REVIEW ARTICLE

Pharmacoeconomics 2011; 29 (1): 35-49
1170-7690/11/0001-0035/$49.95/0

Calibrating Models in Economic Evaluation

A Seven-Step Approach

Tazio Vanni,[1,2] *Jonathan Karnon*,[3] *Jason Madan*,[4] *Richard G. White*,[2] *W. John Edmunds*,[2] *Anna M. Foss*[2] and *Rosa Legood*[1]

1 Health Services Research Unit, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, London, UK
2 Centre for Mathematical Modelling of Infectious Disease, London School of Hygiene and Tropical Medicine, London, UK
3 School of Population Health and Clinical Practice, University of Adelaide, Adelaide, South Australia, Australia
4 Academic Unit of Primary Health Care, University of Bristol, Bristol, UK

Contents

Abstract

In economic evaluation, mathematical models have a central role as a way of integrating all the relevant information about a disease and health interventions, in order to estimate costs and consequences over an extended time horizon. Models are based on scientific knowledge of disease (which is likely to change over time), simplifying assumptions and input parameters with different levels of uncertainty; therefore, it is sensible to explore the consistency of model predictions with observational data. Calibration is a useful tool for estimating uncertain parameters, as well as more accurately defining model uncertainty (particularly with respect to the representation of correlations between parameters). Calibration involves the comparison of model outputs (e.g. disease prevalence rates) with empirical data, leading to the identification of model parameter values that achieve a good fit.

This article provides guidance on the theoretical underpinnings of different calibration methods. The calibration process is divided into seven steps and different potential methods at each step are discussed, focusing on the particular features of disease models in economic evaluation. The seven steps are (i) Which parameters should be varied in the calibration process? (ii) Which calibration targets should be used? (iii) What measure of goodness of fit should be used? (iv) What parameter search strategy should be used? (v) What determines acceptable goodness-of-fit parameter sets (convergence criteria)? (vi) What determines the termination of the calibration process (stopping rule)? (vii) How should the model calibration results and economic parameters be integrated?

The lack of standards in calibrating disease models in economic evaluation can undermine the credibility of calibration methods. In order to avoid the scepticism regarding calibration, we ought to unify the way we approach the problems and report the methods used, and continue to investigate different methods.

In economic evaluation, mathematical models have a central role as a way of combining relevant information about a disease and health interventions, in order to estimate costs and consequences over an extended time horizon.[1] Models incorporate assumptions that allow a simpler representation of a complex reality, and there is always uncertainty around the true values of model input parameters. As George Box[2] famously stated "essentially, all models are wrong, but some are useful." An important step towards proving the credibility and usefulness of a model is the process of calibration. This involves comparing model outputs with empirical data, known as calibration targets (e.g. disease prevalence rate), and exploring variations (within *a priori* plausible bounds) of the parameters of the model to identify combinations that provide a better fit to the data.[3,4]

A common use of calibration in economic evaluation is in situations where mean parameter values to populate the model are not observable, such as rates of clinical presentation in screening models.[5] More recent applications of calibration in economic evaluation have extended the technique to exploring uncertainties, and making adjustments where required, to a broader range of model inputs depending on the consistency between model outputs and observational data.[6,7] Since models are central to economic evaluation, the methods used for model calibration and the way calibration results are incorporated within the analysis have the potential to influence both the base-case cost-effectiveness results and the

variance in estimates of uncertainty. Despite the increasing use of model calibration within economic evaluation,[8] most health technology assessment guidelines and textbooks have little to say on how it should be used.[1,9-13] If calibration is to become more widely used and accepted, it is important that the methods used are coherent, well implemented and clearly reported.

This article first outlines the definitions and rationale for calibration in economic evaluation. Second, the steps for implementing calibration in economic evaluation and the methods for integrating economic evaluation are reviewed with reference to existing examples in the literature, using a selective review. A practical application of the seven stages of calibration is presented by Karnon and Vanni[14] in this issue of *PharmacoEconomics*, in the form of an example model that is available as Supplemental Digital Content 1, to download from http://links.adisonline.com/PCZ/A94.

1. Background

1.1 Definitions

In the literature, terms such as 'model calibration', 'fitting' and 'validation' are sometimes used to describe similar processes.[4,15,16] The simple comparison of model outputs with observed data relates to the concept of validation (or external validation), which is a familiar idea in economic evaluation.[3,10,17] Nonetheless, the use of the term 'validation' is controversial. As argued by Cooper,[4] following in the tradition of Popper,[18] if the model passes the confrontation with data several times, we can gain more credibility in the model, but we cannot be sure that the model is valid. In disease modelling, especially infectious disease modelling, 'fitting' is habitually used to describe the particular process of finding the input parameter values that generate a good fit of the model to observational data.[4,19,20] Calibration is often used as a synonym for fitting,[16,21] even though calibration can be seen as a more comprehensive process that may also take into account different model structures in the fitting procedure.[5,22] In this article, we use the term 'calibration' as a synonym for fitting, as it is more commonly used in the economic evaluation literature.

Although representing a different process, 'cross-validation' is a term that is related to the concept of calibration. It describes the comparison of results of a model with the results of other models built for a similar purpose,[3,12] but has also been used to describe the post-calibration assessment of model outputs with observed data not included in the model calibration.[6]

1.2 Rationale for Calibration

An important part of model development is to check that the predictions of the model are consistent with other data sources describing the model outputs, such as disease prevalence and mortality rates. Calibration has traditionally been seen as a way to make adjustments to 'unobserved' or unavailable parameter values,[5,23,24] in order to achieve a good fit with the data, but as we discuss, it can also be used to adjust all the epidemiological parameters.

Some calibration approaches also generate a number of different sets of plausible estimates that fit with the observed data. Used in this way, a further rationale for model calibration is that it is an additional tool to handle uncertainty surrounding the disease model beyond conventional sensitivity analysis. Importantly, because the calibration process compares the combined output predictions across all the model inputs it gives the analyst further insight into the correlations between input parameter estimates.[15,25] This is particularly beneficial given that it is often difficult to identify and quantify correlation between parameters in disease models.[15]

2. Model Calibration Methods

Model calibration and particularly model fitting resembles the estimation of coefficients in linear regression, where we try to find the coefficients of the regression function (parameter values) that identify outputs that best fit the data. In an approach similar to that of Stout et al.,[8] we have categorized the calibration process into the

following seven stages, which are discussed in this section:

1. Which parameters should be varied in the calibration process?
2. Which calibration targets should be used?
3. What measure of goodness of fit (GOF) should be used?
4. What parameter search strategy should be used?
5. What determines acceptable GOF parameter sets (convergence criteria)?
6. What determines the termination of the calibration process (stopping rule)?
7. How should the model calibration results and economic parameters be integrated?

