

Testing the Validity of Cost-Effectiveness Models

Chris McCabe and Simon Dixon

Sheffield Health Economics Group, School of Health and Related Research, University of Sheffield, Sheffield, England

Contents

Abstract	501
1. Review of Previous Attempts to Establish Validity	502
1.1 The Necessity for Modelling	503
1.2 Validating Cost-Effectiveness Models	503
1.3 Summary	507
2. A Framework for Assessing Validity	508
2.1 Structure of the Model	508
2.2 Inputs to the Model	509
2.3 Results of the Model	509
2.4 Value of the Model to the Decision Maker	510
3. The Broader Model Evaluation Process	511
4. Discussion	511
5. Conclusions	512

Abstract

A growing body of recent work has identified several problems with economic evaluations undertaken alongside controlled trials that can have potentially serious impacts on the ability of decision makers to draw valid conclusions. At the same time, the use of cost-effectiveness models has been drawn into question, due to the alleged arbitrary nature of their construction. This has led researchers to try and identify ways of improving the quality of cost-effectiveness models through identifying 'best practice', producing guidelines for peer review and identifying tests of validity.

This paper investigates the issue of testing the validity of cost-effectiveness models or, perhaps more appropriately, whether it is possible to objectively measure the quality of a cost-effectiveness model. A review of the literature shows that there is much confusion over the different aspects of modelling that should be assessed in respect to model quality, and how this should be done.

We develop a framework for assessing model quality in terms of: (i) the structure of the model; (ii) the inputs to the model; (iii) the results of the model; and (iv) the value of the model to the decision maker. Quality assessment is investigated within this framework, and it is argued that it is doubtful that a set of objective tests of validity will ever be produced, or indeed that such an approach would be desirable. The lack of any clearly definable and objective tests of validity means that the other parts of the evaluation process need to be given greater

emphasis. Quality assurance forms a small part of a broader process and is best implemented in the form of good practice guidelines. A set of key guidelines are presented.

The focus of this paper is the issue of quality assessment and elimination of bias in model-based economic evaluations. In most areas of economics, where modelling is typically undertaken using ordinary least squares regression, validity can be tested in 2 broad ways: first, by the examination of residuals within sample (i.e. tests for non-normality, heteroskedasticity and autocorrelation) and, secondly, by the prediction of data points out of sample. This is made possible by the fact that the model is estimated from both explanatory and dependent data – in other words, a gold standard (i.e. observations of what actually happens) is available. However, such tests are generally not possible for cost-effectiveness models as, typically, the model's aim is to estimate the dependent data (e.g. costs) in the absence of observable dependent data.¹ There is therefore a fundamental question as to whether the validity of such models can actually be tested.

In the absence of the standard approaches to model validation, researchers resort to assessing whether models possess certain desirable properties, such as transparency or robustness. However, there is no guarantee that transparency and robustness will lead to a 'good' model nor, in fact, that a transparent and robust model will be better than an opaque and sensitive model.

In this sense, the process is probably better described as 'assessing the existence of desirable properties' as opposed to 'testing the validity of the outcomes'. However, for the purposes of this paper we shall continue to refer to 'validity' and ask the reader to be tolerant of the limitations of the terminology.

In the first half of the paper we critically review previous work on validating cost-effectiveness

models. The second half of the paper attempts to build upon this work in describing a number of strategies for assessing the quality of cost-effectiveness models.

1. Review of Previous Attempts to Establish Validity

The issue of validation of cost-effectiveness models was brought into sharp focus by Kassirer and Angell, writing on behalf of the *New England Journal of Medicine* (NEJM).^[1] They argued that 'Because of the discretionary nature of the methods used to analyse cost effectiveness and the increasing importance of such analyses, it is incumbent on the authors, journal editors and the funders of these studies to minimise any source of bias' and described 3 policies that the NEJM was adopting to limit the risk of bias. However, this stance does not recognise the need for assessing the validity of models for reasons other than bias. Poor models can also be caused by lack of suitable data, unknown technical relationships and uncertainty. Consequently, the NEJM policy cannot inform the broader debate of measuring issue quality.

These policies (see table I) are justified on the basis that 'the opportunities for introducing bias in to economic studies are far greater (than original scientific studies), given the discretionary nature of model building and data selection in these analyses . . . the cost side is highly artificial.'^[1]

The response of the research community has consisted of attempts to (i) demonstrate that models in cost-effectiveness studies are effectively unavoidable and (ii) explore the potential for agreement on the best methods for model construction and review. The objective of the latter is to move away from the essentially negative discussion about the risk of the pharmaceutical industry having improper influence on design and reporting of cost-effectiveness analyses towards a more constructive dialogue.

1 Some aspects of economic evaluation models may be tested in the conventional way. For example, regression analysis can be used to extrapolate the results of trials (e.g. estimating the reduction in mortality following a reduction in cigarette smoking).

