Common Methodological Flaws in Economic Evaluations

Michael Drummond, BSc, MCom, DPhil, and Mark Sculpher, BA(Hons), MSc, PhD

Abstract: Economic evaluations are increasingly being used by those bodies such as government agencies and managed care groups that make decisions about the reimbursement of health technologies. However, several reviews of economic evaluations point to numerous deficiencies in the methodology of studies or the failure to follow published methodological guidelines.

This article, written for healthcare decision-makers and other users of economic evaluations, outlines the common methodological flaws in studies, focussing on those issues that are likely to be most important when deciding on the reimbursement, or guidance for use, of health technologies.

The main flaws discussed are: (i) omission of important costs or benefits; (ii) inappropriate selection of alternatives for comparison; (iii) problems in making indirect comparisons; (iv) inadequate representation of the effectiveness data; (v) inappropriate extrapolation beyond the period observed in clinical studies; (vi) excessive use of assumptions rather than data; (vii) inadequate characterization of uncertainty; (viii) problems in aggregation of results; (ix) reporting of average cost-effectiveness ratios; (x) lack of consideration of generalizability issues; and (xi) selective reporting of findings.

In each case examples are given from the literature and guidance is offered on how to detect flaws in economic evaluations.

Key Words: cost-effectiveness analysis, decision-makers, reimbursement, health technology assessment

(*Med Care* 2005;43: II-5–II-14)

Economic evaluations are increasingly being used by those bodies such as government agencies and managed care groups that make decisions about the reimbursement of health technologies.[1–4] The growing prominence of economic evaluation in decision-making processes has led to increased scrutiny of studies and their methods. That is, given the importance of the decisions that economic evaluation seeks to inform, can the studies be trusted to deliver reliable results?

From the Centre for Health Economics, University of York, York, United Kingdom.

Mark Sculpher receives a Career Award in Public Health from the NHS Research and Development Programme.

Reprints: Michael Drummond, Centre for Health Economics, University of York, Heslington, York YO10 5DD, UK. E-mail: chedir@york.ac.uk.

ISSN: 0025-7079/05/4300-0005

Several reviews of published economic evaluations are not very encouraging, pointing to numerous deficiencies in the methodology of studies or the failure to follow published methodological guidelines.[5–10] Some commentators argue that economic evaluations are inherently more subject to methodological flaws than clinical trials, which form the basis of registration decisions for drugs and other technologies.[11] Others argue that the mere existence of substantial industrial sponsorship for economic evaluation increases the risk of bias.[12,13]

However, the mere fact that deficiencies exist in published studies may not mean that the analysts concerned have failed or been unduly influenced by the sponsors of the study. Rather, the question we need to ask is whether, given the data available to them at the time of the study, the analysts made the best possible contribution to the decision-making process. This is particularly important in the context of reimbursement decisions for health technologies, because a decision needs to be made at a given point in time, usually around the time of launch of the product. This does not mean, of course, that the decision cannot be revisited as further evidence emerges.

In the context of using economic evaluations for reimbursement decisions, the most relevant review is that by Hill et al[14] of economic submissions made to the Pharmaceutical Benefits Advisory Committee in Australia. Of a total of 326 submissions between January 1994 and January 1997, 218 (67%) had significant problems and 31 had more than 1 problem. Although the authors were initially surprised by the number of submissions with methodological problems, they concluded that this situation was not unexpected given the nature of the process, which requires a complicated synthesis of data from a variety of sources.

This article is written for healthcare decision-makers and other users of economic evaluations. It outlines common methodological flaws in studies, focusing on those issues that are likely to be most important when deciding on the reimbursement, or guidance for use, of health technologies. When possible, suggestions are made for dealing with the methodological problems identified.

COMMON FLAWS IN STUDY DESIGN

It is worth noting that there are 2 basic approaches to economic evaluation. In *trial-based studies,* economic data

(eg, resource utilization, quality of life) are collected alongside a single clinical study, usually a controlled clinical trial. In *modeling studies,* data from a wide range of sources (eg, existing clinical trials, observational studies) are synthesized using an economic model.

Although some commentators[11,15] take the view that, given the choice, trial-based studies are in some sense superior, the view taken here is that rather than being alternative forms of analysis, they are complementary. This is illustrated by the fact that most trial-based studies involve some element of modeling or synthesis.[16]

A classic example of this is the study by Mark et al,[17] undertaken alongside the GUSTO I trial of tPA versus streptokinase in patients with acute myocardial infarction (MI). Many elements of data such as those on survival at 1 year, resource use over 1 year, and health state valuations at 1 year came from the trial. However, to estimate the life-years and quality-adjusted life-years (QALYs) gained, it was necessary to use an observational study of patients experiencing MIs, and cost data were obtained from a mixture of sources.