2.1 Parameters to Include in the Calibration

The most common use of calibration is to estimate unobservable model parameters by only allowing these parameters to vary in the calibration process.[23,26,27] In the case of screening models, a common example is the clinical presentation rate in the absence of screening, because the denominator (the population of undiagnosed individuals) cannot be observed. Moreover, even when parameters have been observed directly, these parameters may have different levels of precision, leading some to advocate that all natural history and other relevant parameters in the model (unobservable and observable) should be allowed to vary in the calibration process.[6,16,22] The comprehensive inclusion of parameters facilitates the representation of correlation between input parameters, and permits the testing and adjustment of the global consistency of the model. However, it does not exclude the need to investigate and to represent the correlation in the model.

2.2 Selection of Calibration Targets

The selection of the calibration targets is another important step in the calibration process. There are no exact criteria to choose the calibration targets that are necessary to the process. However, it is sensible to say that the most important selection issue is the availability of good-quality data to use as calibration targets.[28] 'Good quality' can be basically translated into substantial sample size and lack (or limitation) of study biases. The choice is also determined by the complexity of the model, as simple models only produce a limited number of outputs to be compared with targets.

The intervention being evaluated is also going to determine the choice of target; for example, if we are evaluating a screening test to detect human papillomavirus (HPV) type 16, we should be more concerned that our model produces consistent HPV 16 prevalence rates than rates of other HPV types.

Since there are regional differences in disease epidemiology and management pathways, local data should be preferentially used as targets (e.g. cancer incidence from cancer registries). It is important that a model accurately represents the condition of the population for which the decision is being made and from which empirical data were obtained to use as targets. Where only non-local data are available to be used as targets, consideration should be given to the impact of alternative patterns of disease epidemiology and management pathways on model output.[29] If using non-local data, wider ranges should be used when defining target ranges as the convergence criteria (see section 2.5 for more information).

Calibration targets can be a single summary statistic (e.g. mean disease incidence rate) or a series of statistics (e.g. age-specific disease prevalence curve).[8,22,30] Whenever available, cohort study data should be used as calibration targets, making sure that the study population fairly represents the population for which the decision is being made and that the model's hypothetical cohort population is subjected to the same conditions as the targeted population.

The use of cross-sectional data as calibration targets deserves careful interpretation due to birth cohort effects. For example, in a cervical cancer screening model, in order to use cross-sectional data, it tends to be assumed that each member of the cohort experiences the same pattern of screening and treatment over her lifetime.[22,31] However, such patterns were significantly different before 1990 in most countries, when nation-wide screening programmes were set in place and high coverage rates were achieved across age groups.

We must also make sure that modelled behaviour patterns reflect the conditions of the targeted population. In the case of sexually transmitted diseases such as HPV, changes in sexual behaviour over time are likely to have an impact on the estimates of HPV prevalence. The most common ways to circumvent this problem are to calibrate the model against current data but include in the model changes in behaviour, epidemiology and management of the disease; or to calibrate the model against pre-1990 data to represent natural history in the absence of screening.

2.3 Goodness-of-Fit (GOF) Measures

It is important to evaluate how close the model predictions are to the target data. This can be done in a qualitative way by, for example, visually comparing the age-specific incidence curve predicted by the model and the one derived from observed data. However, this involves subjective judgements that are best avoided if we want model calibration to be more of a science than an art. In the statistics literature, the most commonly used measures of GOF are least squares, chi-squared (χ^2) [or weighted least squares] and the likelihood.[4,32] We first discuss the use of alternative GOF measures in the context of calibrating to a single target, followed by a discussion of fitting multiple targets.

2.3.1 Least Squares

Least squares relies on calculating the sum of square errors, $Q(\theta)$, between the empirical data and the model output for each input parameter value.[5,28] The values that best fit the data are those that minimize this sum.[4] In equations 1, 2 and 3, we are considering a series of statistics, for an age-specific calibration target.

$$Q(\theta) = \sum_a (y(a) - f(a|\theta))^2 \qquad \text{(Eq. 1)}$$

where θ is the input parameter or a vector of parameters (θ), $y(a)$ represents the observed data estimate (e.g. HIV incidence rate) for age a, and $f(a|\theta)$ represents the model output for age a given input parameter θ. The advantages of this approach are that it is intuitive and not very data demanding. The main disadvantage is that it does not take into account the precision of the empirical data; for example, estimates of disease incidence at different ages may come from different studies with different sample sizes and therefore have different levels of certainty.

2.3.2 Chi-Squared

The χ^2 is similar to the above measure, but it overcomes the different levels of certainty problem by dividing the least square error by its standard deviation (σ), as can be seen in equation 2.[32] Therefore, it places more weight on the more reliable estimates, those with large sample size and small standard deviation. Note that this is only one of many χ^2 tests (e.g. Pearson's χ^2) – statistical procedures whose results are evaluated by reference to the χ^2 distribution.

$$\chi^2 = \sum_a \left(\frac{y(a) - f(a|\theta)}{\sigma_a} \right)^2 \qquad \text{(Eq. 2)}$$

2.3.3 Likelihood

One of the most popular GOF measures that, like χ^2, also takes into account levels of certainty of the observed data as well as informing confidence intervals of the GOF when referring to the χ^2 distribution, is the likelihood.[16,21,33] In fact, if the measurement errors are Normally distributed, the χ^2 will give the same results as the likelihood.[4] Unlike the least squares and the χ^2, which try to minimize the result of the above functions, the aim with the likelihood is to maximize how likely a particular set of parameters is, given the empirical data.[32] For example, equation 3 describes the likelihood function for a binomial process[34] such as infection prevalence based on serological data:

$$L(\theta) = \prod_a p(a|\theta)^{y(a)} (1 - p(a|\theta))^{n(a) - y(a)} \qquad \text{(Eq. 3)}$$

where $p(a|\theta)$ represents the proportion of age a seropositive individuals predicted by the model using input parameter θ, where $y(a)$ represents the number of observed seropositives at age a. The second part of the equation, $(1 - p(a|\theta)$, refers in a similar way to the seronegatives, where $n(a)$ is the size of the sample at age a. The set of parameters that gives the maximum value of equation 3 is the best-fit set.[4] For ease of calculation,

it is common practice to work with the logarithm of the likelihood, also known as log-likelihood. Different from the likelihood form, when using the log-likelihood, the best-fit parameter set is defined by the one that best minimizes the log-likelihood estimate.