Table I. *New England Journal of Medicine* policies on economic evaluations^[1]

Any study supported by industry must be funded by a grant to a not-for-profit entity such as a hospital or a university, not to an individual or group of individuals

Written assurance must be given that the agreement between the authors and the funding company ensures the authors' independence in the design of the study, the interpretation of data, writing of the report and decisions regarding publication, regardless of the results of the analysis

The manuscripts must include all the data used in the analysis, all assumptions on which the data are based and any model used in the analysis. There must be a clear explanation of the assumptions made in building the model. The model must be sufficiently straightforward and lucid so that ordinary readers can comprehend it

1.1 The Necessity for Modelling

O'Brien^[2] describes '7 threats to validity' for economic evaluations that are based on randomised controlled trials (see table II). This article is part of a much greater literature assessing the pros and cons of undertaking economic evaluations alongside controlled trials.^[3-13] However, O'Brien was the first to make explicit the main implication of these threats to validity. He argued that 'even if economic questions can be addressed prospectively as part of randomised trials there likely will remain some need for modelling to adjust or project data to address policy-relevant economic questions.'

In arguing that cost-effectiveness models are probably unavoidable, O'Brien directly rejects the implicit preference of Kassirer and Angell^[1] for trial-based cost-effectiveness studies. However, he did not set out any strategies for peer review of models, which is required if the concern of bias in cost-effectiveness studies that use models is to be addressed directly.^[1,2]

Sheldon^[14] provides a strong overview of the problems that may be encountered with modelling in economic evaluation. He accepts the 'principles and potential usefulness of decision analysis' but expresses concern about 'how the values inserted in to the logical structure are derived.' He cites many examples of technical error, poor practice and biased results, as well as poor practice in eliciting expert opinion and, in the analysis of uncer-

tainty, to construct an almost irrefutable case for the poverty of current practice in cost-effectiveness modelling. One of his major conclusions is that 'Until a clear structure for critically appraising decision models is developed, models which produce unrealistic and biased results will continue to be published.'^[14] However, even without the existence of such a structure, reviewers should not accept studies that lack sufficient transparency for quality assessment.

Sheldon briefly considers the role of models in preliminary evaluation and planning comprehensive evaluations. Surprisingly, he appears to set aside the need for a structured approach to the critical appraisal of such models, as these models are not being 'used to provide answers about the cost-effectiveness of an intervention, but rather, to further explore the uncertainty in a structured way.' Given that the structured analysis of uncertainty will lead to decisions about whether it is (i) worth developing a therapy further, and (ii) the design of trials and evaluations for the future development of a therapy, a certain confidence in the model seems desirable. A poor model could lead to the loss of potentially valuable future therapies or the use of limited resources on studies that do not answer the correct question or are not powered to answer the important question. The difference in function, i.e. that models are used to analyse uncertainty rather than provide answers about the cost effectiveness of an intervention, does not appear to justify, *a priori*, a reduction in the quality of the modelling work.

1.2 Validating Cost-Effectiveness Models

The first comprehensive review of the issue of model validation actually predates Kassirer and Angell's editorial^[1] by a number of years. Eddy^[15] identified 4 'orders' of model validation:

1. First-order validation requires expert concurrence.

2. Second-order validation compares the model predictions with data used to estimate the model parameters.

Table II. Threats to validity of trial-based data^[2]

Choice of comparison therapy. Often this is determined by the requirements of licensing authorities. Rarely, if ever, are there head-to-head comparisons for all the relevant alternative therapies
Gold standard measurement of outcomes. The ascertainment of clinical outcome in clinical trials often entails investigations that would not be part of usual care. The classic example is the use of endoscopy to identify the incidence of duodenal ulcers
Intermediate rather than final health outcomes. Many diseases have rare event rates, leading to the researchers reporting biomedical markers, such as blood cholesterol levels in cardiovascular disease
Inadequate patient follow-up or sample size. Clinical trials often end patient follow-up when a clinical event of interest occurs, such as myocardial infarction in cardiovascular disease. The treatment of such events involves the use of resources that the economic analysis needs to take into account
Protocol-driven costs and outcomes. As stated above, clinical trial resource use is unlikely to reflect usual clinical practice. Those aspects of resource use which would not occur in usual practice are referred to as protocol driven and need to be excluded from the estimate of the cost effectiveness of the therapy. Equally, any impact upon the outcome of therapy, using information that would not be available in normal clinical practice, is protocol driven and should be excluded, if possible, from the estimate of cost effectiveness
Geographical transferability of trial evidence. Differences in usual practice and the costs of resources between different health care systems mean that a therapy will have very different cost-effectiveness characteristics in different health care systems. Where medical interventions are the clinical end-point of the trial, the differences in practice can impact upon the effectiveness as well as the cost effectiveness of therapy
Selected patient and provider populations. Trials routinely define patient eligibility criteria. Some of these are driven by safety considerations, others by the desire to maximise the probability that the individuals in the trial will respond to therapy. Equally, for logistical reasons, trials will tend to recruit patients from centres with a track record of involvement in research and access to significant numbers of patients. The result of these strategies is that neither the people recruited to the trials nor the clinicians providing therapy in trials are likely to be representative of the populations in which decision makers are interested

3. Third-order validation compares the model prediction with 'other' observed data, i.e. data not used in the model construction.

