Pharmacoeconomics 2000 May; 17 (5): 461-477
1170-7690/00/0005-0461/$20.00/0

Assessing Quality in Decision Analytic Cost-Effectiveness Models

A Suggested Framework and Example of Application

Mark Sculpher,[1] *Elisabeth Fenwick*[1] and *Karl Claxton*[2]

1 Centre for Health Economics, University of York, York, England
2 Department of Economics and Related Studies, University of York, York, England

Contents

Abstract

Despite the growing use of decision analytic modelling in cost-effectiveness analysis, there is a relatively small literature on what constitutes good practice in decision analysis. The aim of this paper is to consider the concept of 'validity' and 'quality' in this area of evaluation, and to suggest a framework by which quality can be demonstrated on the part of the analyst and assessed by the reviewer and user.

The paper begins by considering the purpose of cost-effectiveness models and argues that the their role is to identify optimum treatment decisions in the context of uncertainty about future states of the world. The issue of whether such models can be defined as 'scientific' is considered. The notion that decision analysis undertaken at time t can only be considered scientific if its outputs closely predict the results of a trial undertaken at time t+1 is rejected as this ignores the need to make decisions on the basis of currently available evidence. Rather, the scientific characteristic of decision models is based on the fact that, in principle at least, such analyses can be falsified by comparison of two states of the world, one where resource allocation decisions are based on formal decision analysis and the other where such decisions are not. This section of the paper also rejects the idea of exact codification of scientific method in general, and of decision analysis in particular, as this risks rejecting potentially valuable models, may discourage the

development of novel methods and can distort research priorities. However, the paper argues that it is both possible and necessary to develop a framework for assessing quality in decision models.

Building on earlier work, various dimensions of quality in decision modelling are considered: model structure (disease states, options, time horizon and cycle length); data (identification, incorporation, handling uncertainty); and consistency (internal and external). Within this taxonomy a (nonexhaustive) list of questions about quality is suggested which are illustrated by their application to a specific published model. The paper argues that such a framework can never be prescriptive about every aspect of decision modelling. Rather, it should encourage the analyst to provide an explicit and comprehensive justification of their methods, and allow the user of the model to make an informed judgment about the relevance, coherence and usefulness of the analysis.

Decision analytic models are widely used as a means of establishing the most cost-effective intervention from amongst mutually exclusive alternatives. The use of models as vehicles for cost-effectiveness analysis has been described,[1-4] and their advantages and limitations highlighted,[5,6] but there is a relatively small literature on how to assess the 'validity' or 'quality' of these methods. Indeed, there are some very important conceptual issues around these terms that have not been fully explored. The aim of this paper is to consider the conceptual problems of assessing validity and quality in decision analytic cost-effectiveness models, and to suggest a framework for judging quality when models are being developed and reviewed.

Section 1 considers the meaning of the term 'validity' in models in the context of scientific method. Section 2 proposes a taxonomy of quality for models which suggests some dimensions of quality and, within each, ideas on what might be accepted as good practice. This culminates in a non-exhaustive list of possible questions that might be applied to a model by both the analyst and users to assess quality. By way of illustration, these questions are then applied to a specific model.

1. The Concept of Validity in Decision Analytic Models

1.1 Falsifying Models

Assessing the validity or quality of decision analytic models poses 2 fundamental questions: (i) can decision analytic modelling be regarded as a scientific activity and (ii) if it can, how can we identify good methods? Before we can address the former we must come to some view of what constitutes scientific method. The notion of falsification (or at least some external test of a theory) has become the generally accepted idea of scientific method.[7] It seems clear that theories that simply cannot be falsified should not be regarded as scientific. Although dogmatic and extreme interpretations of falsification are not useful,[8] the idea that theories – in this case, decision analytic models – cannot be self-validating seems obvious.

So to be regarded as scientific, decision analytic models must be falsifiable, at least in principle. Although we can argue that each of the model inputs are themselves falsifiable, this fails to address the key question of whether the way they are used in the model (the model structure and inputs combined) is valid. One approach would be to test the predictions of the model by comparing them with the results of a traditional clinical trial. However, this will not be an appropriate comparison for a number of reasons. First, clinical trials have no decision context (they create controlled conditions that may not match routine clinical practice); secondly, they do not include all relevant endpoints (arguments in the decision maker utility function); thirdly, the follow-up period may not match the time horizon for the decision; and, finally, there will be non-random selection into the trial (including patients, participating centres and participating clinicians).[9]

Even if we can imagine an idealised pragmatic clinical trial,[10] which includes all relevant endpoints in its design, has a follow-up period sufficient for decision making and randomly selects entrants, it would not provide a valid test of the predictions of a model. The problem is that a model should combine all the information available at the time the decision must be made (t). But the results of the trial at t+1 will be based on the new information generated by the trial which was not available at t. The predictions of the model and the results of the trial are based on entirely different sets of information, so any comparison cannot falsify the model (any differences that arise can be explained).

The real problem is that clinical trials and decision analytic models are not directly comparable because their objectives are fundamentally different: a model combines information already available in an explicit and formal framework; a clinical trial generates new information about one or more parameters of interest. Clinical trials (and observational studies) can provide inputs for decision analytic models but cannot provide a valid test of their predictions. In order to determine whether models can be falsified, the purpose of decision analytic modelling must be carefully considered.

1.2 The Purpose of Decision Models

The purpose of models is not to predict the results of even an ideal pragmatic trial (or observational study) but to inform decision making at a particular point in time. Therefore, the testable and falsifiable hypothesis posed by modelling is that a better decision will be made at time t by using a model than by not using it. This is testable, in principle, by randomly allocating decision makers to use and not to use the model at t. The model is valid (i.e. useful) if the outcomes (costs and effects) are 'better' at t+1 for the group using the model. In this formulation, a model developed and analysed at t that predicts mortality of 20% at t+1 when, at t+1, mortality is actually 40%, may still have been a valid (i.e. useful) model if it improved decisions made at t. Similarly, a model constructed at t that correctly predicted mortality of 20% at t+1 may have been an invalid (i.e. useless) model if it did not improve decisions (when mortality is not the only argument in the decision maker's utility function). Other commentators seem to take a similar position: 'Validity refers to the ability of a decision model to recommend optimal decisions. Short of a clinical trial of a decision model, validity of a recommended decision cannot be assessed because there is no gold standard for the quality of a decision.'[11]

This is not to say that comparisons with clinical trials, observational studies or other models are not important in the process of building a model (see section 2.2.3). However, these checks on external consistency cannot falsify the model. When unexpected discrepancies between the model and external sources occur, this should prompt interrogation of both the model and the external source to account for the apparent inconsistency.