Comparing the three methods mentioned above, it is important to note that, in the likelihood approach, the probability function is often difficult to specify and compute, particularly for complex models. Furthermore, by looking at the equations, we can see that the likelihood approach requires more data than the other two methods. In the case of χ^2, the level of precision of the calibration target is only captured in the σ parameter while the likelihood approach needs, for example, the number of positives and sample size.

2.3.4 Multiple GOF Estimates

It is preferable to calibrate disease models to multiple calibration targets, in which case it is necessary to obtain a combined measure of GOF across all calibration targets. This is also called multi-objective optimization.[35] One option is to treat all calibration targets as independent targets and then sum the GOF measures across the different targets. This task can be performed using different methods (e.g. global criterion method and lexicographic method).[35,36]

The global criterion is given by the sum of GOF of each calibration target, which may be weighted. The most commonly used in the disease modelling field is the weighted GOF approach.[37,38] This approach consists of weighting each calibration GOF estimate (for each calibration target) and then summing across all targets. The weights are usually determined by the analyst or a group of experts based on the importance and/or the existence of biases in the estimate of the target.[38]

In the lexicographic approach, the calibration targets are ranked in order of importance, and the process of finding the optimal parameter values is carried out step by step, starting with the most important calibration target and proceeding according to the order of importance.[28,39]

An alternative general approach involves defining multi-dimensional integrals that represent the joint uncertainty around groups of calibration targets. Ideally, such integrals are solved analytically to identify the set of parameter values that maximize the likelihood across all calibration targets. However, in many cases, the integral cannot be solved analytically, and numerical integration methods are required (as discussed in section 2.4). It is also the case that the process of defining the correct multi-dimensional integrals is a difficult task that requires specialist mathematical expertise.[40]

2.4 Parameter Search Strategies

The terms 'parameter search strategy', 'search algorithm' or 'optimization method' all refer to the method used to search for parameter values or sets of values that produce model outputs that match specified calibration targets most closely. Broadly speaking, optimization is the process of finding the conditions that give the maximum and minimum value of a function.[35] Parameter search or optimization is a large field of operational research. There are various methods for the solution of different types of optimization problems. These methods can be classified according to the existence of constraints, the nature of the design variables, the physical structure of the problem, the nature of the equations involved, the deterministic nature of the variables, the separability of the functions, the number of objective functions, and others.[35] Unfortunately, there is no perfect optimization algorithm. It is advocated that the analyst should consider the most appropriate methods for the problem and even try more than one method, or combinations of methods, in a comparative way.[32] In the case of disease models used in economic evaluation that are usually nonlinear (e.g. Markov models, micro-simulations), with multiple-objective functions (multiple GOF estimates), with and without parameter constraints (*a priori* bounds), there are a number of alternative strategies that could potentially be applied as calibration search strategies.[16,22,30,33,38,41-44]

2.4.1 Grid Search Method

The parameter search takes place across the different possible combinations of parameter values

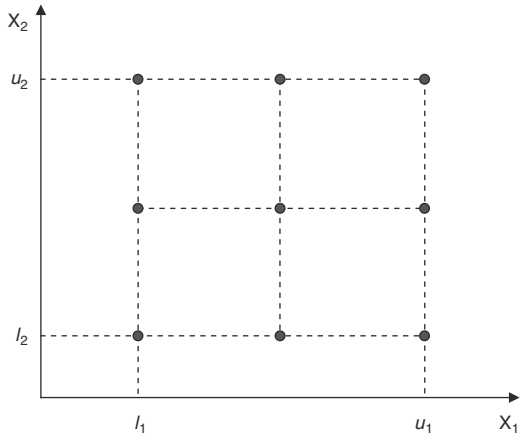

Fig. 1. Parameter space with two variables, X_1 and X_2 ($v_i = 3$). l_i, u_i = lower and upper bounds of the two variables ($i = 1,2$); v = number of possible parameter values.

(i.e. the parameter space). Conceptually, if just two parameters (X_1 and X_2) were varied in the model, the space could be represented in two dimensions. By considering this two dimensional space as in figure 1, it is simple to understand how the grid parameter search method works.[35,42] For example, if the lower and upper bounds of the two variables are l_i and u_i ($i = 1,2$), for simplicity we could divide the ranges into two equal parts, with three considered values per parameter ($v_i = 3$). This method involves setting up a suitable grid in the parameter space, evaluating the GOF estimate at all the grid points (nine points in the example), and finding the grid point that best minimizes the GOF. With each additional parameter, the number of dimensions required to represent the space also increases accordingly and, in most practical problems, the grid search method requires prohibitively large numbers of model evaluations. For example, a model including 20 parameters, only considering the $v_i = 3$, would require $3^{20} = 3\,486\,784\,401$ evaluations.

2.4.2 Random Search Method

To date, the most common approach for parameter searching that has been utilized in economic evaluation is the random search method.[5,28,45] As described in figure 2, in a random search method, distributions are assigned to each parameter in the model and multiple sets of parameter values are sampled using a random number generator.[35] Each set is then used in the model and the GOF is calculated. The set (or sets) that results in the optimum GOF result(s) is selected according to the convergence criteria (see section 2.5).

The main advantages of the random search strategy are that it is intuitive and relatively easy to programme. The main disadvantage is that random searching is not efficient in covering the entire parameter space. With a random search strategy, increasing numbers of searches improves the chance that the global extremum has been identified, but we cannot be certain that the extremum identified is global and not local. In more complex models with more parameters and larger parameter space, random search methods have limitations in the processing time required to search for the global extremum.