4. Fourth-order validation compares pre-implementation model predictions with observed events post implementation.

Eddy argues that both first- and second-order validation should be consistently applied. Third-order validation involves a tension between using data to improve the accuracy of parameter estimates and retaining it for use in validating the model. It also requires data to be available on a scale that is not particularly common. Eddy himself recognises that the relevant data are often not available. He also recognises that fourth-order validation will only be meaningful if the conditions under which an intervention is actually implemented closely reflect those assumed in the model.

Eddy concludes that 'there is no simple and universally applicable procedure for validating a model. Each case must be considered by itself.' To facilitate this approach, he goes on to identify desirable characteristics in the reporting of cost-effectiveness models. These recommended characteristics can be summarised as transparency with regard to the

study question, the model structure, data sources, assumptions, results, sensitivity analyses and value judgements.

Eddy's review foreshadowed much of the subsequent debate and, in emphasising that the specific processes for validating a cost-effectiveness model depend upon the aim of the model and the context in which it is to be used, highlighted the Achilles' heel of cost-effectiveness modelling. This emphasis on the problems was perhaps unduly influential in shaping the tone of the subsequent discussion, whereas the constructive suggestions for good practice have had less of an impact than they deserved.

Buxton et al.^[16] respond directly to the paper by Sheldon.^[14] They argue convincingly that modelling is an unavoidable fact of life, and is likely to become even more important as economic evaluations are required for pharmaceutical reimbursement/purchasing decisions and will entail analyses when data on use in clinical practice is, by definition, not available. Having established this position, they offer the following 5 recommendations for good practice in modelling:

- (i) The model should be kept as simple as possible to aid understanding by decision-makers.
- (ii) The presentation of results should be as transparent as possible (including submission of model and data for thorough scrutiny by reviewers).
- (iii) The quality of all the data used in the model should be made explicit.
- (iv) Uncertainty within the model should be explored thoroughly using sensitivity analysis, not compensated for.
- (v) The model should be validated against the results of other models and/or the results of intervention studies.

Although it sounds reasonable to require a model to be as simple as possible to aid understanding, this may be a flawed recommendation. The obvious question is ‘As simple as possible for what?’. Simplicity may be more persuasive, but life is rarely simple. It would seem sensible to extend the transparency with regard to model results and data quality to choice of model design, i.e. the simplicity of the model should be justified by stating what simplifying assumptions have been made and explaining why such simplifications will not have a material impact upon the results of the model.

The concept of validation is slightly different from that implicit in the work of Mandelblatt et al.^[17] (see later in this section), and complementary to it, using other studies rather than actual outcome as the reference standard. However, without making it clear what action should follow from identifying a study that disagrees with the model’s results, this recommendation is of limited value. It seems sensible to expect that the importance of any observed differences to the decision under consideration should be discussed and, where possible, the source of the difference identified.

In the same year as the Kassirer and Angell editorial for the NEJM,^[1] Sonnenburg et al.^[18] offered a set of principles that could assist journals in reviewing models and identifying common errors in model building (see table III).

Although these appear intuitively reasonable, they are of limited value because they assume a doubtful consensus with regard to the concepts of

‘reasonable treatment options’,^[19] ‘key characteristics of disease’, ‘key clinical outcomes’^[20] and ‘relevant attributes in utility’.^[21,22] They do not identify processes whereby the performance of a model against these principles can be assessed.

In addition to these principles, Sonnenburg et al.^[18] highlighted a number of ‘common’ errors in model construction that reviewers of models should look for (table III). The description of common errors is valuable, but perhaps the most important contribution of this paper was to commence the search for an agreed framework for the peer review or critical assessment of decision analysis/cost-effectiveness models.

Mandelblatt et al.^[17] discussed the role and techniques of modelling in cost-effectiveness analysis at length. They recognise that ‘Models are only as good as their ability to represent reality at the level needed to draw useful conclusions; this, in turn depends upon their structure and on the assumptions that go in to the models.’ The process of assessing how good is a model is divided into ‘model validation’ and ‘peer review’.

Model validation consists of ‘face validation’ and ‘predictive validity’. Mandelblatt et al.^[17] argue that model validation ‘may have to rest solely

Table III. Principles that could assist journals in reviewing models and identifying common errors in model building^[18]

Principles
The type of model should reflect the nature of the clinical problems, e.g. if the clinical problem persists over a long time period, a Markov model is likely to be the most appropriate type of model
All reasonable treatment options, including extremes like watchful waiting, should be included in the model
The key characteristics of the disease should be included in the model
The key clinical outcomes should be included in the model
The utility structure should incorporate all relevant attributes
Errors
Model syntax
Conditioning of action on unobservable states
Violations of symmetry in modelling prognosis
Failure to link variables that are inherently related
Inconsistent bias in assumptions
Modelling results of diagnostic tests
Modelling of treatment

on evaluating the inherent reasonableness of model assumptions as a representation of reality'. Their conclusion on face validation resembles 'buyer beware', concluding that 'users must decide whether the model is sufficiently detailed to capture the important features of the problem.'