Some of the methodological flaws discussed subsequently apply both to trial-based and modeling studies, whereas others apply more to 1 type of study than the other.

Omission of Important Costs or Benefits

Most published methodological guidelines for economic evaluation state that all relevant costs and benefits should be considered in the analysis. Clearly, 1 way of biasing a study would be to omit from consideration items that might count against the treatment or program of interest. For example, in an evaluation of a new drug, one might be tempted to downplay the impact, on resource use or quality of life, of adverse events associated with the therapy.

Of course, the range of costs and benefits that is deemed "important" depends on the viewpoint, or perspective, of the analysis. The adoption of a broad societal viewpoint raises the biggest measurement challenges, and it may be tempting to omit items from consideration. In these situations, it is necessary to make a judgment on whether the omitted items, if included, would make a substantial difference to the study results. Sometimes, the analysts can help by undertaking a sensitivity analysis. For example, in an evaluation comparing tPA with streptokinase, Mark et al included the long-term costs of disabling strokes in a sensitivity analysis and found that this did not have a major impact on the overall study results. From this, one might infer that including the nonmedical costs of those strokes (the only clinical outcome that was inferior for tPA) would not have changed the results of the study.

Although the choice of perspective, and adherence to the chosen perspective, is of general importance in judging the methodological quality of economic evaluations, it is probably not a major source of methodological flaws in submissions to reimbursement agencies, because usually the agency can be quite explicit about the perspective to be adopted and the costs and benefits that should be considered. However, there remain some strong normative arguments in favor of a broad societal perspective,[18] which may conflict with that preferred by the decision-maker. In these circumstances, it can be argued that an appropriate analysis will explore the robustness of the results to the chosen perspective.

Selection of Alternatives for Comparison

The cost-effectiveness of a given therapy can only be judged in relation to 1 or more alternatives, which may include "doing nothing." Therefore, the choice of alternatives is critical because it provides the basis of incremental cost-effectiveness analysis. Clearly, the incentive for the manufacturer or sponsor of a given health technology is to pick an alternative that shows its own therapy in a good light.

Most of the published guidelines for economic evaluation state that the most relevant alternative is "current practice" or the most widely used therapy/therapies in the jurisdiction concerned. Usually this should be possible to determine, so the risk of flaws is probably small. However, problems could occur if "current practice" is itself inefficient. Therefore, some guidelines state that "a viable low-cost alternative" should be considered.[6,19] More generally, it can be argued that all relevant alternative therapies should be considered, including no treatment, although this may not be straightforward in terms of measurement.

Another problem, most frequently occurring in trial-based studies, is that the alternatives compared in the economic evaluation are inappropriate because the comparison is constrained by the clinical trial itself. This arises because economic evaluations are often conducted alongside trials being undertaken for drug registration purposes, in which the comparator to the drug of interest may be a placebo or an older therapy, neither of which may represent "current practice" in the jurisdiction concerned.

Sometimes current practice may differ by jurisdiction. For example, in their economic evaluation of enoxaparin (a low-molecular-weight heparin) for prophylaxis against deep-vein thrombosis (DVT), O'Brien et al[20] found that all the trials used standard heparin as a comparator. However, in Canada, the setting for their study, warfarin was the most commonly used drug. That is, the trials, which had been predominantly conducted in Europe, had the "wrong" comparator. Therefore, a modeling study was undertaken using *indirect comparisons*, that is, comparisons using clinical data that do not come from head-to-head studies of the 2 therapies concerned. (More on this subsequently.)

As in the case of the choice of study perspective, it should be possible (in a reimbursement environment) for the payer or decision-maker to specify which alternatives should

be compared. In the United Kingdom, this is usually part of a scoping exercise undertaken by the National Institute for Clinical Excellence (NICE).

COMMON FLAWS IN DATA COLLECTION AND ANALYSIS

Making Indirect Clinical Comparisons

It was mentioned previously that clinical trials may not always compare the relevant alternatives. In the context of reimbursement decisions, this problem becomes most acute when there are a number of new therapies for a given condition (eg, glycoprotein IIb/IIIa antagonists for treatment of acute coronary syndromes, anti-TNF therapy for rheumatoid arthritis). In these situations, the payer or decision-maker may want to assess not only whether the newer therapies (in general) represent good value for the money, but also whether there are differences in cost-effectiveness among them. (Indeed, manufacturers may want to make the same distinctions to justify a premium price for their product.)

