 1 confirms that this bias is small (although it would be larger if a smaller PSA had been conducted; $N = 500$ and $n = 300$ is adequate to estimate σ^2 reasonably accurately). Note, however, that both $\hat{P}_N$ and $\hat{P}_H$ underestimate the true value of $P = 0.705$ slightly. The discrepancy is because of the non-normality of the underlying output distribution. The two estimates are very similar and both are much better that $\hat{P}_S$, which shows the anticipated bias due to not having a large enough n.

The similarity of the two estimates $\hat{P}_N$ and $\hat{P}_H$ reflects the fact that the hybrid method does not recover much of the rather marked non-normality of the underlying distribution, as shown in Figure 1. This is not really a failure of the method, but is a consequence of using a relatively small n. This leads to the sample means $\bar{z}_i$ having a large random variability around the true y_is, and this error is effectively normally distributed. Hence, the sampled $\bar{z}_i$s do not retain the underlying non-normal shape of the y_is, and the gain from using the individual $\hat{y}_i$s in the hybrid method is much smaller than that of using the nonparametric estimator $\hat{P}_S$ in the large n case.

4.4 CEAC

The above analysis assumes that the model output y is incremental net benefit, which requires that the willingness to pay coefficient λ is known. In practice, it is usual to consider a range of values of λ by computing the cost-effectiveness acceptability curve (CEAC), which plots the probability $P(\lambda)$ that incremental net benefit is positive against λ. The above analysis can be applied separately for each λ in order to plot the CEAC; however, it is possible to derive estimates of the CEAC directly, using both the normal-distribution and hybrid methods, by generalising the above analysis to two outputs.

Let y be a vector comprising the two outputs y_e and y_c, representing respectively incremental efficacy and incremental cost. Now we identify $\mu = E(y(\mathbf{X}))$ as also a vector comprising μ_e and μ_c, while $\sigma^2 = \mathrm{var}(y(\mathbf{X}))$ is a 2×2 matrix. Similarly, the between patient variance $\bar{\tau}^2$ is a 2×2 matrix. The data now give rise to the mean vector $\bar{z}_i$ at the i-th input configuration $\mathbf{x}_i$ and the overall mean vector $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} \bar{z}_i$, as before. The sums of squares S_b and S_w are now also 2×2 matrices of sums of squares and cross-products, defined by

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

The same algebra applies as in Section 3.2, and we still find that $\bar{z}$ is an unbiased estimator of μ, $S_w/\{N(n-1)\}$ is an unbiased estimator of $\bar{\tau}^2$, while (9) gives the unbiased estimator of σ^2.

The normal-distribution method is now readily applied by computing the estimates $\hat{\mu}_\lambda$ and $\hat{\sigma}^2_\lambda$ for the incremental net benefit at a given λ. These are $\hat{\mu}_\lambda = L_\lambda\bar{z}$ and $\hat{\sigma}^2_\lambda = L_\lambda\hat{\sigma}^2 L_\lambda^T$, where L_λ is the vector $(\lambda, -1)$. So the CEAC is estimated by $\hat{P}_N(\lambda) = \Phi(\hat{\mu}_\lambda/\hat{\sigma}_\lambda)$.

To derive the hybrid estimate, essentially the same Bayesian theory applies for estimating y_i, although now this is again in the form of matrix algebra. The Bayesian posterior distribution of y_i is (bivariate) normal with mean

$$\hat{y}_i = w\bar{z}_i + (1-w)\mu$$

and variance

$$v = \frac{1}{n}w\bar{\tau}^2 \ .$$

Note, however, that w is now a 2×2 matrix

$$w = (n\bar{\tau}^{-2} + \sigma^{-2})^{-1} n\bar{\tau}^{-2} \ ,$$

where $\bar{\tau}^{-2}$ and σ^{-2} denote the matrix inverses of $\bar{\tau}^2$ and σ^2.