Predictive validity is identified as important but problematic. Mandelblatt et al.^[17] state that the accuracy of the model's cost and outcome predictions should be assessed when data are available, for either 'intermediate or final numerical predictions'. However, it is important that the predictive validity of the model focuses on the modelled relationship between inputs and outputs. A model should not be criticised for inaccuracy due to ignorance about the true value of inputs. Care must be taken to identify the source of any inaccuracy in a model's predictions. The implications of poor model structure and poor data are likely to be quite distinct and, therefore, they should be assessed separately.

For Mandelblatt et al.,^[17] verification is concerned with the technical accuracy of the model and should identify 'programming errors, data entry errors, and logical inconsistencies in the model specification.' They recommend that verification include testing the performance of the model under hypothetical conditions such as 100% and 0% efficacy. It should also include the comparison of intermediate outputs from the model with the data entered into the model to check consistency. They do not describe in detail what this means but recommend that all 'Reports based on models should contain assurances that the model has been verified in this manner.'

With regard to peer review, Mandelblatt et al.^[17] state that it is 'incumbent . . . to provide for the possibility of peer review and replication by colleagues who are able to examine the inner workings of the model.' They say that peer review can normally be satisfied by providing detailed structural assumptions and data for the model, whilst recognising there is no agreed approach to the presentation of this information. This said, they go on to argue that 'a willingness to release model software and data for peer review under appropriate

protection must exist on the part of CEA (cost-effectiveness analysis) investigators in order to guarantee the integrity of modelling.'

Mandelblatt and colleagues are to be applauded for making some clear recommendations about tests that cost-effectiveness models should be subjected to before they, or their results, are used in the real world. However, given the extensive knowledge of good modelling practice demonstrated by the authors, it is disappointing that 'face validation' was reduced to 'buyer beware'.

Rittenhouse^[23] produced a detailed exploration of the role of modelling in economic evaluations in healthcare for the Office of Health Economics. Although saying much about the lack of trained and/or talented people to produce quality models, Rittenhouse said little specific on the way forward to a more 'rigorous and transparent review policy'.

The most recent contribution to the debate on the critical appraisal of models in economic evaluations was produced by Halpern et al.^[24,25] They offer a framework that was 'intended to help model designers and reviewers focus on the key criteria . . . of the development/evaluation process.'

Their 'model evaluation checklist' (table IV) identifies 3 stages of model development and evaluation: model approach, model specifics and model analysis. They use this structure to present a detailed description and discussion of the processes involved in model development and in so doing argue convincingly for transparency in process. Like Mandelblatt et al.,^[17] they briefly consider model validation and verification.

For Halpern and colleagues,^[24,25] model verification is 'the use of sensitivity analysis to determine whether the model is performing appropriately, which is a somewhat subjective judgement.' This form of verification appears similar to face validity, in requiring the model's results to make intuitive sense and differs from Mandelblatt and colleagues' definition of verification,^[17] which is concerned with identifying technical errors within a model.

Halpern et al.^[24,25] suggest that verification will often involve 'running the model under simplify-

Table IV. Model evaluation checklist (from Halpern et al.,^[24] with permission)

Model approach
Study question specified
Need for modelling vs alternative methodologies discussed
Type of model identified
Reason for use of this model type discussed
Model scope specified
time frame
perspective
comparator(s)
setting/country/region
Basis of scope discussed
Model specifics
Source and strength of model data specified
Model assumptions discussed
Model parameters available in technical appendix
Values and sources for model parameters specified
event probabilities
rates of resource utilisation
costs
Criteria for evaluating quality of data specified
Relevant treatment strategies included
Relevant treatment outcomes included
Biases discussed and explored
Model analysis
Base case results presented and described
Sensitivity analysis performed
unidimensional
multidimensional
best/worst case
threshold
Key cost drivers identified
Verification performed
Validation performed

ing assumptions . . . to convince healthcare personnel and policy makers of the usefulness of the model.’ They acknowledge that any unexpected changes in results identified in the verification process should be fully evaluated, because ‘It could reflect model errors or unsuspected linkages between model values and outcomes.’

With regard to model validation, Halpern et al.^[24,25] differentiate between structural validity and content validity. Structural validity asks how well the model represents the patterns seen in real-world decisions; content validity asks how well the

data used reflect current knowledge and practice. Although these are sensible questions, how they should be operationalised is not made clear.

Halpern et al.^[24,25] also recognise the desirability of validating the model predictions against either real-world data or previously developed models. As with Mandelblatt et al.,^[17] the problem with the recommendation is that they do not make it clear what action should follow from identifying a study that disagrees with the model’s results.^[24]

The importance of the data sources for a model was recognised by Nuijten in a recent paper in *PharmacoEconomics*.^[26] Having discussed the problems of using data from different sources, Nuijten makes a number of recommendations for good practice, which operationalise the concept of transparency identified by many other authors (see table V).