Models should combine all the information readily available at a particular point in time (model inputs) with current clinical/biological theory about disease processes (reflected in the structure). They are updated as information and theory changes. Since information accumulates over time, all models developed at t will be 'wrong' (in terms of their predictions) at t+1, and we know this at t. All models may be 'wrong' but some are useful, and it is the usefulness of models that is the appropriate test of validity not the accuracy of predictions. Indeed, this case has been made about scientific activities in general.[12]

1.3 Assessing the Quality of Models

The fact that the usefulness of models is testable, at least in principle, means that we can claim this activity to be scientific, rather than simply telling irrefutable stories or myths. However, in practice, models are developed to inform decisions that cannot be deferred (without substantial resource and health costs), so it may not be efficient to test every model in the way in section 1.2. In fact, it may be self defeating, because once the model has been tested (which takes time) it may be of limited use for subsequent decision making. What is required

is some means of assessing the quality of a model (the structure and the inputs), and some criteria to judge competing models before decisions about service provision must be made. However, in developing an assessment of quality, a universal abstract codification (guideline) of what constitutes good method may not be possible or desirable: 'There does not exist (at least at present) a complete codification of scientific rationality, and we seriously doubt that one could ever exist.'[13] 'The idea that science can, and should, be run according to fixed and universal rules, is both unrealistic and pernicious.'[14]

Sokal and Bricmont[13] and Feyerabend[14] are addressing the natural sciences, and physics in particular. If it is not possible to develop universal codes or guidelines in physics, then it is even less likely to be available in the social sciences, which address more complex and uncertain problems. Methodological guidelines are available for some activities, such as the reporting of randomised controlled trials (RCTs).[15] However, a clinical trial addresses a very closely defined question and, although it may be possible to codify good methods when making the very modest claim that 'A is more efficacious than B within the experiment', it may not be so easy when addressing the fundamentally more complex question, 'will A be more effective then B in routine clinical practice?'. There are some suggestions about how this complex issue of external validity should be addressed: 'A better approach than rigidly using the study's inclusion and exclusion criteria is to ask whether there is some compelling reason why the results should *not* be applied to the patient. A compelling reason usually won't be found . . . '[16] However, they hardly constitute a well-defined codification of appropriate methods. Even the guidelines and codes for establishing internal validity are not rigid or universal and in practice all commentators emphasise the importance of judgement and circumstances in monitoring the results of clinical trials.[17] In these circumstances, it would be foolish to expect any universal code to address problems that are even more complex and uncertain than the issue of the external validity of a clinical trial.

This does not mean that no guidance is possible and 'anything goes', or that it is impossible to choose between models on the basis of quality. However, it does mean that we should recognise the limitations of any guidelines or codes of practice and be conscious that the application of rigid methodological guidelines may do more harm than good for 3 reasons. First, the stringent application of guidelines will lead to the rejection of valuable models and will discourage the development of novel methods (we will become prisoners of our own orthodoxy). Secondly, it may distort research priorities by focusing on those issues where codified method and more precise answers may be available. These may be relatively simple questions where good evidence is already available. If we wish to address important, more complex and uncertain questions we must accept less precise answers, no universal methodological guidelines and the discomfort of exposing how little evidence is available. Finally, if the proposal of codes and guidelines is motivated by a desire to placate critics who mistakenly believe that universal codes for science are possible (and therefore decision analytic modelling is not scientific), then we will be conceding the whole argument to this naive view of science. The real argument against these critics would be to point out the complexities of the decision problems that must be addressed, the limited evidence available for key parameters and the fact that no other area of scientific activity can hold itself to the standards proposed.

2. A Framework for Quality Assessment

2.1 The Purpose of a Framework

Although there cannot and should not be an attempt to codify acceptable methods in decision analysis, there is a need to address more clearly the question of how to distinguish a good from a bad model. A small number of papers have considered how quality might be assessed in decision modelling.[11,18,19] The objective of the remainder of this paper is to extend and further develop this work as part of a continuing process of developing a frame-

work for quality assessment in decision analytic cost-effectiveness models and mapping-out what constitutes 'good practice'. This process is necessary if users of the outputs of models (e.g. decision makers at various levels in healthcare systems) are to have confidence in the results of studies and to value formal analysis as a means of reaching decisions in complex situations surrounded by uncertainty.

Most people familiar with cost-effectiveness models will probably be able to identify 'bad' models or inappropriate assumptions or treatment of data. However, describing what constitutes a 'good model' is a difficult undertaking. Such a description needs to be sufficiently generic to apply across diseases, interventions and model types, but to avoid being too general and hence of little value in any given context. Given that it is not possible to prescribe the appropriate approach at every step in building, populating and analysing a decision model, assessing the quality of a cost-effectiveness model requires the analyst to have adequately described their methods and to be able to provide clear, honest and transparent justification for the numerous components of their approach. Without this, systematic critical appraisal of decision models will be particularly difficult to undertake. Users can then make an informed judgement about whether they believe these answers to be justified and coherent.

To be able to agree a series of questions addressing different aspects of quality in cost-effectiveness models would be a valuable advance in the area: to analysts by indicating the sort of factors they should consider when developing a model and detailing the methods of their analysis in technical reports and journal papers; and to users of these outputs, including scientific referees, other researchers and decision makers, in terms of critical appraisal. Lists of questions about quality exist in other areas of evaluative health services research which are also seeking to be of value to analyst and appraiser alike. For example, the Consort guidelines is a series of questions about quality in RCTs.[15] Although their focus is the reporting of such studies, they can be seen as defining a set of dimensions along which the analyst has to justify their methods and results. The critical appraiser can then assess the extent to which the analyst's statements are reasonable. In the context of economic evaluation in general, the 'Drummond checklist' also serves this dual purpose, being of value to those designing and undertaking a study and to the various users or reviewers of these analyses.[20]

2.2 The Dimensions of Quality

Given their purpose and the myriad forms they can take, decision analytic cost-effectiveness models are inherently complex. Therefore, quality needs to be assessed at many different levels. The aim here is to begin the process of defining a taxonomy of quality in cost-effectiveness decision models. To be of value, this taxonomy requires 3 elements:

- A list of dimensions of quality for a model. This list of dimensions builds on earlier work[11,18,19] but cannot be exhaustive. This is because models involve numerous steps and components, many of which are specific to the disease area and/or technology under consideration. Furthermore, the methods of decision modelling are developing rapidly, such that the definition of 'a good model' will inevitably alter over time.
- Ideally it should be possible, for each dimension, to reach agreement on what constitutes good practice. For some dimensions, an unequivocal statement about the 'correct approach' may be possible; for others no rigid definition of good practice is possible, and defining quality is more about showing that the methods employed are reasonable and justifiable.
- To be a practical tool of critical appraisal for the users of models, the dimensions and descriptions of quality need to be distilled into a series of questions that an appraiser can apply to a given model. It should be emphasised that these questions focus on the modelling rather than the economic evaluation that the model will support. Therefore, the list would not seek to replace existing economic evaluation guidelines and checklists.[20,21]

The aim of the list should be to promote a way of thinking about modelling, with the onus on the analyst to identify the difficult methodological questions and justify their chosen solutions. Assessing and indicating quality is not an afterthought that, although inconvenient, is added to the end of the modelling process to satisfy those who review models and make decisions. Rather, it should be seen as an inherent part of the modelling process. It should be the way that analysts approach models during their inception, construction and presentation.