Many parameter search strategies such as random search, as well as probabilistic sensitivity analysis, employ sampling methods in order to obtain values from the parameter distributions. There are various sampling methods that can be used to sample from distributions. The random

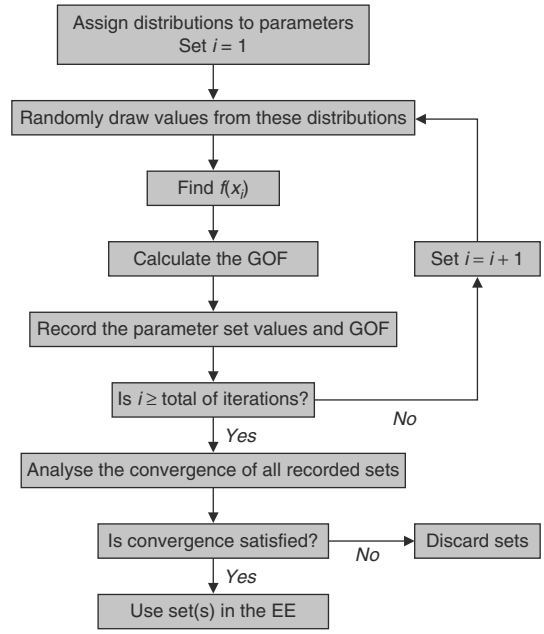

Fig. 2. Random search method in economic evaluation models. **EE** = economic evaluation; $f(x_i)$ = output of the model for iteration i; **GOF** = goodness of fit; i = number of current iteration.

sample is the most obvious alternative, even though it is not the most efficient way to sample the parameter space. As shown in figure 3a, it may not widely cover the parameter space. A more efficient and increasingly popular sampling method is Latin hypercube, which was introduced to the field of disease modelling by Blower and Dowlatabadi.[46] For each parameter, a probability density function is defined and divided into *n* intervals with the same probability (figure 4). A parameter value is picked randomly from every interval and this procedure is performed for every parameter. As can be seen in figure 3b, a parameter value from each sampling interval is used only once in the analysis.

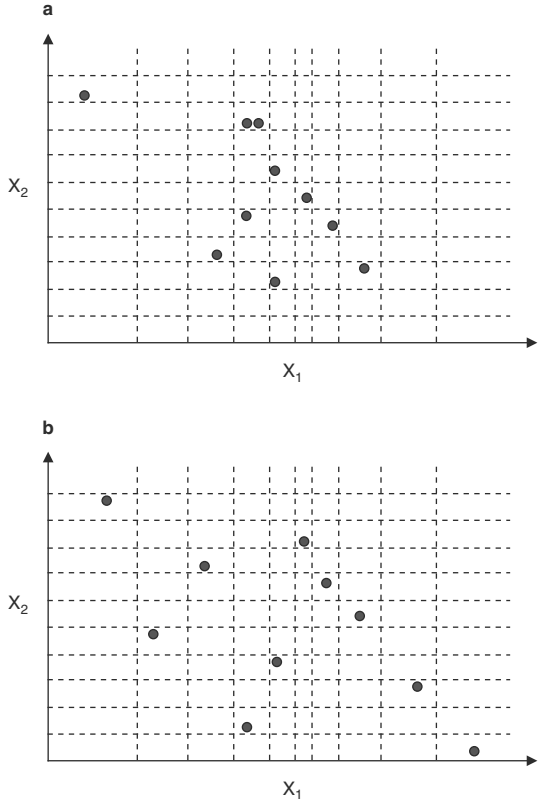

Fig. 3. Examples of two sampling methods: (**a**) random sampling and (**b**) Latin hypercube sampling for a simple case of ten samples (samples for parameter X_1 that follows a Normal distribution and X_2 that follows a Uniform distribution that can be found in figure 4). In the random sampling, some areas are not sampled and others are more greatly sampled; in the Latin hypercube, a value is chosen once and only once from each interval.

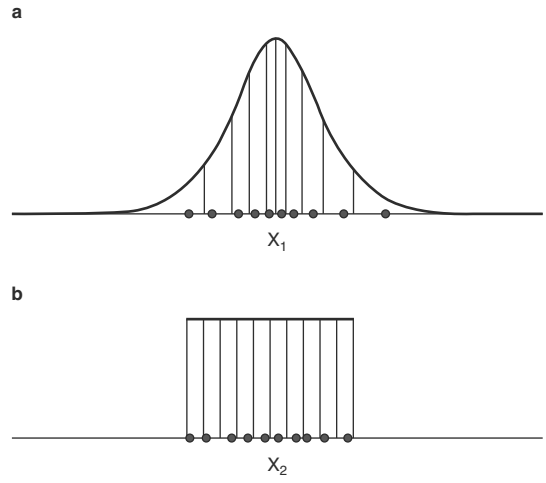

Fig. 4. Latin hypercube sampling in a (**a**) Normal distribution (parameter X_1) and (**b**) Uniform distribution (parameter X_2). Examples of probability density functions associated with parameters X_1 and X_2 used in figure 3. Since Latin hypercube sampling was used, the distributions were divided in intervals with equal probability, and one sample was obtained from each of those intervals.

2.4.3 Generalized Reduced Gradient Method

One of the most widely used optimization tools is the Microsoft® Excel Solver. In the case of nonlinear models, it employs a generalized reduced gradient method, as implemented in the GRG2 code.[47,48] The gradient methods make use of the gradient of a function, which is an *n*-parameters vector given by equation 4:

$$\nabla f = \begin{Bmatrix} \frac{df}{dx_1} \\ \frac{df}{dx_2} \\ \vdots \\ \frac{df}{dx_n} \end{Bmatrix} \quad \text{(Eq. 4)}$$

The gradient has an important property. If we move along the gradient direction from any point in the parameter space, the function value increases at the fastest rate. Therefore, the negative of the gradient vector represents the direction of steepest descent. Optimization methods that use the gradient vector can be expected to find the minimum point faster. As the name suggests, the generalized reduced gradient method is a modified version of the reduced gradient method that

was presented originally for solving problems with linear constraints only.

In order to solve the optimization problem that is presented in a spreadsheet format, Microsoft® Excel Solver extracts the problem from the spreadsheet cells and internally builds a representation of the model that is suitable for the generalized reduced gradient method. In more general terms, this is the Jacobian matrix of partial derivatives of the problem functions (objective and constraints) with respect to the decision variables.[47] In linear models, the matrix entries are constant, and only need to be evaluated once at the start of the optimization. In nonlinear models, the Jacobian matrix entries are variable and have to be re-calculated at each new trial point. The Jacobian matrix is approximated using the finite differences method.[49]

When using the Microsoft® Excel Solver, it is important to remember that it assumes the model to be nonlinear as default. The path and scaling factors used by the generalized reduced gradient method will depend on the starting point. It is recommended that different starting points are tried. If the software reaches roughly the same final point, we can be fairly confident that this is a global extremum. Otherwise, the best results of the solutions obtained can be selected or other optimization methods can be tried.