1.3 Summary

The literature to date has identified the importance of:

- transparency in model building
- verification of models with regard to technical implementation
- validation of models with regard to structure and content
- validation of models with regard to outcomes.

Nuijten’s work^[26] has operationalised the concept of transparency, and Mandelblatt et al.^[17] have indicated a valuable way forward with regard to verification. Brief consideration has been given to what is meant by validation for the structure, content and predictions of models. However, these concepts are not well developed, and a significant proportion of the literature suggests that these are issues of professional judgement rather than tests that can be objectively applied.^[7,23-25]

Modelling is a flexible and complex evaluative technique, which can be used for a number of purposes. No one should argue for a ‘one-size-fits-all’ approach to the critical appraisal of models. However, it is difficult to argue for the scientific nature of economic modelling when the principles of critical appraisal are as unformed as they presently are.

Table V. Recommendations for good practice in the selection of data sources for use in modelling studies^[26]

The sources of study data should be recommended and explained in sufficient detail

For clinical outcomes, the general rule may be to assume that data are not country specific. For each study this assumption has to be controlled

For economic measures and information on therapeutic choices, the general rule may be that country-specific data sources have to be used

For each location in the model (e.g. a Markov state), the patient subpopulation has to correspond as much as possible with the population in the data source(s) being used

To describe for each data source, the type, number of patients, study population, countries, date of data collection, cost of access to database and data abstraction. A justification for the final 'Yes' or 'No' decision to use a data source should be presented, based on the advantages and disadvantages of the specific source

2. A Framework for Assessing Validity

This section attempts to develop a framework for assessing the validity of economic evaluation models.

At the broadest level, it must be recognised that while a comparison of modelled outcomes with actual outcomes is desirable, very few appropriate observational data sets exist. Consequently, a gold standard test of validity is rarely possible.

This leads to the need for assessing not only the results of models but also the structure and inputs of models.

It also needs to be recognised that the aim of a model may not simply be to produce an accurate estimate of cost effectiveness. Models are used for many different reasons, and hence their characteristics will vary between applications. However, their overarching purpose is essentially the same: to help the decision-maker reach a better informed and rational decision. Consequently, the model may also be assessed in terms of its value to the decision-maker, which can be seen as much broader than the figures produced from a mathematical algorithm.

It is therefore proposed that 4 different aspects of the modelling process should be assessed: (i) the structure of the model; (ii) the inputs of the model; (iii) the results of the model; and (iv) the value of

the model to the decision maker. Each of these aspects can then be assessed in different ways, as discussed in sections 2.1 to 2.4.

2.1 Structure of the Model

At the most fundamental level, it is important that the possible pathways described by the model are feasible and sensible. More specifically, current practice in the decision-making context to which the model is being applied should be capable of being described by the model's structure.

Although such 'descriptive validity' is intuitively appealing, it is not straightforward. Not all models require absolute detail for aspects of the clinical area. For example, mapping out comprehensive treatment pathways for each individual adverse effect from treatment may not alter the results of the model over and above a simpler representation, and may only confuse the decision maker through its added complexity.

Consequently, the intuitively appealing notion of descriptive validity cannot be described in a clear and unambiguous manner that is capable of providing a simple test. It is inextricably linked to the purpose of the model and the knowledge about the processes that are being modelled. Construction of any 'test' will therefore be open to subjectivity. A good test should avoid subjectivity in its calculation, although no test can remove the role of subjectivity in its interpretation.

It may be possible to require that any simplification be explicitly justified by demonstrating that increased complexity cannot logically have a significant impact upon the output/decision implications of the model. Absence of data is not in itself a justification for simplifying important issues. Rather, the model should make explicit assumptions that can be challenged and explore the impact through sensitivity analysis. This may be merely a development of being specific about the purpose of the model. However, there is a risk that this will lead to extensive models requiring the expansion of models and the search for ever more uncertain data, only to find that the model is sensitive to uncertain data.

2.2 Inputs to the Model

Given the existence of established methods for assessing the validity and relative worth of study data,² an attempt at defining 'internal validity' appears more promising than the highly context-specific nature of 'descriptive validity'. However, it should be noted that the relative importance of one data source over another is not constant – using an 'inferior' data source for an insignificant parameter may not weaken the model. It could therefore be argued that a hierarchy of data is a secondary consideration to identifying a hierarchy of parameters within the model.

The work of Nuijten^[26] is very useful with regard to assessing the validity of data inputs. His approach requires a very detailed description of all data identified as potentially relevant for the model, together with some critical appraisal of each item of data, finished with a formal justification of the choice of each item of data. Such an approach to the process of populating a model would be welcomed by many who would see this process as a way to enforce rigour, and even emulate the movement toward evidence-based medicine. However, there are significant risks associated with this approach.

One possible danger is that if it encouraged a more formal appraisal of data sources along the lines of evidence-based medicine, it could damage modelling by incorporating within models the threats to validity outlined by O'Brien^[2] (and which he saw as the main reasons for requiring models). For example, the methodological hierarchy proposed by the US Task Force on Preventative Healthcare 'rewards' internal validity ahead of external validity and does not discriminate against explanatory trials, which typically incorporate features such as

placebos and blinding that can invalidate economic models.