As a starting point in this taxonomy, it seems reasonable to group the dimensions of quality into 3 categories: the structure of the model, data and consistency. These are considered in turn below, together with some suggestions about what can be defined as good practice in each.

2.2.1 Structure

Determining structure is the key initial stage in the development of a model. At the outset, there should be a clear statement about the decision problem prompting the analysis. The analyst then translates this into an appropriately structured and specified model. This phase should be seen as entirely separate from the data population phase, with the theoretical model[11] reflecting the current level of understanding of the disease process rather than being constrained by the quality, level and availability of data. All clinically and economically relevant events should be included within the model and it '. . . should correspond as much as possible to the real life situation of a disease.'[22] However, clearly there must be some constraints placed upon a model to ensure that it is sufficiently tractable to make it comprehensible to users. The assumptions and compromises that are necessary to satisfy these constraints must be transparent and justified under a series of headings such as those detailed below.

Disease States

Cost-effectiveness decision models are typically developed around states. This is most clearly seen in state transition models, such as Markov models,[1,4] where discrete states should be clearly defined. However, simpler models such as decision trees can also be seen as having a range of states within them, defined by each chance node. The difference between these 2 general groups of model relates to the way they deal with time, with state transition models having advantages with chronic conditions where the disease process has a clear time dimension. The analyst should choose the type of model with reference to the time dependence of the disease events,[11] selecting the simplest format possible that adequately reflects the disease. The user should be able to satisfy themselves that the model type is applicable to the area under consideration and does not restrict the model.

How should these states be identified and defined? In principle, states should be chosen to reflect the underlying biological process of the disease in question rather than health service inputs. There are 2 related reasons for this. First, the nature of the disease should be the underlying process that drives all decisions about service delivery, so to structure a model around the latter is to get the decision-making process the wrong way around. Secondly, as discussed below, an important role of cost-effectiveness decision models is to estimate the costs and effects of a whole range of options regarding health service delivery. To develop a model on the basis of states that define different aspects of service, rather than of disease, loses sight of the purpose of the exercise and constrains what can be achieved.

For example, Davies and Drummond[23] developed a model of drug treatment in schizophrenia without considering the nature of the disease process. Others have started to develop economic models founded on an epidemiological model of the disease process which is based on theories of the pathophysiology of schizophrenia.[24,25] One advantage of this type of approach is that the core epidemiological model of the disease process should be generic and generalisable to all settings and treatment options. This requires a model to have a foundation in the clinical/biological theory of the pathophysiology of the disease. In the case of schizophrenia, there are a number of competing theories (2, 3, 4 or 5 dimension syndrome) which are more or less supported by evidence.[26]

Underlying the choice of disease states should be an accepted theory about how a disease can be categorised and its process characterised. These pathological theories are clearly evident in the clinical literature and invariably reflect what can be measured and/or patient-reported symptoms. For example, a range of clinical measures of pathology underlie theories of the process of ischaemic heart disease, including the extent of coronary stenosis or occlusion and the number and identification of arteries involved. In addition, the disease progress is understood in terms of symptoms of angina as reported by the patient.[27]

Once a theory of disease has been selected (which is not simply an empirical question), measures, definitions and the number of disease states can also be identified. These must be consistent with the theory of disease. For example, in developing a model of disease process in schizophrenia, 2 clinical scales are commonly used [Brief Psychiatric Rating Scale (BPRS) and Positive and Negative Symptom Assessment Scale (PANSS)], but the question remains of how to define disease states using these measures and how may disease states should be incorporated into the model. This will not be simply an empirical question but an iterative process of theoretical development (including clinical judgement) and empirical tests of whether the disease states have desirable properties.

Although theories of disease have to be the starting point in identifying appropriate model states, selecting a manageable number of states which, taken together, are sensitive to change in the underlying condition (as a result of intervention or natural history) has to be a balance between theory and empiricism. The analytical tools used in the development of descriptive health-related quality of life (HR-QOL) instruments may be of value in demonstrating appropriateness of disease state selection.[28] For example, cluster analysis has been used to help identify health states for HR-QOL instruments by clarifying the trade-off between the degree of variation explained and the number of health states.[29] Methods that have been used in the HR-QOL literature to assess the sensitivity to change of instruments, such as receiver operating curves,[30] may also be useful in identifying appropriate disease states for a model.

As for HR-QOL measures, where there is no universally accepted threshold that an instrument needs to pass before it is accepted as 'validated', the selection of disease states in a model should be shown to be sensitive to change and a reasonable reflection of what is understood about the levels and progression of a disease. Even in cluster analysis, the efficient number of means (states) will always be a clinical (normative) rather than empirical (positive) issue. This is because, even if maximising explanatory power was the only objective (which it is not), then we would still need to decide the cost (in terms of the number of states) per improvement in explanatory power we are willing to pay. This suggests that the number and definition of states will continue to be a trade-off between descriptive realism and analytic simplicity, which is not simply an empirical question. Although analytical methods such as cluster analysis may be of value in demonstrating an appropriate choice, more methods research is required in this area and, for the time being at least, a more descriptive justification may have to suffice.

In principle, the structure of a model, including the selection of disease states, should not be driven by the availability of data with which to populate the model. However, the process of 'pruning' models to fit in with what data are available is commonplace in decision analytic cost-effectiveness models. For example, in an economic model of alternative treatments for toenail onychomycosis,[31] the authors decided to exclude adverse events because 'the reporting of adverse events was inconsistent with respect to description, causality, assessment methods, extent of reporting, and provision of numerator and denominator data'. However, the nature of some of the pharmaceutical treatments indicates (according to the Committee on the Safety of Medicines) a very small but measurable possibility of liver damage as well as a number of other more probable but less serious events. Because of the nature of its action, one of the treatments evaluated

in the model carries no theoretical risk. Excluding adverse events from the model on the basis of poor data can only bias the results away from this treatment. In fact, some adverse event data are available from the clinical trials that were used to estimate effectiveness. The fact that these data are incomplete and difficult to interpret (and may need to be supplemented with expert judgement) does not seem to justify ignoring this potentially important clinical event entirely.