2.4.4 Downhill Simplex Method (Nelder-Mead)

Downhill simplex also known as the Nelder-Mead method does not require the evaluation of derivates like the gradient methods, only function evaluation. It is not as fast as some gradient methods. Nonetheless, it is a very popular optimization method, because it requires concise code, and makes almost no special assumption about the function being minimized.[32] A simplex is a geometrical figure consisting, in N dimensions, of $N+1$ points (or vertices) and all their interconnecting line segments. In two dimensions the simplex is a triangle. In three dimensions it is a tetrahedron, as represented in figure 5. As previously mentioned, the number of dimensions is determined by the number of input parameters varied in the optimization process.

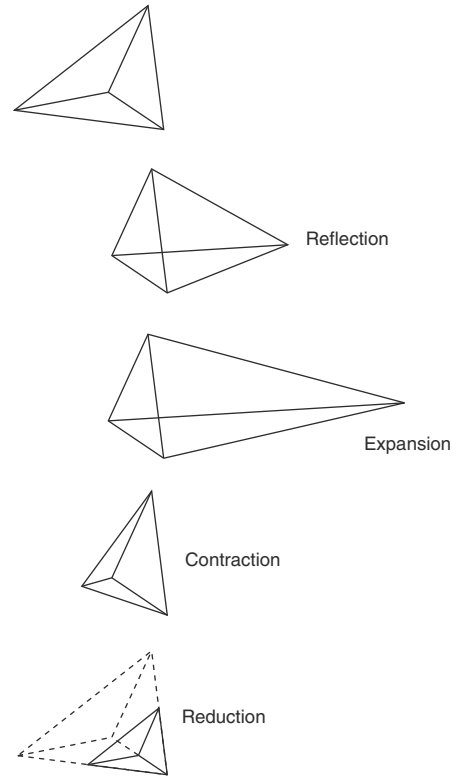

Fig. 5. Downhill simplex steps.

The downhill simplex method must be initialized not just with one point (set of parameter values) but with $N+1$ points, in order to constitute an initial simplex. By conceptualizing the disease model's GOF as a surface with peaks (poorly fitting parameter sets) and valleys (better fitting parameter sets), the downhill simplex method takes a series of steps (reflection, expansion, contraction, reduction), as represented in figure 5. Most steps just move the point of the simplex where the GOF is largest ('highest point') through the opposite face of the simplex to a lower point to search for potential areas of the parameter space that might better fit the data. The movement of the simplex resembles an amoeba searching for a 'valley floor'. The main disadvantage is that it can be slow and only one best-fit parameter set emerges at the end of the process. In order to gain more confidence that the best-fit parameter set does not represent a local extremum, the

algorithm is usually run a few times from different starting points (different simplexes).[35,50]

2.4.5 Simulated Annealing Method

Simulated annealing is a more complex parameter search method that has attracted significant attention as an efficient alternative for large-scale optimization problems,[7,35,38] particularly those where a desired global extremum is hidden among many poorer local extrema. Simulated annealing is based on the thermodynamics of the crystallization of metal, where parameter searching involves the introduction of an artificial parameter (called temperature) that determines the probability of accepting a set of random parameter values. At initial high temperatures, the probability of accepting a new set of parameter values is higher, which means that the algorithm is allowed to widely explore the parameter space. Like in the downhill simplex, by conceptualizing the model's GOF as a surface with peaks (poorly fitting parameter sets) and valleys (better fitting parameter sets), it is apparent that bigger 'jumps' avoid the algorithm falling into a local minimal GOF. Slowly decreasing the temperature allows the algorithm to find the parameter set with the lowest GOF.[35,38]

As in the downhill simplex method, in simulated annealing, only one parameter set emerges at the end of the process. However, simulated annealing is more efficient than the downhill simplex and it can also be used in problems of combinatorial optimization. In the case of disease models, this would allow us to consider sets of possible model structures in the calibration process.[32] In a recent study, Chung Yin et al.[38] found that simulated annealing outperformed genetic algorithm in the calibration of a micro-simulation model, the Lung Cancer Policy Model.[51]

2.4.6 Mixed Approaches

Mixed approaches have been proposed where methods such as random search or grid search can be used to predict the region of the parameter space in which the global extremum is placed. Once this region is located, more efficient guided techniques can be used to find the precise location of the global extremum.[35] In general, if time allows, analysts should consider the application of more than one method or combinations of methods in a comparative way.[32,35]

2.5 Convergence (or Acceptance) Criteria

Convergence criteria, acceptance criteria and the acceptance threshold are terms that describe the process of defining acceptable sets of input parameter values. In the example of the random search method described in section 2.4.2, if the analyst is only looking for the parameter set that best minimizes (or maximizes, depending on the GOF measure used) the GOF estimate, this is the acceptance criterion.[23] However, there are potentially more than one parameter set that can give the same GOF estimate.

Moreover, to inform analysis of uncertainty, it is necessary to identify sets of input parameter values that produce an acceptable fit according to the analyst's objectives.[35] Analysts often define a GOF threshold based on 'plausible' visual fit.[22,28] This means that the predicted output parameters of many parameter sets are plotted and the analyst arbitrarily defines the worst fitting set that is acceptable. The GOF of this parameter set is used as the threshold value and all the parameter sets that produce a better GOF in comparison with the threshold are deemed acceptable.

Another approach is to define target ranges based on the data informing the calibration target(s) and select those parameter sets that produce model output within those ranges.[6,52] An alternative approach is to define a confidence interval around the GOF of the best-fit parameter set and to deem acceptable (or statistically indistinguishable) all the parameter sets with GOF estimates within that interval.[16,37]

2.6 Stopping Rule

The stopping rule or termination criteria determine whether the calibration process (or the search for parameters) is complete. There are two broad criteria that can be used: acceptability of the convergence of the model outputs to the observed calibration targets and/or completion of a specified number of searches (or iterations within the parameter space).[22,35,38]

A simple calibration objective (or convergence criterion) may require that one parameter set is identified for which the model outputs are within the 95% confidence intervals of the observed calibration target values, or that a specified number of parameter sets achieve that level of accuracy. Currently, there is a lack of empirical evidence on how to specify the number of parameter sets to terminate the calibration process; in practice, this number is determined by the parameter space, the parameter search strategy, the convergence criteria and the available computational power.