Despite the dangers of such an approach, we feel that it should be adopted, but with significant amendments.

First, it should be made clear that a full systematic review is not required for each parameter. Even a fully systematic approach to data identification and critical appraisal will not eliminate the need for a sensitivity analysis, which can incorporate more aspects of uncertainty than just interstudy variability. Although not inherently harmful, a fully systematic approach to data identification can lull the researcher and the decision maker into a false sense of security when assessing the model, and lead them away from taking full consideration of the sensitivity analysis and other aspects of the modelling process.

Secondly, the issue of how the different sources of data relate to one another must be given added emphasis. Nuijten^[26] mentions the need for patient populations to correspond as much as possible to one another. However, this issue needs to be investigated more fully. It is likely that trade-offs will arise when competing data sources are available. For example, one source may be more appropriate in terms of its study population, but less appropriate in terms of the study methods (e.g. resource use based on patient recall rather than third-person observation). These trade-offs should be explicitly identified, and perhaps the implications of choosing one combination of data over another should be explored empirically.

Finally, the critical appraisal process implied by Nuijten,^[26] and its resultant justification of the data utilised in the model, needs to be along different lines to the evidence-based medicine approach. An important aspect of the justification is to measure the value of the data jointly with the importance of the parameter.

2.3 Results of the Model

The results of the model, e.g. the estimate of cost effectiveness, can only truly be validated in one way: through comparing the modelled estimates

2 Scales for ranking the methodological quality of a study have been proposed by several researchers. For example, the US Task Force on Preventative Health Care ranked studies in the following order: (i) meta-analyses; (ii) randomised controlled trials; (iii) non-randomised controlled trials; (iv) quasi-experimental studies; (v) descriptive studies; and (vi) opinion.

with those produced in real life, which could be termed 'predictive validity'. Although such an approach can rarely be implemented at the time of the modelling process, it is feasible to design future studies to generate the data necessary for a such a test. However, such an approach is potentially expensive and self-defeating. By indicating that future studies that will help in the assessment of the model are forthcoming, the model validation process may encourage decision makers to give little weight to the model results and adopt a 'wait and see' policy, awaiting the results of the subsequent work.

In addition, without careful design, the observational studies may suffer from the threats to validity that O'Brien^[2] identified, and consequently, the model would be 'validated' against the sort of flawed estimates that it was designed to replace. This is not to argue that model results should not be compared to real world data for validation purposes. Where data exist on the outcomes, the results of the model should be compared to them and any variation should be explained. The applicability of the 'validation data' to the decision-making context for which the model is being developed should be made explicit. This should take account of patient characteristics, healthcare system characteristics and cost structure.

Finally, when assessing the validity of the results of the model it is essential that the computational correctness of the model is assured. In this regard, the work by Mandelblatt et al.^[17] describing the verification process is useful. Strategies for quality assurance with regard to data handling, data input and computational formulation need to be agreed upon. It may be that the good practices in data handling developed for trial-based research, such as double data entry, can be adopted with little or no modification.

2.4 Value of the Model to the Decision Maker

Previously (in the introduction to section 2), we argued that the value of the model to the decision maker goes beyond its ability, or otherwise, to produce accurate predictions. A model that is too com-

plex for the decision maker to understand (and/or trust) is likely to result in reversion to simple, and probably imperfect, rules-of-thumb, with all the attendant risks of irrational and opaque decision making. Likewise, models that are too simple are unlikely to have credibility.

A model's ability to influence the behavioural change necessary for the optimal patterns of practice to be adopted adds value beyond its mathematical predictions. In this regard, it is important that the model is:

- Appropriate to the decision making context.
- Understandable. Its level of complexity needs to be tailored to its main audiences. This may be achieved by producing several 'reduced form' models from the master model, which are targeted at particular audiences.
- Believable. The key decision-makers need to 'buy into' the model's predictions, so that they genuinely try and promote the appropriate behaviour. For this to be achieved, it is essential that the model is clinically credible and transparent. Models primarily seek to change the behaviour of clinicians, and without their tacit approval it is likely that they will be ignored.

One possible way in which the model can be judged in this regard is that the structure, inputs and results of a model should be compared with those of existing models, and differences should be capable of being explained and justified. Some papers have adopted this approach.^[27,28] Such an approach is analogous to the notion of 'encompassing' that is supported by some econometricians, and this ensures that research in any particular area is progressive.

This issue is difficult to evaluate in any formal way. It could be argued it is a presentational issue that lies beyond any formal evaluation of a model. For example, the results of trials have been criticised for producing measures of effect that are meaningless to physicians and, as a response, the notion of 'number needed to treat' was developed.