However, at other times data will drive model structure. For example, we often do expect differences in outcome and cost by age, gender and comorbidities, but we seldom have enough data for any meaningful subgroup analysis. In these circumstances, we often focus on an average patient or a homogeneous group with clearly defined characteristics (often selected to match trial protocols). It seems to be a question of a balance between theory and evidence that is difficult to codify. The best we can say is that decisions should be clearly justified with an honest statement of the data and prior information that was available. It would be very dangerous to specify the quality of data that is acceptable, since this would lead to the arbitrary pruning of important clinical events from models simply because data of a certain quality are not yet available.

If expert opinion is the only means of estimating a parameter, a theoretically valid structure will facilitate an analysis of how sensitive the results are to changes in a particular parameter and of the benefit of generating additional information about the value of the parameter. This concept of quality in models is apparently not shared by others, however. Sonnenberg et al.[11] state that the 'detail of the model should be no more than can be supported by the available data', suggesting that the model should be adapted on the basis of the data available. If a general agreement is to be reached on what quality in modelling is to mean, these differences in viewpoint must be ironed out.

Options

Assuming that the purpose of decision modelling is to help to establish (economically) optimal solutions to particular decision problems, the options and strategies incorporated within the model should not be limited by the constraints of currently accepted clinical practice. It is a tautology to use a model founded in clinical practice to inform optimal clinical management; such an approach could bias the results of a model and limit its usefulness, especially over time. Hence the analyst should encourage 'thinking outside the box' when seeking clinical opinion concerning options for inclusion in the model. This is possibly best handled when a clinically uninformed analyst questions the logic and basis for the strategies identified by the experts and prompts thinking about unconventional approaches. The inclusion of extreme strategies of clinical management is particularly useful to serve as an anchor for other strategies and to assess consistency (see section 2.2.3).[11] However, the model cannot conceivably contain every possible option, so the analyst will need to justify the reasonableness of decisions about inclusion given the context and perspective of the decision problem.

Time Horizon

Although partly conditioned by the perspective of the study, the principles behind the selection of an appropriate time horizon for a cost-effectiveness analysis are reasonably clear:[3,20] analysis should continue until the differences between options in terms of costs and effects (and hence cost effectiveness) are stable and unlikely to alter significantly. However, in a similar way to clinical trials, identifying this point of stability in cost and effect differences is an empirical output of a model rather than something that can be stated before the analysis is undertaken.

For this reason, many longitudinal cost-effectiveness models relating to chronic conditions choose to model their cohort until the vast majority have entered the death state. Examples exist in peripheral vascular disease[32] and hormone replacement therapy.[33] For diseases (and treatment implications) that are very unlikely to exist over a lifetime, the choice of a shorter time horizon can be justified. For example, menorrhagia (heavy menstrual bleeding) is a condition that will disappear naturally at

the menopause. Women tend to present with the condition in their early 40s[34] and the menopause takes place in the early 50s,[35] suggesting a time horizon of about a decade. However, a 10-year time horizon would only be a reasonable justification if there were no differences between treatment options in terms of mortality (if there was, a time horizon of longer than 10 years would be necessary to estimate differential life-years) or in the incidence of other conditions that have excess mortality risks and implications for HR-QOL. In the menorrhagia example, both of these situations may exist: a difference in mortality because some treatments are surgical whilst others are pharmaceutical, and a potential difference in uterine cancer rates because one of the treatment options is hysterectomy. This example indicates how difficult *priori* decisions about time horizon can be.

Cycle Length

Within the state transition decision models that predominate in the cost-effectiveness literature, the selection of the appropriate cycle length is an important feature of the model's structure. As for the selection of appropriate disease states, the choice of cycle length should be driven by what is known about the underlying disease process. More specifically, the length of a cycle should be the minimum interval over which pathology and/or symptoms in patients is expected to alter. Hence, to indicate to users that a reasonable decision has been taken regarding cycle length, the analyst should justify the selection in terms of the disease process. This contrasts with a choice which is based on the intervals between clinical reviews, which is inappropriate because finding the optimal period of follow-up/review should, ideally, be one aspect of the options and strategies being evaluated by the model. Similarly, cycle length should not be determined by the availability of data with which to populate the model. At a technical level, this is true because methods exist to translate rates over any interval available in the literature to model probabilities for a given cycle length.[36] More fundamentally, as detailed above, there are strong reasons not to let data availability determine model structure; instead, knowledge about the underlying disease should be the driving force.

2.2.2 Data

As indicated in section 2.2.1 the process of populating the model with data should be seen as separate from decisions about model structure. There are 2 main issues to consider when approaching an assessment of the quality of the data within a model. The first deals with the quality of the data identified for use within the model; the second deals with the methods used to incorporate the data into the model.

Data Identification

The purpose of decision analytic cost-effectiveness models is to indicate the most cost-effective management option on the basis of best currently available data. Therefore, it is inappropriate to criticise a model because of a dearth of data or because existing data fall short of the ideal in terms of the scientific rigour of the methods that produced them.[37] An important potential function of cost-effectiveness models is to facilitate formal analysis of the value of additional research. That is, they can be used to assess the 'cost' (in terms of resources used or health benefits forgone) of uncertainty in a given decision context and hence the value of perfect information and the optimal design of future primary research.[38] However, this crucial role for models in determining what data should be available at time t+1 does not obviate the need to reach decisions on the basis of best available data at time t.

So the 'best' available data may fall some way short of the ideal, but careful consideration needs to be given to what 'best' means, even in the context of existing evidence. A number of authors have discussed the identification and use of data when populating a model without explicitly defining what constitutes 'best' available data.[21] An important rationale for the development of methods for systematic literature review was that even existing evidence may require considerable time and effort to identify. This resource intensive process is, of course, research in its own right and should be subject to the same sort of appraisal process as new primary research. In other words, 'best available' data

should in fact be referred to as 'optimal available' data, as it is an empirical question whether acquiring all existing evidence in order to determine the 'best' is a good use of resources.

How can an analyst developing a model (and a user appraising its value) be clear regarding whether parameter values are based on optimal available data? In part, this depends on how many iterations of the model exist.[38] If the model is the latest in a series the analyst has developed to address a particular decision problem, with each iteration including new primary and/or secondary evidence deemed cost-effective to generate by an earlier iteration, the justification of the data inputs selected should emerge from a careful description of the methods and results of the overall process.