2.7 Integrating the Results of the Calibration and the Economic Parameters

The last step of the process is to integrate the results of the model calibration within the full economic model. There are many ways of doing this, and the choices made in the previous steps of the calibration process will determine the most sensible way. The simplest approach is to use the point estimates derived from the best fitting set of calibrated input parameters. However, where probabilistic sensitivity analysis (PSA) is used, a more elaborate approach is required. Treating the calibrated parameters as independent parameters, fitted values that passed the acceptance criteria may be used to derive an independent probability distribution for each parameter. However, the ability to represent parameter correlation is an important attribute of the calibration process and so a PSA should reference all parameter sets deemed acceptable in the calibration process. Two broad alternatives are to report the range of cost-effectiveness results associated with multiple parameter sets within the acceptance region, implicitly assigning an equal probability of relevance to all included parameter sets,[6] or to sample acceptable parameter sets one at a time, with the probability of a parameter set being sampled defined as a function of its overall GOF.[53-57]

3. Bayesian Methods

In section 2.5, we introduced the idea of generating a plausible range for parameter values, rather than a single 'best' value. This implicitly introduces the concept of parameter uncertainty, without providing a formal theoretical basis to analyse it. 'Classical' or 'frequentist' statistical theory cannot provide such a framework, as it assumes uncertainty can only be assigned to data, not parameters. If a theoretical underpinning for dealing with parameter uncertainty is required, this can be achieved by the use of Bayesian methods.[58]

Bayesian updating involves defining a prior distribution for the model parameter set. Once defined, the prior can be updated via Bayes' theorem to reflect the additional information given in the likelihood function for the data, to give the posterior (i.e. updated) distribution reflecting the remaining uncertainty around the true values of the parameter. An approach known as Markov Chain Monte Carlo (MCMC) can be used to generate a sample from the joint posterior density function of the model parameters.[58] The resulting samples will capture the degree of correlation between parameter values implied by the data and the model structure, so that uncertainty around cost effectiveness is accurately stated.[59] Software packages such as WinBUGS can be used to fit models to data and generate MCMC samples.

Whilst MCMC methods overcome some of the computational difficulties involved in Bayesian model calibration, many challenges remain. One issue is regarding how the prior distribution should be specified. A common approach is to set 'vague' priors, so that their influence on the posterior distribution is negligible. Ideally, this should be confirmed with sensitivity analysis using several alternative vague priors, to assess whether this choice has any meaningful impact on the results. Whilst some see the subjective nature of prior distributions as undermining the approach,[60] it can also be seen as one of its strengths. Priors can be chosen to reflect sources of information beyond the dataset, thereby more appropriately reflecting the available evidence base. Elicitation of expert opinion is a common source of this additional information.[60]

A further challenge is that, despite the benefits of MCMC, computation of the posterior distribution may still be computationally expensive. This can be a particular issue where the likelihood is a complex function of the model parameters, as

may often be the case when calibration is required. De Angelis et al.[61] provide an example where Bayesian methods are used to estimate the prevalence of hepatitis C virus. The disease is asymptomatic for most of its long incubation time, and direct data on prevalence are not available. The authors developed a WinBUGS model to estimate this parameter from indirect information such as the results of screening programmes in at-risk populations. As well as only informing the desired parameter indirectly, the available data sources were potentially biased. The authors demonstrated how Bayesian evidence synthesis can be used to explore the impact of alternative models of this bias, and presented several measures of GOF that can be used in Bayesian calibration models. Other examples of this recent approach include Welton and Ades[62] and Goubar et al.[63]

4. Discussion

Calibration of computer models is being used actively in many research fields. Disease models used in economic evaluation require information of various types and sources with different levels of certainty. Calibration is not only a useful tool for estimating parameters but also a way of dealing with model uncertainty by testing and adjusting the consistency of the model when compared with empirical data. We have divided the model calibration process in economic evaluation into seven steps and examined the different methods used in each step. The seven steps are (i) Which parameters should be varied in the calibration process? (ii) Which calibration targets should be used? (iii) What measure of GOF should be used? (iv) What parameter search strategy should be used? (v) What determines acceptable GOF parameter sets (convergence)? (vi) What determines that the calibration process should stop? (vii) How should the model calibration results and economic parameters be integrated?

We selectively reviewed examples from the current economic evaluation and operational research literature. A limitation of our review is that it is not systematic and there are further examples and methodological research that we were not able to cover.

There is a lack of guidance on calibration methods for economic evaluation. For example, we reviewed the main health economics textbooks for guidance on calibrating decision models.[1,13,64-66] Only one[13] addressed calibration methods: "There may be several parameters within the model … which can be dialled up and dialled down to try to achieve calibration. While statistical methods can, in principle, be used to achieve optimal calibration … in practice the process of calibration is more art than science." This sole and rather sceptical statement is a good illustration of how the lack of standards can undermine the credibility of a methodology.

We also reviewed the guidelines in economic evaluation and decision analysis in health from the UK,[9] Canada,[10] Australia[11] and the International Society for Pharmacoeconomics and Outcomes Research (ISPOR)[12] for guidance on calibration methods. The UK[9] and Canadian[10] guidelines recommend only that the model should be validated. The Canadian guidelines particularly recommend that the results of the model (e.g. health outcomes) should be calibrated and compared against reliable independent data sets (e.g. national cancer statistics). Any difference should be explained, or used to inform adjustments in the model. The UK guidelines make a strong case about the need to explore uncertainty in the model; they also suggest that a PSA is the preferred method to deal with parameter uncertainty. A brief recommendation could be found in the report of the ISPOR Task Force on Good Research Practice in Modelling Studies:[12] "Models should be calibrated … when there exist data on both model outputs and model inputs, over the time frame being modelled … The calibration data should be from sources independent of the data used to estimate input parameters in the model." Nothing was found in the Australian guideline.[11]

An accompanying article in this issue of *PharmacoEconomics* presents an applied example of a calibrated model. Karnon and Vanni[14] work through the seven stages of calibration, providing practical guidance on a widely applicable calibration process for health economic

decision models given current evidence on calibration processes in economic evaluation.

5. Conclusions

As presented in this review, a considerable number of studies that apply calibration methods in economic evaluation can already be found in the literature. However, if we are to change the current sceptical view of calibration processes and incentivize good practice in usage, we ought to unify the way we approach the problems and report the methods used, and continue to investigate different methods. Further research is required to systematically investigate the engineering and operations research literature for calibration methods that could be used in economic evaluation. Empirical research is also required to assess the impact of different ways of calibrating economic models on the final economic results. This should be done for the different levels of the calibration process. Additional investigation of the performance of different calibration methods in the particular case of disease modelling is also important. All this further evidence will permit us to better define what constitutes good practice when calibrating models for economic evaluation, and to improve guideline recommendations.