It follows that the posterior distribution of the incremental net benefit $L_\lambda y_i$ for given λ is normal with mean $L_\lambda\hat{y}_i$ and variance $L_\lambda v L_\lambda^T$. We can then apply the method of Section 4.2 to get the hybrid estimate of the CEAC as

$$\hat{P}_H(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\Phi\left(\frac{L_\lambda\hat{y}_i}{\sqrt{L_\lambda v L_\lambda^T}}\right) \ . \tag{21}$$

4.5 Example 3: rheumatoid arthritis

This example concerns an application of the Sheffield model for TNF inhibitor treatments in rheumatoid arthritis [10][17][18]. There is no cure for this disease. TNF inhibitors are a recent addition to the armoury of drugs used to ameliorate symptoms in the short-term, and slow the longer-term progression of rheumatoid arthritis. The model examines the cost-effectiveness of using these drugs rather than the next best treatment (DMARDs). TNF inhibitors are currently only indicated for patients with severe rheumatoid arthritis, who have failed to respond to front-line therapies. The Sheffield model is a patient-level simulation model for the impact of treatment on this patient group. Characteristics for each individual patient are simulated by sampling from a national registry of rheumatoid arthritis sufferers. The long-term quality of life of each patient is simulated through models of initial improvement from treatment, longer term disease progression during treatment, duration of drug effectiveness, patient lifetime and days spent in hospital as functions of the patient's simulated characteristics and the specified treatment. The uncertain model inputs are the coefficients in each function. Simulated costs comprise drug expenditure and associated monitoring costs, as well as general treatment costs for the disease.

For this example, costs and benefits were both discounted at 3.5% per annum, and uncertainty about model input coefficients was expressed through multivariate normal joint distributions.

Initial exploration of the between patient variability in this model suggested that it was much smaller than was experienced in the osteoporosis model. An initial run using $N = 100$ randomly chosen input parameter combinations with $n = 100$ patients per run obtained the following results. For the incremental QALY output, the estimates of $\bar{\tau}^2$ and σ^2 were respectively 0.8888 and 0.03232, suggesting an optimal $n = 1 + 0.8888/0.03232 = 29$. For the incremental cost output, the optimal value was estimated as $n = 1+(6.112\times 10^8)/(1.248\times 10^7) = 49$. On this basis, is was decided to make the main PSA run with $n = 50$ patients per run. A sample size of $N = 1000$ was chosen.

The theory for two outputs was implemented fully for the main PSA. The 1000 runs yielded the following estimates.

$$\bar{z} = \left(\begin{array}{c} 1.2639 \\ 42594 \end{array} \right) , \qquad \bar{\tau}^2 = \left(\begin{array}{cc} 0.84532 & 18741 \\ 18741 & 6.0766 \times 10^8 \end{array} \right) ,$$

$$\hat{\sigma}^2 = \left(\begin{array}{cc} 0.046619 & 332.19 \\ 332.19 & 1.1937 \times 10^7 \end{array} \right) .$$

On the basis of these estimates, the TNF inhibitor is estimated to produce 1.26 more QALYs, with a standard deviation of $\sqrt{0.046619} = 0.216$, so we are very sure that it is more effective than the DMARD. It has an estimated incremental cost of 42600 with a standard deviation of $\sqrt{1.1937 \times 10^7} = 3455$, so again we are very sure that is is more expensive. The question is whether it is cost-effective, at a range of willingness to pay values λ. The estimated incremental cost-effectiveness ratio is $42594/1.2639 = 33700$ UK pounds per QALY, so there is some doubt over its cost-effectiveness for the National Health Service. The estimated CEAC using both methods is plotted in Figure 2 over the range $\lambda \in [20000, 50000]$.

The two curves are very close, but the normal-distribution curve $\hat{P}_N(\lambda)$ is slightly flatter, implying less information about cost-effectiveness. The estimated probability that incremental net benefit is positive is 0.52 at $\lambda = 33700$ (the estimated ICER), but falls to 0.21 at $\lambda = 30000$ and to effectively zero at $\lambda = 20000$.