In current practice, most model-based cost-effectiveness analyses represent the first (and often only) application of a model to a decision problem. In this situation it may be more difficult for an analyst to show – and a user to be confident – that the model has incorporated optimum available data. A pragmatic approach to indicating and assessing quality of data inputs in this context may be to make clear that all sources of information that are available at relatively low cost, in terms of researcher time, have been searched for the most appropriate parameter values. Although this may be easier to stipulate for clinical and epidemiological information than resource use, cost and utility data, these sources are likely to include the NHS Centre for Reviews and Dissemination's Database of Abstracts of Reviews of Effectiveness (DARE), the Cochrane Library and Medline. In effect, this is laying down some minimal (but low-cost) thresholds that the analyst should report in how they have identified quality.

It is often the case that no data can be identified upon which to base a particular parameter. In these circumstances it is appropriate to use informed opinion to estimate parameter values, but eliciting expert opinion can take many forms.[19] Although there is little guidance in the methods literature, some reasonable aspects of good practice in this area may include:

- a clear statement of the inclusion criteria and sample size for the survey of experts
- a careful description of what the experts were asked and the format of the questions (e.g. questionnaire or interview, group- or individually-based)
- the avoidance of an attempt to forge a group consensus, the results of which may be more a reflection of the personalities in the group than genuine agreement; the search for consensus also ignores the likelihood of genuine variation in individual experience and hence opinion.

Data Incorporation

In addition to identifying optimal data for a decision model, quality assessment should also be concerned with the methods used to translate information in the literature into an appropriate form for incorporation into the model. In some ways this is the most difficult area of quality assessment in modelling about which to be prescriptive. This is largely because the development of most models requires the interpretation, manipulation and translation of many different data items, and the extent to which each of these is appropriate will invariably depend on the context. In general, the process of data incorporation should follow the accepted methods of epidemiology and statistics (including data synthesis methods such as meta-analysis) but, beyond this, the incorporation of each item of data should be treated on its merits.

However, there are some common steps concerning incorporation of data that occur in a large proportion of decision models and where there is a high risk of using incorrect methods. It may be argued that the appropriateness of these steps should be subject to particularly careful scrutiny on the part of the analyst and user. A (far from exhaustive) list of common steps is considered below.

Handling Uncertainty

Uncertainty is ubiquitous in all cost-effectiveness models. Indeed, part of the rationale for the use of models is that they enable decisions to be reached through the systematic handling of uncertainty. The dimensions of uncertainty in economic evaluation have been described elsewhere.[39] In the con-

text of decision models, it is helpful to distinguish between uncertainty resulting from the process of sampling from a population and variability due to heterogeneity.[40] For example, the incidence of adverse effects after a particular treatment will typically be based on a sample from a population of patients and a distribution of incidence rates can be defined that allows for the sampling uncertainty. On the other hand, heterogeneity exists when the samples are actually being drawn from different populations (e.g. adverse events might be quite different in those aged below 40 years than in older individuals). In this case, it would be inappropriate to characterise this variability in terms of a single distribution of incidence rates and the analysis should be undertaken for each subgroup of patients.

Unit costs can be seen as a special case of heterogeneity. For example, the unit cost of a day in hospital may appear to be uncertain from the perspective of the analyst but, for a specific decision maker, it should be known at any given point of time. One way of reflecting uncertainty in decision models is to make them stochastic; that is, instead of incorporating point estimates of parameters, a distribution of values can be incorporated. Although there are some good reasons for suggesting that a stochastic approach to modelling should be considered good practice as a means of dealing with uncertainty, there is no clear consensus on this.[3,19] However, if a stochastic analysis is undertaken, the analyst should be clear about, and justify, their choice of statistical distributions to model the uncertainty.[41] Furthermore, it is important to be clear that, given the purpose of the model to estimate mean costs and effects of options, it is second order uncertainty that is being reflected in the distributions – that is, uncertainty in probabilities and in mean values.[42]

Probabilities

Care should be taken when translating data about the rates of particular events into transition probabilities for a given cycle length in order to correctly incorporate the effect of time in the estimate. There has been some confusion in the literature about the appropriate methods for affecting this translation, although a generally accepted approach now seems to have emerged.[36]

Half-Cycle Correction

Any time-related estimates from state transition models may be affected by the implicit assumption that all events occur at the start or end of the cycle. An unbiased estimate of these outputs should assume that events occur mid-way through a cycle, and the half-cycle correction is a way of achieving this adjustment.[1]

2.2.3 Consistency

Consistency describes whether the model was correctly doing what it set out to do. It is useful here to distinguish between internal and external consistency. To ensure internal consistency, the model should be checked and tested by the analyst during the modelling process to trap any errors relating to data incorporation and modelling syntax. A range of tests can be used and providing details of these may be an important way in which the analyst can indicate quality to a user. Simple techniques to test for internal consistency include movement of results in sensitivity analysis (including analysis of extreme values) follow *a priori* expectations, and the construction of the model, using the same data, by another analyst and/or using a different software package to check that the results are the same.

External consistency relates to whether the results of the model are consistent with information contained in relevant primary research studies. One approach to this is to look at intermediate outputs from the model, such as time in particular states, under one or more options and to compare this with the results of longitudinal studies in the published literature. For example, in a cost-effectiveness model of different treatments for peripheral vascular occlusions,[32] where quality-adjusted life-years (QALYs) was the ultimate measure of effectiveness, 5-, 10- and 15-year survival rates predicted by the model were compared with the results of primary research in the literature. A similar approach was adopted in a model of hormone replacement therapy,[43] where

Table I. Suggested framework for assessing the quality of decision analytic cost-effectiveness models