The problem is that, at the time of the reimbursement decision, it is highly unlikely that there will be head-to-head clinical trials of the therapies concerned. In part, this may be because such trials are costly and time-consuming to undertake. In part, it may be because it is not possible to undertake a trial comparing 2 investigational therapies. Whatever the reason for the lack of head-to-head studies, an economic evaluation seeking to provide relevant information to a reimbursement agency will have to incorporate an estimate of treatment effect that is based on indirect comparisons.

It has long been argued that indirect comparisons are potentially subject to methodological flaws, mainly because, outside the confines of a single randomized, controlled clinical trial, one cannot be sure that the patients enrolled in the various trials are equivalent in terms of baseline risk, that the settings for the trials are comparable, and that end points are measured in the same way. Therefore, an apparent superiority for 1 therapy, over another, in an indirect comparison may be as much the result of differences in the trials as to differences between the therapies themselves.

These concerns have led the Australian Pharmaceutical Benefits Advisory Committee to specify that claims of superiority (for 1 drug over another) can only be substantiated through head-to-head studies.[1] In the United Kingdom, the position adopted by NICE is more flexible, and it is expected that data synthesis will be required, because of the absence of head-to-head comparisons of the relevant alternatives.[3]

If indirect comparisons are to be considered, how should they be made? Often they are made using a common therapy, to which the therapies of interest have been compared in separate trials. (For example, in the case of new drugs, it is likely that they have both been compared with placebo, or an older therapy.)

Some analysts have made these comparisons without adjustment based on a common comparator when available. For example, O'Brien et al[20] compared the results of enoxaparin and warfarin placebo-controlled trials in terms of the pooled rates of DVT on single arms of the trial. If there were important differences in the trials such as the enrollment of patients with different baseline risks of DVT, the comparison could be misleading because the analysis involves breaking randomization.

One improvement on this is the approach used by Palmer et al[21] in a study of glycoprotein IIb/IIIa antagonists (GPAs) in acute coronary syndromes. They used a meta-analysis of 18 trials, each of which compared GPAs with a common comparator (standard therapy). The metaanalysis generated a pooled estimate of the relative risk reduction of cardiac events compared with standard therapy. In a decision analytic model, this was then used as the treatment effect for GPAs by adjusting an estimate of the baseline risk without GPAs from an observational study relevant to the setting where the study was required (the United Kingdom).

The use of estimates of treatment effect based on indirect comparisons when there is a common comparator has recently been shown on many occasions to agree with the results of head-to-head clinical trials.[22] Clearly, a more challenging situation exists in which there is not a common comparator. This was the case in a recent study of the relative cost-effectiveness of newer drugs for treatment of epilepsy.[23] In this case, Bayesian Markov Chain Monte Carlo models for multiparameter synthesis were used.[24] Although the use of such models cannot guarantee the absence of bias (neither can head-to-head trials!), they can give decision-makers the best available estimate of relative effectiveness (and cost-effectiveness) within the constraints of data availability.

Inadequate Representation of the Effectiveness Data

In making the case for a given therapy or health technology, the temptation is to present the subset of evidence that best supports the argument. This is why the proponents of systematic review go to considerable lengths to identify all clinical studies, both published and unpublished.

Clearly, any economic evaluation should be based on an adequate representation of the underlying effectiveness data. One of the most widely discussed examples of problems in this regard concerns the study by Jönsson and Bebbington[25] on selective serotonin reuptake inhibitors (SSRIs) for treatment of depression. They used, as the basis for their decision analytic model, a clinical study by Dunbar et al[26] published in the *British Journal of Psychiatry*. The clinical study was in fact a summary of 6 separate trials of paroxetine (the SSRI) compared with imipramine (a tricyclic antidepressant).

Using the clinical data from the Dunbar et al study, which showed superiority for paroxetine in terms of withdrawal from therapy, Jönsson and Bebbington were able to construct an economic argument in favor of the SSRI based on the cost of managing "dropouts" from therapy. However, their study was criticized by Freemantle and Maynard,[27] who argued that metaanalyses of *all* the studies of paroxetine compared with conventional antidepressants showed a much smaller advantage for the new drug. In fact, the study by Dunbar et al was the *only* one showing a clear advantage judged by the conventional criteria of statistical significance. At least reasons for the exclusion of some of the trials should be given.