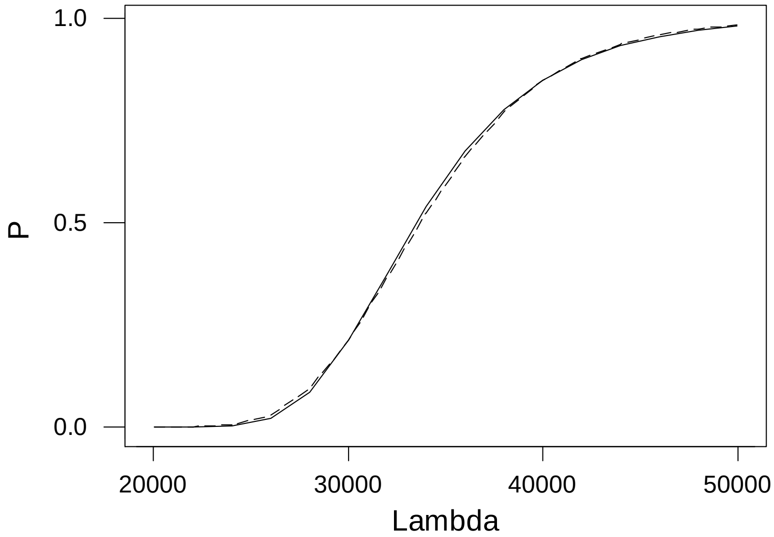

Figure 2. Estimated CEAC for rheumatoid arthritis example using normal-distribution (dashed line) and hybrid (solid line) methods.

4.6 Implementation

The rheumatoid arthritis example raises another question about implementing this approach. The CEACs shown in Figure 2 are not only very similar but are in fact almost identical to the one that would have been obtained by simply using the proportion of positive sample mean net benefit values $L_\lambda \bar{z}_i$ for each λ. The reason is that although $n = 49$ was estimated as optimal for the incremental cost output it is unnecessarily large for PSA of incremental net benefits in the range of λ of interest. For λ around 33700, the corresponding values of $\bar{\tau}^2$ and σ^2 are estimated to be

$$\begin{pmatrix} 33700 & -1 \end{pmatrix} \begin{pmatrix} 0.84532 & 18741 \\ 18741 & 6.0766 \times 10^8 \end{pmatrix} \begin{pmatrix} 33700 \\ -1 \end{pmatrix} = 3.045 \times 10^8$$

and

$$\begin{pmatrix} 33700 & -1 \end{pmatrix} \begin{pmatrix} 0.046619 & 332.19 \\ 332.19 & 1.1937 \times 10^7 \end{pmatrix} \begin{pmatrix} 33700 \\ -1 \end{pmatrix} = 4.249 \times 10^7 ,$$

so that the optimal n is only about 8. With 50 patients per run, the sample mean net benefit values $L_\lambda \bar{z}_i$ are relatively accurate. As a result, the corresponding Bayesian estimates $L_\lambda \hat{y}_i$ are close to those means and their variances $L_\lambda v L_\lambda^T$

are small. Had we used $n = 8$ instead of $n = 50$, the simple CEAC estimate based on the proportion of positive sample mean net benefits would have been appreciably biased, making the methods of Section 4.4 necessary.

If the primary objective is, as in most PSA analyses of cost-effectiveness, to examine incremental net benefit over a range of values of λ, then for maximum efficiency n should be chosen on the basis of initial estimates according to the above analysis, rather than by looking separately at incremental QALYs and incremental costs as was done in the example. In effect, this was how n was derived for the osteoporosis model, and the runs obtained in that example could have been used to estimate a CEAC for a range of λ values around the assumed £30,000/QALY.

5 Value of information

5.1 EVPI

In the context of decision making about which of a number of treatments to adopt, an important measure of overall decision uncertainty is the expected value of perfect information (EVPI). This is defined as the expected increase in expected net benefit that could be obtained if we were able to learn the true values of all the uncertain model inputs $\mathbf{x}$. For simplicity of exposition, we assume that there are just two treatments to be compared, and the model output is incremental net benefit of treatment 2 relative to treatment 1.