Dimensions of quality	Attributes of 'good practice'	Questions for critical appraisal
Structure	The structure of the model should be consistent with the stated decision problem Structure should be dictated by a theory of disease, not by data availability	Is there a clear statement of the decision problem, the context and the perspective? Is a theory of the underlying disease detailed? Are the assumptions involved in the model clearly specified? Are they justified? Are the implications of relaxing these assumptions described?
Disease states	Analysts should choose the simplest model type that adequately reflects the time dependence of the disease process States should be chosen to reflect the underlying biological process of the disease in question, and the impact of interventions, rather than health service inputs. Models need to reflect accepted clinical classifications of disease The number of states should be manageable but sensitive to change in the underlying condition States should not be omitted on the basis of lack of data	Is the chosen model type appropriate for the time dimension of the disease process? Is a justification of the choice of states within the model provided? If so, does this accord with a theory of the disease process? Is any empirical evidence provided on the suitability of the states (e.g. sensitivity to change in the underlying disease)? Have any important states been omitted from the model?
Options	Options and strategies should not be limited by the constraints of currently accepted clinical practice A balance needs to be struck between including a full range of logical and feasible options and keeping the decision problem manageable	Is there a clear statement of the options being evaluated? Do these appear to cover the full range of logical and feasible options?
Time horizon	The time horizon of a study should be sufficient to indicate when cost and effect differences between options are stable However, this is an output of a model rather than something knowable *ex ante* Lifetime time horizons will be appropriate in many longitudinal models, but shorter time dimensions can be justified according to understanding of the disease process and the effect of interventions	Is the time horizon of the analysis stated? If so, is this justified in terms of the underlying disease and the effect of interventions?
Cycle length (if relevant)	The length of a cycle should be the minimum interval over which pathology and/or symptoms in patients is expected to alter To indicate to users that a reasonable decision has been taken regarding cycle length, the analyst should justify the selection in terms of the disease process	If relevant, is the cycle length used in the model stated? Is justification offered on the choice of cycle length? If so, does the justification relate to the disease process?
Data identification	It is inappropriate to criticise a model because of a dearth of data or because existing data fall short of the ideal in terms of the scientific rigour of the methods that produced them 'Best available' data should to referred to as 'optimal available' data as it is an empirical question whether acquiring all existing evidence in order to determine the 'best' is a good use of resources Models can be used to undertake formal value of information analysis to determine the optimal data to incorporate For the first iteration of a model, the analyst should make clear that all sources of information that are available at relatively low cost in terms of researcher time have been searched for the most appropriate parameter values In the context of no data being identified for a particular parameter, the methods used to elicit expert opinion should be fully detailed	Are the sources of parameter values in the model clearly stated? Is reasonable empirical justification, from earlier iterations of the model, offered that these data are optimal? For the first iteration of a model, has satisfactory justification been offered that data are based on a search of all the low-cost data sources (e.g. Medline, DARE, Cochrane Library)? Are ranges specified for parameters? Is there evidence to suggest selective use of data? If some parameter values are based on elicitation of expert opinion, have the methods used for this purpose been adequately described (e.g. inclusion criteria, sample size, elicitation methods)? Are the claims made about the model results tempered by the limitations of the data?

Table I. Contd

Dimensions of quality	Attributes of 'good practice'	Questions for critical appraisal
Data incorporation	In general, the process of data incorporation should follow the accepted methods of epidemiology and statistics (including data synthesis methods such as meta-analysis) but, beyond this, the incorporation of each item of data should be treated on its merits Some aspects of data incorporation are particularly important. For example, the different sources of uncertainty should be distinguished, especially sample uncertainty versus heterogeneity; stochastic analyses should set distributions for parameter values to reflect second order uncertainty; interval rates from the literature should be translated into transition probabilities using the appropriate formula; time estimates from the model should include a half-cycle correction	For each parameter value, is there a clear and reasonable justification of how data have been incorporated into the model? Has a stochastic analysis been undertaken? If so, do the distributions in parameter values reflect second order uncertainty? Have appropriate distributions been selected for each parameter? Have interval rates been translated into transition probabilities using the appropriate formula? If appropriate, has a half-cycle correction been applied to adjust time-related estimates in the model?
Internal consistency	To ensure internal consistency, the model should be checked and tested by the analyst during the modelling process to trap any errors relating to data incorporation and modelling syntax. A range of tests can be used and providing details of these may be an important way in which the analyst can indicate quality to a user (e.g. movement of results in sensitivity analysis follow *a priori* expectations; the construction of the model, using the same data, by another analyst and/or using a different software package to check that the results are the same)	Is there a statement about the tests of internal consistency that were undertaken?
External consistency	External consistency relates to whether the results of the model are consistent with information contained in relevant primary research studies. One approach to this is to look at intermediate outputs from the model, such as time in particular states, under one or more options and compare this with the results of longitudinal studies in the published literature	Are any relevant studies and/or models identified by the analyst for the purpose of comparison? Have any comparisons of the outputs of the model with independent external sources been reported? If so, are the conclusions justified? Have discrepancies been investigated and explained?

DARE = Database of Abstracts of Reviews of Effectiveness.

the model's estimate of age at first hip fracture was compared with estimates from a cohort study.

This is a useful way of increasing confidence in a model. However, some caveats are needed regarding this sort of approach to assessing external consistency. First, the external sources of output estimates should not have been used to estimate the parameter values in the model: to have been used for this purpose would clearly limit its value for assessing external consistency. Secondly, it is important to be clear that an external source of output estimates is not a gold standard when compared with model outputs. This is because the purpose of the model is to assist in decision making, which is unlikely to have been the purpose of the longitudinal study. Therefore, the term 'validation' should perhaps be considered too strong and the results of the comparison need to be interpreted carefully. The absence of convergence in their estimates should cause the analyst to reassess the structure, data and internal consistency of the model, but it need not mean that the model should be abandoned. Tosteson et al.[43] undertook sensitivity analysis to examine the association between bone mineral density and fracture risk when the relative risks for hip fracture estimates from their model did not correspond to those from a prospective study. Similarly, the presence of convergence of output estimates does not ensure a quality model on all the dimensions considered here.

Table II. Illustrative application. The model is that of Briggs et al.;[44] the perspective is that of a reviewer of the model

Dimensions of quality	Questions for critical appraisal	Consideration of the 'hip model'
Structure	Is there a clear statement of the decision problem, the context and the perspective?	The model aims to determine the reduction in revision rates required for new design prostheses to be considered cost effective from the perspective of the NHS
	Is a theory of the underlying disease detailed?	Background information concerning total hip replacement prognosis is provided
	Are the assumptions involved in the model clearly specified? Are they justified? Are the implications of relaxing these assumptions described?	A major assumption within the modelling is the linear failure risk based upon time since THR. This assumption is justified on the basis that it produces similar results to a published study for prosthesis survival. However, there does not appear to be any biological, clinical or technical theory behind this assumption. In fact, technical theory suggests that the age at THR would be an important factor in failure risk. However, this is not taken into account and there is no examination of the impact of uncertainty surrounding this relationship
Disease states	Is the chosen model type appropriate for the time dimension of the disease process?	A Markov model with 4 distinct states (successful primary THR; revision THR; successful revision THR; death) is used. The Markov model involves an assumption of 'no memory' - past states do not affect the probabilities of future states. It may be inappropriate for subsequent revisions to be modelled in this way if, as has been suggested, revisions of revisions have an increasing failure rate
	Is a justification of the choice of states within the model provided? If so, does this accord with a theory of the disease process?	Indirectly. The health states seem to adequately describe the main process of the disease
	Is any empirical evidence provided on the suitability of the states (e.g. sensitivity to change in the underlying disease)?	No evidence is provided. Gradual deterioration of a patient's health as a hip fails is taken into account in the year prior to a move to the revision THR state, and recovery from surgery is taken into account in the first cycle following a move into either of the successful surgery states
	Have any important states been omitted from the model?	Complications arising as a result of surgery have been omitted from the model and may be important: 1-5% of patients are reported to have serious complications after surgery and the number is higher for revision surgery than primary surgery
Options	Is there a clear statement of the options being evaluated?	The model evaluates total hip replacement using established and new design prostheses
	Do these appear to cover the full range of logical and feasible options?	The specific nature of the model is such that the examination of 2 main groups of prostheses is sufficient
Time horizon	Is the time horizon of the analysis stated?	Yes – 60 years from primary THR surgery
	If so, is this justified in terms of the underlying disease and the effect of interventions?	This time horizon is justified on the basis that 98% of patients, in the youngest cohort, are dead after 60 years and therefore differences in cost and effect will be minimal
Cycle length	If relevant, is the cycle length used in the model stated?	Cycle length – 1 year
	Is justification offered on the choice of cycle length?	No justification is given for the cycle length
	If so, does the justification relate to the disease process?	Not applicable
Data identification	Are the sources of parameter values in the model clearly stated?	Resource use estimates are provided by the NOC and clinical opinion. Cost estimates are provided by the NOC (overheads) and published list prices (materials and salaries). Patient benefits are determined according to pain levels and associated utility values are estimated from published studies. Transition probabilities are from published studies and statistics (operative and underlying mortality), clinical opinion (risk of re-revision) and models to extrapolate available data (risk of failure)