If the use of a subset of the clinical evidence is misleading, under what circumstances could an economic study based on a single clinical trial be valid? In a response to Freemantle and Maynard, Jönsson[28] argued that, in good faith, the economic study was based on clinical data published in a good-quality peer-reviewed journal. It seems that, although it may be admissible for clinical researchers to publish the results of a single clinical trial, reimbursement decisions are, by definition, based on a synthesis of all the available data.

Although a full answer to the question posed here is beyond the scope of this article, it seems that a trial-based economic evaluation would only be acceptable as a basis for a reimbursement submission under very specific circumstances. These would include when the trial concerned represented the only effectiveness evidence on the health technology or that its design was considered far superior to previous studies. For example, it might be much larger or conducted under conditions more representative of regular clinical practice. These situations are quite rare, which suggests that in trial-based studies, there should always be some element of modeling or interpretation of results in relation to the broader range of clinical studies.

Certainly, in the United Kingdom, the independent teams providing technology assessment reviews (TARs) for NICE undertake a thorough systematic review of the clinical data as a basis for any subsequent economic modeling. Examples of a close link between systematic review and economic evaluation also exist in the published literature, although such examples are still relatively rare. (See the paper by Jefferson et al[8] for an example.)

Inappropriate Extrapolation Beyond the Period Observed in Clinical Studies

In the discussion of the study by Mark et al,[17] it was mentioned that benefits were extrapolated beyond the period observed during the GUSTO I clinical trial (which lasted for 1 year). Of course, some economic evaluations extrapolate from an intermediate end point to a final outcome. This was the case in the early studies of cholesterol-lowering therapies, which used an epidemiologic model to link reductions in serum cholesterol to reductions in congenital heart disease risk.[29]

The issue of extrapolation is probably the most debated point in economic submissions to reimbursement agencies, because only in a minority of cases (eg, studies of antibiotics) are the full benefits (and harms) of therapy observed during the period of the clinical trial.

In some cases, the need for extrapolation is self-apparent such as in diseases like diabetes, in which the major consequences occur in the long term. Here, the validity of extrapolation rests on the quality of the epidemiologic data, linking risk factors that can be modified by therapy to the long-term outcomes. Clearly, it would not be possible to wait 20 or 30 years for the definitive clinical trial demonstrating that modification of the risk factor led to a superior outcome.

However, in other cases, the justification for extrapolation and the methods for undertaking it are much more judgmental. For example, in chronic, disabling conditions like rheumatoid arthritis and Alzheimer disease, the clinical trials may last for only 1 year. In studies of therapies to prevent disease such as those of hypertension or cholesterol-lowering, the clinical trials may last for 5 years, but the benefits of therapy are likely to continue beyond the period of the trial. Therefore, a full analysis of the costs and benefits of therapy is likely to require some elements of extrapolation.

The view taken on extrapolation beyond the period of the trial can have a major impact on cost-effectiveness results. The most important judgment probably relates to the clinical benefit beyond the duration of the trial, but other important judgments relate to the incidence of rare side effects (not observed during the trial) and long-term compliance with therapy. The judgment on the latter may not be critical if lack of compliance, or discontinuation of therapy, means a reduction in cost of the therapy (eg, reduced drug consumption) as well as potential loss of benefits. However, in a study of combination antiretroviral therapy versus monotherapy in patients with HIV, Chancellor et al[29a] assumed that as soon as combination therapy ceased to have an advantage, it would be discontinued. This would not be the case unless the patients were closely monitored. There is, therefore, a risk that the economic advantages of combination therapy would be overstated.

The problems inherent in predicting future clinical benefits are well illustrated by the study by Schulman et al[30] on the cost-effectiveness of early zidovudine therapy for patients with HIV. The study used clinical data from the 019 trial, which was stopped on ethical grounds after 1 year, when it was observed that the patients randomized to zidovudine therapy had a slower progression to AIDS, as measured by CD4 counts. Schulman et al estimated the cost-effectiveness of zidovudine, compared with no therapy, based on 2 assumptions: 1) that the total benefit was only that observed during

the first year, this 1 year difference in the survival curves being projected into the future; and 2) that there was a continuous benefit of therapy, namely the benefit observed in the first year was repeated in subsequent years (ie, divergent survival curves).

In the absence of any other data, one might argue that the first assumption was reasonable (and probably conservative). However, longer-term studies showed that there was a "catch-up" effect such that after 3 years, there was no difference in life expectancy irrespective of whether the patient received zidovudine.