Acknowledgements

No sources of funding were used to conduct this study or prepare this manuscript. The authors have no conflicts of interest that are directly relevant to the content of this review.

For the invaluable advice provided, the authors thank Michael Pickles and Andrew Cox.

References

1. Drummond MF SM, Torrance GW, et al. Methods for the economic evaluation of health care programmes. 3rd ed. New York: Oxford University Press, 2005
2. Box G, Draper N. Empirical model-building and response surfaces. 1st ed. New York: John Wiley & Sons, 1987
3. Philips Z, Ginnelly L, Sculpher M, et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technol Assess 2004 Sep; 8 (36): iii-iiv, ix-xi, 1-158
4. Cooper BS. Confronting models with data. J Hosp Infect 2007 Jun; 65 Suppl. 2: 88-92
5. Weinstein MC. Recent developments in decision-analytic modelling for economic evaluation. Pharmacoeconomics 2006; 24 (11): 1043-53
6. Van de Velde N, Brisson M, Boily M-C. Modeling human papillomavirus vaccine effectiveness: quantifying the impact of parameter uncertainty. Am J Epidemiol 2007 Apr 1; 165 (7): 762-75
7. Jit M, Choi YH, Edmunds WJ. Economic evaluation of human papillomavirus vaccination in the United Kingdom. BMJ 2008 Jul 17; 337: a769
8. Stout NK, Knudsen AB, Kong CY, et al. Calibration methods used in cancer simulation models and suggested reporting guidelines. Pharmacoeconomics 2009; 27 (7): 533-45
9. National Institute for Health and Clinical Excellence. Guide to the methods of technology appraisal. London: NICE, 2008 [online]. Available from URL: http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf [Accessed 2009 Jun 23]
10. Canadian Agency for Drugs and Technologies in Health. Guidelines for the economic evaluation of health technologies: Canada. 3rd ed. Ottawa (ON): CADTH, 2006 [online]. Available from URL: http://www.cadth.ca/media/pdf/186_EconomicGuidelines_e.pdf [Accessed 2009 Jun 23]
11. Pharmaceutical Benefits Advisory Committee. 1995 guidelines for the pharmaceutical industry on preparation of submission to the Pharmaceutical Benefits Advisory Committee. Woden (SA): PBAC, 1995 [online]. Available from URL: http://www.health.gov.au/internet/main/publishing.nsf/Content/health-pbs-general-pubs-pharmpac-part1.htm [Accessed 2009 Jun 23]
12. Weinstein MC, O'Brien B, Hornberger J, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices – Modeling Studies. Value Health 2003 Jan-Feb; 6 (1): 9-17
13. Drummond MF, McGuire A. Economic evaluation in health care: merging theory with practice. 1st ed. New York: Oxford University Press, 2001
14. Karnon J, Vanni T. Calibrating models in economic evaluation: a comparison of alternative measures of goodness-of-fit, parameter search strategies and convergence criteria. Pharmacoeconomics 2011; 29 (1): 51-62
15. Ades AE, Claxton K, Sculpher M. Evidence synthesis, parameter correlation and probabilistic sensitivity analysis. Health Econ 2006; 15: 373-81
16. Kim JJ, Kuntz KM, Stout NK, et al. Multiparameter calibration of a natural history model of cervical cancer. Am J Epidemiol 2007 Jul 15; 166 (2): 137-50
17. Kim L, Thompson S. Uncertainty and validation of health economic decision models. Health Econ 2010; 19 (1): 43-55
18. Popper K. Conjectures and refutations: the growth of scientific knowledge. London: Routledge, 1963
19. Foss AM, Watts CH, Vickerman P, et al. Could the CARE-SHAKTI intervention for injecting drug users be maintaining the low HIV prevalence in Dhaka, Bangladesh? Addiction 2007; 102 (1): 114-25
20. Williams JR, Foss AM, Vickerman P, et al. What is the achievable effectiveness of the India AIDS initiative

intervention among female sex workers under target coverage? Model projections from southern India. Sex Transm Infect 2006; 82 (5): 372-80