EVPI is calculated in two stages, first finding the expected incremental net benefit if no extra information is available, and then finding the expectation if we were to learn the true value of $\mathbf{x}$. First, if no extra information is available we should prefer treatment 2 to treatment 1 if and only if $\mu > 0$. The resulting expected incremental net benefit is

$$U = \max\{\mu, 0\} \ .$$

Second, if we can learn the true value of $\mathbf{x}$, then we will choose treatment 2 if and only if $y(\mathbf{x}) > 0$, obtaining expected incremental net benefit of $\max\{y(\mathbf{x}), 0\}$. However, prior to actually obtaining this information we do not know the value of $\mathbf{x}$, and the appropriate comparison with U is the expectation

$$U^* = E\left[\max\{y(\mathbf{X}), 0\}\right] \ .$$

Notice that we formally recognise that $\mathbf{x}$ is uncertain here by using the symbol $\mathbf{X}$. The expectation in U_1 is with respect to the uncertainty in $\mathbf{X}$. Finally, the EVPI is the difference

$$\text{EVPI} = U^* - U \ ,$$

and it can be shown that this is necessarily non-negative. The larger the EVPI, the more appreciable is the uncertainty in the choice of treatment.

It is usual to compute EVPI in cohort models by Monte Carlo sampling. Given a suitably large number N of runs, U is estimated as $\max\{\bar{y}, 0\}$, and

U^* by $\frac{1}{N}\sum_{i=1}^N \max\{y_i, 0\}$. For patient-level models, the standard Monte Carlo estimates are then given by

$$\hat{U}_S = \max\{\bar{z}, 0\} \ , \qquad \hat{U}_S^* = \frac{1}{N}\sum_{i=1}^N \max\{\bar{z}_i, 0\} \ ,$$

evaluated using a large number n of patients in each run. Now it is important to recognise that because of the maximisation steps in these calculations both estimators are biased (and indeed the usual estimate of U is biased in the case of cohort models) [19]. Essentially, the bias arises from the fact that because of the possibility of estimation errors in $\bar{z}$ and $\bar{z}_i$ we are not certain whether the corresponding true values μ and $y(\mathbf{x}_i)$ are positive, and the operation of taking the maximum will tend to overestimate the true values of U and U^*. Because there is more uncertainty in each $\bar{z}_i$ than in $\bar{z}$, the bias is larger in U^*, and so the estimate of EVPI will be biased upwards. The answer to minimising these biases is again to use large samples. Both N and n need to be large enough to be almost certain whether μ or $y(\mathbf{x}_i)$ is positive.

Note that if the interest is simply in estimating U then the result of Section 3.1 applies and it is most efficient to use $n = 1$. However, the standard Monte Carlo method uses the same sampled inputs $\mathbf{x}_i$ for both U and U^*, and to estimate U^* accurately it is necessary to estimate each $\mu(\mathbf{x}_i)$ accurately, and hence we must use large n.

5.2 Partial EVPI

A measure of the decision uncertainty induced by uncertainty in a subset of the model inputs is the so-called partial EVPI for those inputs. Let $\mathbf{x}^{\boldsymbol{I}}$ denote the subset of inputs of interest, and let $\mathbf{x}^{-I}$ denote the remaining inputs, so that $\mathbf{x}^{\boldsymbol{I}}$ and $\mathbf{x}^{-I}$ together partition $\mathbf{x}$. If we were able to learn the true value of $\mathbf{x}^{\boldsymbol{I}}$ before making a decision about which treatment to use, then the decision would give utility $\max\{\mu(\mathbf{x}^{\boldsymbol{I}}), 0\}$, where $\mu(\mathbf{x}^{\boldsymbol{I}})$ is the expected incremental net benefit with respect to uncertainty in the remaining inputs $\mathbf{x}^{-I}$, conditional on the revealed value of $\mathbf{x}^{\boldsymbol{I}}$. Since this value is not known at the present time, it is a random variable $\mathbf{X}^{\boldsymbol{I}}$, and we need to evaluate the expectation with respect to that uncertainty:

$$U^I = E\left[\max\{\mu(\mathbf{X}^{\boldsymbol{I}}), 0\}\right] \ . \tag{22}$$

Then the partial EVPI for $\mathbf{x}^{\boldsymbol{I}}$ is $U^I - U$.