One lesson here is that the seemingly benign approach of projecting out the survival curves on a parallel basis can overstate the true benefits of therapy if there is any decline in effectiveness over time. This is also illustrated by the study by Sharma et al[31] on the cost-effectiveness of photodynamic therapy for subfoveal choroidal neovascularization secondary to age-related macular degeneration. Cost-effectiveness estimates based on a follow up of 2 years (the period observed during the clinical trial) ranged from $87,000 (Canadian) to $174,000 (Canadian) per QALY. The corresponding estimates for a follow up of 11 years were $43,000 (Canadian) to $87,000 (Canadian), the difference being mainly accounted for by the additional QALYs gained by projecting the area under the curve over a longer period.

Clearly, there are many different approaches to the projection of benefit beyond the trial. The most conservative would be the so-called "stop and drop" approach, in which no benefit is assumed beyond that observed. This may be appropriate in some diseases such as chronic renal failure, in which cessation of therapy is likely to result in death, but is probably inappropriate for most chronic diseases. Here the most likely possibility is a gradual decline in effect over time. In some cases, there may also be a "catch-up" effect when the patient discontinues therapy.

Sometimes, analysts modeling the same therapy have made different assumptions. For example, in 2 separate economic evaluations of donepezil for Alzheimer disease, O'Brien et al[32] assumed that, beyond the trial period, that the treatment effect was maintained, whereas Neumann et al[33] assumed that the curves jump back to the original point by varying the time horizon. As a result, the Neumann et al cost-effectiveness estimates are less optimistic for the drug.

Beyond recognizing the fact that different assumptions about the maintenance of treatment effect will generate different results, it is difficult to argue (in the absence of data) that 1 approach is more flawed than another. However, it might be possible to identify cases of extreme optimism. For example, in assessing the cost-effectiveness of beta-interferon in multiple sclerosis, Kendrick and Johnson[34] used a simple regression analysis to estimate a relationship between the clinical outcome measure (the Expanded Disability Status Scale) and time using data from a clinical trial. They then simply extrapolated by drawing a straight line out into the future. That is, they assumed that all patients would stay on therapy and would all continue to receive the level of benefit that they had received in the past. This contrasts with a model commissioned by NICE, which showed that a key assumption concerned what happened to patients once they discontinued beta-interferon, a parameter for which some data were already available.[35] The cost-effectiveness estimates for the Chilcott et al model were at least twice as high as those for the Kendrick and Johnson model.

Finally, another study by Caro et al[36] illustrates the difficulties of even defining the within-trial period. Their study was an economic evaluation of primary prevention with pravastatin for people with elevated cholesterol. Although the trial lasted for a considerable period of time (with an average follow up of approximately 5 years), only 10% of the benefit (in life-years gained) used in the cost-effectiveness estimates was actually observed during the period of the trial. Most of the benefit was extrapolated from events that occurred during the trial period (eg, unstable angina) but whose consequences would mainly manifest themselves in the future.

The approach taken by Caro et al was, in some respects, a within-trial analysis, because it considered only the costs and events occurring during the period of the trial. However, the benefits attributable to some of these events were projected into the future.

There are 2 problems with this approach. First, it is unreasonable to expect that drug therapy for people with elevated cholesterol would be stopped after the period of the trial. Therefore, the costs of therapy could be expected to continue into the future, apart from those patients who discontinue therapy over time owing to side effects or inconvenience. Second, for the cost-effectiveness estimates of Caro et al to be valid in the long-term, it would be necessary to assume that the benefits observed during the first 5 years would be repeated in the next 5 years, and so on. This may not be realistic.

From a methodological standpoint, it is important for analysts not to present just 1 set of cost-effectiveness estimates using a single method of extrapolation. Rather, a series of scenarios should be presented based on different extrapolation assumptions. This will provide an indication of how robust the cost-effectiveness results are to the extrapolation approach. When results are sensitive to the choice of method, the onus is on the decision-maker to identify which methods he or she considers the most reasonable.

Excessive Use of Assumptions Rather Than Data

One of the greatest attributes of economic analysis in any field is the ability to work with imperfect data. In many branches of economics, this attribute is borne out of necessity, because it is often difficult to conduct controlled

experiments (eg, into the impact of fiscal measures on the economy). However, in contrast, the biomedical field is characterized by extensive experimentation (eg, clinical trials) and large observational studies (eg, epidemiologic studies).