Table II. Contd

Dimensions of quality	Questions for critical appraisal	Consideration of the 'hip model'
	Is reasonable empirical justification, from earlier iterations of the model, offered that these data are optimal?	No previous analysis is referred to
	For the first iteration of a model, has satisfactory justification been offered that data are based on a search of all the low-cost data sources (e.g. Medline, DARE, Cochrane Library)?	'There is a dearth of data of long-term follow-up on effectiveness of hip prostheses'. However, studies have been identified and used for parameter estimation. No details are given on how this identification was undertaken
	Are ranges specified for parameters?	Not explicit
	Is there evidence to suggest selective use of data?	No
	If some parameter values are based on elicitation of expert opinion, have the methods used for this purpose been adequately described (e.g. inclusion criteria, sample size, elicitation methods)?	Clinical judgement is used to determine some aspects of resource use (x-rays and outpatients visits). There is no discussion of the methods or experts used
	Are the claims made about the model results tempered by the limitations of the data?	The authors explicitly identify the lack of data available to estimate the parameters of the model and suggest the model represents 'preliminary economic modelling'. The results of the model are framed in this context and refer mainly to the worth of further data acquisition (through clinical trials) given the results of the model. There is some discussion about the implications of the results for service provision decisions
Data incorporation	For each parameter value, is there a clear and reasonable justification of how data have been incorporated into the model?	The model is well detailed and it is clear how data have been incorporated within the model
	Has a stochastic analysis been undertaken? If so, do the distributions in parameter values reflect second order uncertainty rather than heterogeneity? Have appropriate distributions been selected for each parameter?	No stochastic analysis has been undertaken. Uncertainty has been examined through the use of 1-way sensitivity analysis for the main parameters of the model and presented as an elasticity estimate. However, this information is of limited use to the user without information concerning the range for each parameter. There was no examination of the uncertainty surrounding the functional relationship of the hazard rate
	Have interval rates been translated into transition probabilities using the appropriate formula?	Unknown
	If appropriate, has a half-cycle correction been applied to adjust time-related estimates in the model?	A half-cycle correction has been applied to the model
Internal consistency	Is there a statement about the tests of internal consistency that were undertaken?	There is no discussion of the tests undertaken to check the internal consistency of the model
External consistency	Are any relevant studies and/or models identified by the analyst for the purpose of comparison?	Studies are identified to compare the intermediate outcome of prosthesis failure. Other economic evaluations of THR are identified although not used for formal comparison purposes.
	Have any comparisons of the outputs of the model with independent external sources been reported?	The prosthesis survival curves predicted and used within the model were compared with actual curves from the Swedish National Hip Arthroplasty Registrar
	If so, are the conclusions justified? Have discrepancies been investigated and explained?	Conclude that 'On balance, the hazard function predictions...do not seem inconsistent ...'. The use of different categories within the data makes direct comparison difficult and discrepancies are assumed to be due to this

DARE = Database of Abstracts of Reviews of Effectiveness; **NHS** = UK National Health Service; **NOC** = Nuffield Orthopaedic Centre, Oxford, England; **THR** = total hip replacement.

2.3 Moving Towards a List of Questions on Quality

Section 2.2 has sought to make some suggestions about the various dimensions of quality in a decision analytic cost-effectiveness model. It should be reiterated that these dimensions are not exhaustive and will no doubt alter in time as the methods of decision modelling develop. For each dimension, an attempt has been made to describe what might be considered good practice, although this is often couched in general terms given that the appropriateness of methods often cannot be firmly assessed outside of the specific disease and treatment context. However, to be of value to users of such models, the dimensions and descriptions of quality should be expressed in terms of a series of questions that can be asked of a model by the analyst in developing and analysing it and by the range of possible users in appraising and interpreting it.

Alongside each dimension of quality identified in section 2.2, and a summary of the main tenets of what might be considered good practice for each, table I suggests a series of questions that might be used in the critical appraisal of cost-effectiveness models. By way of an illustration of how these questions might be used in practice, table II applies them to a specific model that evaluates the cost effectiveness of alternative hip prostheses.[44] Given that they focus on the modelling process, these suggested questions will not replace checklists designed to assess the quality of the economic analysis.[21]

3. Conclusions

Any progress in developing a framework to assess the quality of decision analytic cost-effectiveness models requires some agreement on the purpose of this type of model. It is a key argument in this paper that models are concerned with rational decision making on the basis of existing information rather than with generating new information and are, therefore, not substitutes for clinical trials. Hence, it is inappropriate to compare the results of decision models with those subsequently generated by clinical trials and to describe this as a form of validity test. In principle, it is possible directly to test the 'validity' of decision models (in terms of scientific method, to falsify a model) by comparing its value as a mode of decision making with other modes such as implicit decision making, but this is not feasible or indeed desirable for each and every decision model.

However, there remains much to say about quality in modelling, as it has to be possible to distinguish good from bad analysis. Although it will never be practical (or indeed desirable) to codify every step in the development, analysis and reporting of decision models, it is important to move towards an agreed framework for quality assessment. This will be a major development if it can influence the way analysts think about model building and offer the full range of users a systematic means of appraising these methods. Given the complexity of models and the specificity of the structure, data and assumptions with respect to a given decision problem, the focus of any framework is likely to be the analyst providing an explicit and comprehensive justification of their methods, with users encouraged to make an informed judgement about whether they believe these methods to be justified and coherent.

The framework suggested in this paper is an early step in the process of developing a systematic approach to quality assessment in decision modelling, and should be considered with other contributions in this area.[11,18,19] Its purpose is to stimulate discussion and contributions from others in the hope of continued development and refinement.