As has been pointed out by Brennan *et al* [19], to evaluate this by Monte Carlo, even in the case of a cohort model, requires a two-level simulation. In an outer loop, we simulate many values of $\mathbf{x}^{\boldsymbol{I}}$, then in an inner loop we simulate many values of $\mathbf{x}^{-I}$ for *each* simulated value of $\mathbf{x}^{\boldsymbol{I}}$. The inner loop computes $\mu(\mathbf{x}^{\boldsymbol{I}})$ while the outer loop evaluates the expectation in (22). For a patient-level simulation model, it now becomes optimal to use $n = 1$, because the inner computation to evaluate $\mu(\mathbf{x}^{\boldsymbol{I}})$ is analogous to the estimation of μ, except that we fix $\mathbf{x}^{\boldsymbol{I}}$ and only simulate $\mathbf{x}^{-I}$. The argument of Section 3.1 applies and we should use just one patient per run. In practice, this means that estimating a partial EVPI for a patient-level model entails relatively little extra computation beyond that needed to compute EVPI.

6 Discussion

6.1 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. In addition to simple formulae for estimating the PSA mean μ and variance σ^2, we have extended the theory to provide methods for estimating the cost-effectiveness acceptability curve and measures of expected value of information.

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

6.2 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 [20]. 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 ICER is not an average of patient-level ratios, and therefore all of the above theory is inapplicable.

6.3 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 [21]. 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, σ^2 and the CEAC can be calculated using the methods of Oakley and O'Hagan [22] and Stevenson *et al* [23], while the theory is extended to EVPI by Oakley [24]. A similar approach has been proposed based on the more restrictive idea of fitting a response surface to the model output instead of a Gaussian process [25].

6.4 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 [26]. If patient outcomes are not independent, then a more appropriate modelling approach is by discrete event simulation [27][28][29]. 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.

6.5 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.

6.6 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, and this is anther topic for future research. For instance, 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 ideas in Section 4.4 concerned with two outputs.

Acknowledgement

This research was stimulated by a Focus Fortnight on patient-level simulation models organised by the Centre for Bayesian Statistics in Health Economics 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 on the presentation of this work.

References

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

[2] Claxton K, Sculpher M, McCabe C, Briggs A, Buxton M, Brazier J, Akehurst, O'Hagan A. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. *Health Economics* 2005; **14**, 339–347.

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

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

[5] O'Hagan A, McCabe C, Akehurst RL, Brennan A, Briggs A, Claxton K, Fenwick E, Fryback D, Sculpher M, Spiegelhalter DJ, Willan A. Incorporation of uncertainty in health economic modelling studies. *PharmacoEconomics* 2005; **23**, 529–536.

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

[7] Paltiel AD, Scharfstein JA, Seage GR 3rd, Losina E, Goldie SJ, Weinstein MC, Craven DE, Freedberg KA. A Monte Carlo simulation of advanced HIV disease: application to prevention of CMV infection. *Med Decis Making* 1998; **18**(Suppl):S93-105.

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

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

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

[11] Barton P, Jobanputra P, Wilson J, Bryan S, Burls A. 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* 2004; **8**: 1–104.

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

[13] Stinnett AA, Mullahy J. Net health benefits: A new framework for the analysis of uncertainty in cost-effectiveness analysis. *Med Decis Making* 1998; **18** suppl., S68–S80.

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

[15] Van Hout BA, Al MJ, Gordon GS, Rutten F. Costs, effects and C/E ratios alongside a clinical trial. *Health Econ* 1994; **3**, 309–319.

[16] Stevenson MD, Lloyd Jones M, de Nigris E, Brewer N, Oakley JE. A systematic review and economic evaluation of alendronate, etidronate, risedronate, raloxifene and teriparatide for the prevention and treatment of postmenopausal osteoporosis. *Health Technol Assess* 2005; **9** (22), 1 -160.

[17] Bansback NJ, Brennan A, Ghatnekar O. Cost effectiveness of adalimumab in the treatment of patients with moderate to severe rheumatoid arthritis in Sweden. *Ann Rheum Dis* 2005; **64**: 995–1002.