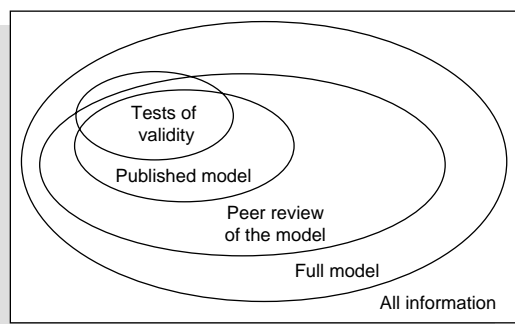


Fig. 1. Model evaluation.

3. The Broader Model Evaluation Process

The literature to date has attempted to assess the validity of models, and our preceding analysis has pointed to several promising aspects of the literature. However, it is doubtful that a set of objective tests of validity will ever be produced (or indeed that such an approach would be desirable). It is therefore important to assess whether and how models will be evaluated in the absence of strictly defined tests. Objective tests do not exist for clinical studies, such as randomised clinical trials, yet they command great trust and respect within medicine when designed and conducted correctly.

Figure 1 shows schematically how validity testing is only part of a much broader evaluation process. Any cost-effectiveness model resides within the universal set of all information. The model (and associated documentation) that is published is a subset of the full model. However, more information than is published may need to be peer reviewed prior to publication. Any tests of validity that can be developed would need to form part of any full cost-effectiveness model, although would not necessarily be incorporated as part of all the associated models.

The absence of clearly defined objective tests of validity does not signify the death of model evaluation. Rather, it highlights the importance of other aspects of model evaluation. These include the need to develop quality assurance during the model

building process (e.g. good practice guidelines) and quality assessment during the peer review and dissemination process (e.g. critical appraisal guidelines). Although these components of the model evaluation process are subjective, real improvements are possible, whereas the search for the definitive test of model validity continues to elude researchers.

4. Discussion

The structure we have outlined does not have an explicit place for the assessment of the robustness/sensitivity of the model, even though many commentators mention the desirability of robustness with respect to a model's results. Some in the field of modelling see robustness as a measure of validity.^[24] This has the attraction of annexing sensitivity analysis into the territory of validity testing. However, such a belief is misplaced. Although robustness is a desirable property of a model, this does not necessarily imply that a robust model is in some way 'better' than a 'sensitive' model. The robustness/sensitivity of the model is separate from its validity, as it is a characteristic of the uncertainty around the true values of the parameters within the model, not of the construction of the model. Therefore, sensitivity should be considered separately from the validity of the model, although the analysis of uncertainty is obviously an area where good practice is desirable.

Equally, we have not given an explicit role to the assessment of expert opinion within the validation of cost-effectiveness models. Expert opinion has several potential roles within modelling. First, opinion may be required to produce various estimates for the model in the event of observational data not being available. This is particularly likely in the early stages of technology development.^[29] Secondly, if professionals accept the model as plausible, their support will enhance the model's ability to effect behavioural change. However, professional opinion cannot be regarded as a formal test of the validity of a model; their endorsement is preferred but neither necessary nor sufficient for the existence of a good model.

The literature to date, and our own suggestions here, demonstrate the difficulty in assessing quality of model-based economic evaluations. However, the success of clinical trials in obtaining such an influential position demonstrates that the absence of simple objective tests is not an insurmountable barrier to acceptance by the research and decision-making communities. What is necessary is the identification and general adoption of standards of good practice, alongside a critical debate around the principles of good practice to ensure that these develop to reflect improvements in the methodology of economic evaluation and decision analysis modelling.

5. Conclusions

We propose the following set of statements of good practice. Adherence to these proposals will not produce more valid models in a quantifiable sense, and indeed we have argued that this is not generally possible, but will instead produce models whose purpose, structure and data are more clearly presented and justified. This should facilitate assessment of whether the model is fit for its stated purpose, whether the correct model has been built and whether it has been built correctly.

- The purpose of a model must be clearly specified, as this will have a fundamental impact on several aspects of the model, such as its perspective, the comparators, its complexity, data sources and outputs.
- A clear justification of the need for a model must be made, together with a justification of the approach taken (in terms of the model type and its structure). This will require an indication of the lack of any alternative information or an appraisal that demonstrates its weakness.
- All data that is relevant to the model need to be appraised in terms of its appropriateness to the purpose of the model and its compatibility with the other data sources. The reasons for choosing certain data over other possible data should be justified.
- The structure, inputs and results of a model should be compared with those of existing models, and differences explained and justified.

- Intermediate outputs of a model should be compared to existing data if available, and final outputs of a model should be compared to prospectively collected data if appropriate. Any differences should be explained and justified.
- The mathematical correctness of the model needs to be verified through quality assurance procedures.

Acknowledgements

We wish to thank Dr David Veenstra for his excellent and constructive discussion of this paper, and Andrew Mitchell for his observation about the relative importance of parameters versus data in the section on inputs to the model. We would also like to thank all the participants at the conference for their constructive comments and contributions.