References

1. Sonnenberg FA, Beck JR. Markov models in medical decision making. Med Decis Making 1993; 13: 322-38
2. Keeler E. Decision trees and Markov models in cost-effectiveness research. In: Sloan F, editor. Valuing health care: costs, benefits and effectiveness of pharmaceuticals and other medical technologies. Cambridge: Cambridge University Press, 1995: 185-205
3. Gold MR, Siegel JE, Russell LB, et al. Cost-effectiveness analysis in health and medicine. New York (NY): Oxford University Press, 1996
4. Briggs A, Sculpher MJ. An introduction to Markov modelling for economic evaluation. Pharmacoeconomics 1998; 13: 397-409
5. Sheldon TA. Problems of using modelling in the economic evaluation of health care. Health Econ 1996; 5: 1-11

6. Buxton MJ, Drummond MF, Van Hout BA, et al. Modelling in economic evaluation: an unavoidable fact of life. Health Econ 1997; 6: 217-27
7. Popper KR. The logic of scientific discovery. London: Hitchinson, 1972
8. Lakatos I. Falsification and the methodology of scientific research programmes. In: Lakatos I, Musgrave, A, editors. Criticism and the growth of knowledge. Cambridge: Cambridge University Press, 1970: 91-195
9. Manning R, Claxton K. Experimental and econometric solutions to selection bias. The International Health Economics Association Meeting, 1999 Jun 6-9; Rotterdam
10. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. J Chronic Dis 1967; 20: 637-48
11. Sonnenberg FA, Roberts MS, Tsevat J. Toward a peer review process for medical decision analysis models. Med Care 1994; 32 Suppl.: JS52-64
12. Russell B. My philosophical development. London: Routledge, 1959
13. Sokal A, Bricmont J. Intellectual impostures. London: Profile Books, 1998
14. Feyerabend P. Against method. London: New Left Books, 1975
15. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials. JAMA 1996; 276: 637-9
16. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature: II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? JAMA 1994; 271: 59-63
17. Pocock SJ. When to stop a clinical trial. BMJ 1992; 305: 235-40
18. Eddy DM. Technology assessment: the role of mathematical modeling. In: Mosteller F, editor. Assessing medical technologies. Washington, DC: National Academy Press, 1985: 144-53
19. Halpern MT, Luce BR, Brown RE, et al. Health and economic outcomes modeling practices: a suggested framework. Value Health 1998; 1: 131-47
20. Drummond MF, O'Brien BJ, Stoddart GL, et al. Methods for the economic evaluation of health care programmes. New York (NY): Oxford University Press, 1997
21. Drummond M, Jefferson T. Guidelines for authors and peer reviewers of economic submissions to the BMJ. BMJ 1996; 313: 275-83
22. Nuijten MJC, Starzewski J. Applications of modelling studies. Pharmacoeconomics 1998; 13: 289-91
23. Davies LM, Drummond MF. Assessment of costs and benefits of drug therapy for treatment-resistant schizophrenia in the United Kingdom. Br J Psychiatry 1993; 162: 38-42
24. Palmer CS, Revicki DA, Genduso LA, et al. A cost-effectiveness clinical decision analysis model for schizophrenia. Am J Manag Care 1998; 4: 345-55
25. Mohr P, Neumann P, Claxton K. Letter to the editor. Am J Manag Care 1998; 4: 1200-1
26. Peralta V, Cuesta MJ, de Leon J. An empirical analysis of latent structures underlying schizophrenic symptoms: a four-syndrome model. Biol Psychiatry 1994; 36: 726-36
27. Campeau L. Grading of angina pectoris. Circulation 1976; 54: 522-3
28. Patrick DL, Erickson P. Health status and health policy: allocating resources to health care. New York (NY): Oxford University Press, 1993
29. Sugar CA, Sturm R, Lee TT, et al. Empirically defined health states for depression from the SF-12. Health Serv Res 1998; 33: 911-28
30. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis 1986; 39: 897-906
31. Van Doorslaer EKA. Economic evaluation of antifungal agents in the treatment of toenail onychomycosis. Dermatology 1996; 193: 239-44
32. Sculpher M, Michaels J, McKenna M, et al. A cost-utility analysis of laser-assisted angioplasty for peripheral arterial occlusions. Int J Technol Assess Health Care 1996; 12: 104-25
33. Daly E, Vessey MP, Barlow D, et al. Hormone replacement therapy in a risk-benefit perspective. Maturitas 1996; 23: 247-57
34. Coulter A, Bradlow J, Agass M, et al. Outcomes of referrals to gynaecology outpatient clinics for menstrual problems: an audit of general practice records. Br J Obstet Gynaecol 1991; 98: 789-96
35. Luoto R, Kaprio J, Uutela A. Age at natural menopause and sociodemographic status in Finland. Am J Epidemiol 1994; 139: 64-76
36. Miller DK, Homan SM. Determining transition probabilities: confusion and suggestions. Med Decis Making 1994; 14: 52-8
37. Dowie J. Evidence-based, cost-effective and preference-driven medicine: decision analysis based medical decision making is the pre-requisite. J Health Serv Res Policy 1996; 1: 104-13
38. Fenwick E, Claxton K, Sculpher M, et al. Improving the efficiency and relevance of health technology assessment: the role of iterative decision analytic modelling. Centre for Health Economics (CHE) discussion paper. York: CHE, University of York. In press
39. Briggs A, Sculpher M, Buxton M. Uncertainty in the economic evaluation of health care technologies: the role of sensitivity analysis. Health Econ 1994; 3: 95-104
40. Thompson KM, Graham JD. Going beyond the single number: using probabilistic risk assessment to improve risk management. Hum Ecological Risk Assess 1996; 24: 1008-34
41. Thompson KM, Burmaster DE, Crouch EAC. Monte Carlo techniques for quantitative uncertainty analysis in public health risk assessments. Risk Anal 1992; 12: 53-63
42. Stinnett A, Paltiel D. Estimating CE ratios under second-order uncertainty: the mean ratio versus the ratio of the means. Med Decis Making 1997; 17: 483-9
43. Tosteson ANA, Rosenthal DI, Melton J, et al. Cost-effectiveness of screening perimenopausal white women for osteoporosis: bone densitometry and hormone replacement therapy. Ann Intern Med 1990; 113: 594-603
44. Briggs A, Sculpher M, Britton A, et al. The costs and benefits of primary total hip replacement: how likely are new prostheses to be cost-effective? Int J Technol Assess Health Care 1998; 14 (4): 743-61

Correspondence and offprints: Dr *Mark Sculpher*, Centre for Health Economics, University of York, Heslington, York YO10 5DD, England.