[18] Brennan A, Bansback NJ , Nixon R. Modelling the cost effectiveness of TNF-a inhibitors in the management of rheumatoid arthritis: Results from the British Society for Rheumatology Biologics Registry. Report to the British Society of Rheumatology, May 2005.

[19] Brennan A, Kharroubi S, O'Hagan A, Chilcott J. Calculating partial expected value of information in cost-effectiveness models. Submitted to *Medical Decision Making* 2005. http://www.shef.ac.uk/~st1ao/other/EVPI.doc.

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

[21] O'Hagan A, Kennedy MC, Oakley JE. Uncertainty analysis and other inference tools for complex computer codes (with discussion). In *Bayesian Statistics 6*, J. M. Bernardo *et al* (eds.). Oxford University Press, 1999; 503–524.

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

[23] 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. *Med Decis Making* **24**, 89–100.

[24] Oakley JE. Decision-theoretic sensitivity analysis for complex computer models. Statistics research report, University of Sheffield, 2005. http://www.sheffield.ac.uk/st1jeo/#_Papers

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

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

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

[28] Barton P, Bryan S, Robinson S. Modelling in the economic evaluation of health care: selecting the appropriate approach. *J Health Serv Res Policy* 2004; **9**, 110–118.

[29] Brennan A, Chick SE, Davies R. A taxonomy of model structures for economic evaluation of health technologies.

[30] O'Hagan A, Leonard T. Bayes estimation subject to uncertainty about parameter constraints. *Biometrika* 1976; **63**, 201–203.

[31] Azzalini A. The skew-normal distribution and related multivariate families (with discussion). *Scandinavian Journal of Statistics* 2005; **32**, 159–200.

Appendix: Further details

A.1 Optimal sample sizes

First, using the definitions of $k = \bar{\tau}^2/\sigma^2$ and $M = Nn$ we can rewrite (10) 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

$$2\sigma^4 \left(\frac{(M+k)^2}{M(M-n)^2} - \frac{k^2}{M(n-1)^2} \right) \ . \tag{23}$$

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, $n = \frac{M+k+Mk}{M+2k}$ and $n = -\frac{-M-k+Mk}{M}$. It is straightforward to confirm that the first of these, equation (11), is the required solution, by differentiating (23) again with respect to n and checking that it is negative at $n = \frac{M+k+Mk}{M+2k}$.

With this choice of n, (10) reduces to

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

We now require to choose M so that $\mathrm{var}(\hat{\sigma}^2_A) \leq c_2^2\sigma^4$. Setting $\mathrm{var}(\hat{\sigma}^2_A) = c_2^2\sigma^4$ 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 (12).

Now suppose that we also wish to have $\mathrm{var}(\hat{\mu}_S) \leq c_1^2\sigma^2$. 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

$$M = \frac{1}{2c_1^2} \left(1 + 2k - 2c_1^2 k + \sqrt{1 + 4k + 4k^2 + 4c_1^4 k^2} \right) \ . \tag{24}$$

If we suppose that k is large and c_1^2 small, then the square root is approximately $2k$, and (24) is approximately $2k/c_1^2$. In the case where the output is incremental net benefit, we have argued in Section 2.2 that we should set $c_1 = c_2/2$, in which case this equates to (13). Therefore, the same values (14) and (13) 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 Section 3.6 have approximately the same width.

A.2 Construction of the population in Example 2

The population shown in Figure 1 is a large sample from the distribution of a random variable Y constructed as follows. Let Z_1 and Z_2 be independent random variables with the skew-normal density

$$f(z) = \Phi(\sqrt{2}z)\phi(z/\sqrt{2}) \ ,$$

where ϕ and Φ denote the density and distribution functions of the standard normal distribution. The skew-normal distribution was first described by O'Hagan and Leonard [30] and more recently explored by Azzalini [31]. Then

$$Y = Z_1^2 \text{sign}(Z_1) + Z_2^2 \text{sign}(Z_2) \ .$$