21. Goldhaber-Fiebert JD, Stout NK, Ortendahl J, et al. Modeling human papillomavirus and cervical cancer in the United States for analyses of screening and vaccination. Popul Health Metr 2007; 5 (1): 11
22. Jit M, Gay N, Soldan K, et al. Estimating progression rates for human papillomavirus infection from epidemiological data. Med Decis Making 2010 Jan-Feb; 30 (1): 84-98
23. Suárez E, Smith JS, Bosch FX, et al. Cost-effectiveness of vaccination against cervical cancer: a multi-regional analysis assessing the impact of vaccine characteristics and alternative vaccination scenarios. Vaccine 2008; 26 Suppl. 5: F29-45
24. Berkhof J, Bruijne MCD, Zielinski GD, et al. Evaluation of cervical screening strategies with adjunct high-risk human papillomavirus testing for women with borderline or mild dyskaryosis. Int J Cancer 2006; 118 (7): 1759-68
25. Welton NJ, Ades AE. Estimation of markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis, and model calibration. Med Decis Making 2005 Nov-Dec; 25 (6): 633-45
26. Anderson R, Haas M, Shanahan M. The cost-effectiveness of cervical screening in Australia: what is the impact of screening at different intervals or over a different age range? Aust N Z J Public Health 2008; 32 (1): 43-52
27. Mandelblatt J, Schechter CB, Lawrence W, et al. The SPECTRUM population model of the impact of screening and treatment on US breast cancer trends from 1975 to 2000: principles and practice of the model methods. J Natl Cancer Inst Monogr 2006; (36): 47-55
28. Karnon J, Goyder E, Tappenden P, et al. A review and critique of modelling in prioritising and designing screening programmes. Health Technol Assess 2007 Dec; 11 (52): iii-iv, ix-xi, 1-145
29. Pickles M, Foss AM, Vickerman P, et al. Interim modelling analysis to validate reported increases in condom use and assess HIV infections averted among female sex workers and clients in southern India following a targeted HIV prevention programme. Sex Transm Infect 2010 Feb; 86 Suppl. 1: i33-43
30. McMahon PM, Kong CY, Weinstein MC, et al. Adopting helical CT screening for lung cancer. Cancer 2008; 113 (12): 3440-9
31. Kim JJ, Wright TC, Goldie SJ. Cost-effectiveness of human papillomavirus DNA testing in the United Kingdom, the Netherlands, France, and Italy. J Natl Cancer Inst 2005; 97 (12): 888-95
32. Press WH, Teukolsky SA, Vetterling WT, et al. Numerical recipes in C: the art of scientific computing. 2nd ed. Cambridge (NY): Cambridge University Press, 1992
33. Trotter CL, Edmunds WJ. Modelling cost effectiveness of meningococcal serogroup C conjugate vaccination campaign in England and Wales. BMJ 2002; 324 (7341): 809
34. Keeling M. Modeling infectious diseases in humans and animals. 1st ed. Princeton (NJ): Princeton University Press, 2007
35. Rao SS. Engineering optimization: theory and practice. 4th ed. Chichester: John Wiley & Sons, 2009
36. Freitas AA. A critical review of multi-objective optimization in data mining: a position paper. SIGKDD Explorations 2004; 6 (2): 77-86
37. Goldie SJ, Kim JJ, Kobus K, et al. Cost-effectiveness of HPV 16, 18 vaccination in Brazil. Vaccine 2007; 25 (33): 6257-70
38. Chung Yin K, Pamela MM, Gazelle GS. Calibration of disease simulation model using an engineering approach. Value Health 2009; 12 (4): 521-9
39. Tappenden P, Chilcott J, Eggington S, et al. Option appraisal of population-based colorectal cancer screening programmes in England. Gut 2007 May 1; 56 (5): 677-84
40. Wong S. Computational methods in physics and engineering. 2nd ed. Singapore: World Scientific Publishing Co. Pte. Ltd, 1997
41. Forrester M, Pettitt A, Gibson G. Bayesian inference of hospital-acquired infectious diseases and control measures given imperfect surveillance data. Biostat 2007 Apr 1; 8 (2): 383-401
42. Goldie SJ, Weinstein MC, Kuntz KM, et al. The costs, clinical benefits, and cost-effectiveness of screening for cervical cancer in HIV-infected women. Ann Intern Med 1999; 130: 97-107
43. Dayhoff JE, DeLeo JM. Artificial neural networks. Cancer 2001; 91 (S8): 1615-35
44. Tan SYGL, van Oortmarssen GJ, Piersma N. Estimating parameters of a microsimulation model for breast cancer screening using the score function method. Ann Operat Res 2003; 119 (1): 43-61
45. Vanni T, Legood R, Franco E, et al. Cost-effectiveness of strategies for managing women presenting atypical squamous cells of unknown significance in Brazil. London: London School of Hygiene and Tropical Medicine, 2008
46. Blower S, Dowlatabadi H. Sensitivity and uncertainty analysis of complex-models of disease transmission – an HIV model, as an example. Int Stat Rev 1994; 62 (2): 229-43
47. Fylstra D, Lasdon L, Watson J, et al. Design and use of the Microsoft Excel Solver. INTERFACES 1998 Sep 1; 28 (5): 29-55
48. Lasdon LS, Waren AD, Jain A, et al. Design and testing of a generalized reduced gradient code for nonlinear programming. ACM Trans Math Softw 1978; 4 (1): 34-50
49. Gill PE, Murray W, Wright MH. Practical optimization. San Diego (CA): Academic Press, 1981
50. Taylor DCA. Methods of model calibration: a comparative approach. ISPOR 12th Annual International Meeting; 2007 May 18-23; Arlington (VA) [online]. Available from URL: http://www.ispor.org/awards/12meet/MC1-Taylor.pdf [Accessed 2009 Dec 12]
51. Vanni T, Legood R, White RG. Calibration of disease simulation model using an engineering approach. Value Health 2010; 13 (1): 157
52. Fryback DG, Stout NK, Rosenberg MA, et al. The Wisconsin breast cancer epidemiology simulation model. J Natl Cancer Inst Monogr 2006; (36): 37-47
53. Karnon J, Czoski-Murray C, Smith KJ, et al. A hybrid cohort individual sampling natural history model of age-related macular degeneration: assessing the cost-effective-

ness of screening using probabilistic calibration. Med Decis Making 2009 May 1; 29 (3): 304-16

54. Karnon J, Campbell F, Czoski-Murray C. Model-based cost-effectiveness analysis of interventions aimed at preventing medication error at hospital admission (medicines reconciliation). J Eval Clin Pract 2009; 15 (2): 299-306
55. Carlton J, Karnon J, Czoski-Murray C, et al. The clinical effectiveness and cost-effectiveness of screening programmes for amblyopia and strabismus in children up to the age of 4–5 years: a systematic review and economic evaluation. Health Technol Assess 2008 Jun; 12 (25): iii, xi-194
56. Karnon J, Jones R, Czoski-Murray C, et al. Cost-utility analysis of screening high-risk groups for anal cancer. J Public Health 2008; 30 (3): 293-304
57. Karnon J, McIntosh A, Dean J, et al. A prospective hazard and improvement analytic approach to predicting the effectiveness of medication error interventions. Saf Sci 2007; 45 (4): 523-39
58. Gelman A, Carlin JB, Stern HS, et al. Bayesian data analysis. 2nd ed. Boca Raton (FL): Chapman and Hall/CRC, 2004
59. Welton NJ, Ades AE. Estimations of Markov transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis and model calibration. Med Decis Making 2005; 25 (6): 633-45
60. O'Hagan A, Buck CE, Daneshkhah A, et al. Uncertain judgements: eliciting expert probabilities. Chichester: Wiley, 2006
61. De Angelis D, Sweeting M, Ades A, et al. An evidence synthesis approach to estimating hepatitis C prevalence in England and Wales. Stat Methods Med Res 2009; 18: 361-79
62. Welton NJ, Ades AE. A model of toxoplasmosis incidence in the UK: evidence synthesis and consistency of evidence. J R Stat Soc Ser C Appl Stat 2005; 54 (2): 385-404
63. Goubar A, Ades AE, Angelis DD, et al. Estimates of human immunodeficiency virus prevalence and proportion diagnosed based on Bayesian multiparameter synthesis of surveillance data. J R Stat Soc Ser A Stat Soc 2008; 171 (3): 541-80
64. Morris S, Devlin N, Parkin D. Economic analysis in health care. 1st ed. Chichester: Wiley, 2007
65. Gold MR, Siegel JE, Russell LB, et al. Cost-effectiveness in health and medicine. 1st ed. New York: Oxford University Press, 1996
66. Briggs A, Sculpher M, Claxton K. Decision modelling for health economic evaluation. 1st ed. Oxford: Oxford University Press, 2007

Correspondence: Dr *Tazio Vanni*, Health Services Research Unit, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK.
E-mail: tazio.vanni@lshtm.ac.uk