References

1. Kassirer JP, Angell M. The journal's policy on cost effectiveness analyses [editorial]. *N Engl J Med* 1994; 331 (10): 669-70
2. O'Brien B. Economic evaluation of pharmaceuticals: Frankenstein's monster or vampire of trials? *Med Care* 1996; 34 (12): DS99-108
3. Drummond MF, Stoddart GL. Economic analysis and clinical trials. *Control Clin Trials* 1984; 5: 115-28
4. Drummond MF, Davies L. Economic analysis alongside clinical trials: revisiting the methodological issues. *Int J Technol Assess Health Care* 1991; 7: 561-73
5. Adams ME, McCall NT, Gray DT, et al. Economic analysis in randomised controlled trials. *Med Care* 1992; 30: 231-43
6. Morris J, Goddard M. Economic evaluation and quality of life assessments in cancer clinical trials: the CHART trial. *Eur J Cancer* 1993; 5: 766-70
7. Bonsel GJ, Rutten FFH, Uyl-de Groot CA. Economic evaluation alongside cancer trials: methodological and practical aspects. *Eur J Cancer* 1993; 29A: S10-4
8. Drummond MF. Economic analysis alongside clinical trials: problems and potential. *J Rheumatol* 1995; 22: 1403-7
9. Powe NE, Griffiths RI. The clinical-economic trial: promise, problems, and challenges. *Control Clin Trials* 1995; 16: 377-94
10. Mauskopf J, Schulman K, Bell L, et al. A strategy for collecting pharmacologic data during phase II/III clinical trials. *Pharmacoeconomics* 1996; 9: 264-77
11. Ellwein LB, Drummond MF. Economic analysis alongside clinical trials: bias in the assessment of economic outcomes. *Int J Technol Assess Health Care* 1996; 12: 691-7
12. Rigby K, Silagy C, Crockett A. Can resource use be extracted from randomized controlled trials to calculate costs? *Int J Technol Assess Health Care* 1996; 12: 714-20
13. Rittenhouse BE. Exorcising protocol-induced spirits: making the clinical trial relevant for economists. *Med Decis Making* 1997; 17: 331-9
14. Sheldon T. Problems of using modelling in the economic evaluation of health care. *Health Econ* 1996; 5 (1): 1-11
15. Eddy D. Technology assessment: the role of mathematical modelling. In: Mosteller F, editor. *Assessing medical tech-*

- nologies. Washington, DC: National Academy Press, 1985: 144-60
16. Buxton MJ, Drummond MF, van Hout BA, et al. Modelling in economic evaluation: an unavoidable fact of life. *Health Econ* 1997; 6 (3): 217-28
 17. Mandelblatt JS, Fryback DG, Weinstein MC, et al. Assessing the effectiveness of health interventions. In: Gold MR, Siegel JE, Russell LB, et al., editors. *Cost-effectiveness analysis in health and medicine*. New York (NY): Oxford University Press, 1996: 135-64
 18. Sonnenberg FA, Roberts MS, Tsevat J, et al. Toward a peer review process for medical decision analysis models. *Med Care* 1994; 32 (7): JS52-64
 19. Ellis J, Mulligan I, Rowe J, et al. Inpatient general medicine is evidence based. *Lancet* 1995 Aug 12; 346 (8972): 407-10
 20. Morris S, McGuire A, Caro J, et al. Strategies for the management of hypercholesterolaemia: a systematic review of the cost-effectiveness literature. *J Health Serv Res Policy* 1997; 2: 231-50
 21. Deverill M, Brazier J, Green C, et al. The use of QALY and non-QALY measures of health related quality of life: assessing the state of the art. *Pharmacoeconomics* 1998; 13 (4): 411-20
 22. Dolan P. Valuing health related quality of life: issues and controversies. *Pharmacoeconomics* 1999; 15 (2): 119-27
 23. Rittenhouse B. *Use of models in economic evaluations of medicines and other technologies*. London: Office of Health Economics, 1996
 24. Halpern MT, Luce BR, Brown RE, et al. Health and economic outcomes modeling practices: a suggested framework. *Value Health* 1998; 1 (2): 131-47
 25. Halpern MT, McKenna M, Hutton J. Modeling in economic evaluation: an unavoidable fact of life [letter]. *Health Econ* 1998; 7 (8): 741-2
 26. Nuijten MJC. The selection of data sources for use in modelling studies. *Pharmacoeconomics* 1998; 13 (3): 305-16
 27. Tosteson AHA, Rosenthal DI, Melton LJ, et al. Cost effectiveness of screening perimenopausal white women for osteoporosis: bone densitometry and hormone replacement therapy. *Ann Intern Med* 1990; 113: 594-603
 28. Sculpher M, Michaels J, McKenna M, et al. A cost utility analysis of laser-angioplasty for peripheral arterial occlusions. *Int J Technol Assess Health Care* 1996; 12 (1): 104-25
 29. Sculpher M, Drummond M, Buxton M. The iterative use of economic evaluation as part of the process of health technology assessment. *J Health Serv Res Policy* 1997; 2 (1): 26-30
-
- Correspondence and offprints: Mr *Chris McCabe*, Sheffield Health Economics Group, School of Health and Related Research, University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA, England.
E-mail: c.mccabe@sheffield.ac.uk