Therefore, the economist's approach, of making a series of assumptions backed up by sensitivity analysis, is often considered alien by those with a biomedical training. An example of the use of multiple assumptions is the study by Kristiansen et al[37] on the cost-effectiveness of a community intervention program to lower cholesterol in the Norwegian male population. The authors made 10 key assumptions regarding:

Participation in the community intervention program;
The number of physician visits per year for dietary treatment;
The reduction in serum cholesterol obtained by community interventions;
The effectiveness of dietary interventions;
The reduction in coronary heart disease achieved by a given reduction in serum cholesterol;
The proportion of MIs that were fatal;
The change in the rate of MIs over time (for a given population cohort);
The cost of a mass media campaign;
The reduction in quality of life through being identified as being at high risk of heart disease; and
The gain in quality of life from avoiding a nonfatal MI.

Although many of the assumptions were based on existing data and (for all we know) may have been reasonable, the extensive use of assumptions often raises concerns and may be viewed as a methodological flaw by those trained in the biomedical sciences.

However, it should be emphasized that the purpose of economic evaluation is to assist decision making at a point in time based on existing evidence. In some circumstances, particularly for new technologies which have hardly been used in routine practice, little or no data may exist for particular parameters, and the use of appropriately elicited expert opinion may be appropriate. However, it is crucial that the uncertainty that is then associated with these parameters is reflected in sensitivity analysis.

Inadequate Characterization of Uncertainty

Until relatively recently, the usual approach to handling uncertainty in economic evaluations was to conduct a *sensitivity analysis*. Here, the estimates of key parameters are varied individually or collectively to assess the sensitivity of the study result to the various assumptions.

Although almost all economic evaluations contain a sensitivity analysis, several reviews have shown that the approaches used are often inadequate.[38] The 2 main flaws are as follows. First, the choice of parameters to vary, and the range over which they are allowed to vary are often not adequately justified. Therefore, the authors of a study can give an appearance of stability in their findings by omitting important parameters from consideration or by only varying them by a small amount.

The ranges for variation of some parameters can easily be prescribed. For example, one might argue that an estimate of clinical effect size should be varied over a range corresponding to that bounded by the 95% confidence interval around the estimate. However, the ranges for other parameters such as the cost of a hospitalization are less easily determined.

The other major flaw is the failure to account for the combined effect of several parameters varying at the same time. That is, the majority of analysts conduct only a series of 1-way sensitivity analyses. Of course, whereas it may be possible to conclude that variation in individual parameters makes no difference to the overall result, the combined variation in a number of parameters could make a difference. Some approaches to sensitivity analysis such as multiway sensitivity analysis, threshold analysis, and scenario analysis partly deal with this problem. Another approach, called probabilistic sensitivity analysis is now becoming widely used in decision analytic modeling studies.[39] Here, probability distributions are applied to the specified ranges for the key parameters and samples drawn at random from these distributions to generate an empiric distribution of the cost-effectiveness ratio.

Fortunately, the analysis of uncertainty is 1 area in which economic evaluation has experienced important methodological developments in recent years. These include the analysis of stochastic data from economic clinical trials, the development of probabilistic models, and Bayesian interpretations of data, including the use of cost-effectiveness acceptability curves. (See Briggs[40] for a good review.)

Most of these approaches are now becoming the accepted standard, and it is likely that users of studies will have a much better appreciation of the uncertainty in study results in the future. However, it should be noted that probabilistic models still embody analysts' judgments about the range of variation in key parameters and their distributional form; hence, the importance, in modeling, or maintaining the transparency of methods.[41]

COMMON FLAWS IN INTERPRETATION OR REPORTING OF RESULTS

Problems in Aggregation of Results

Most economic evaluations, whether they are trial-based studies or modeling studies, usually present an aggregate result such as an incremental cost-effectiveness ratio (ICER) or a net benefit. Although such a synthesis is helpful, because it can be related to a given decision rule (eg, the ICER must be below a

TABLE 1. Average Cost-Effectiveness Ratios

	Costs (C)	Effects (E)	C/E
Asthma Drug A	€4500	15	€300
Asthma Drug B	€2750	10	€275

(Effects were measured in symptom-free days.) In the submission, the argument was put forward that "an extra €25 was not a lot to pay for the superior effectiveness." Of course, in an incremental analysis the extra symptom-free days are being bought at a cost of €350 each [(4500 − 2700)] ÷ 5].

given threshold), it also causes problems if the individual components of data are not reported.

The main problem is that the decision-maker may find it difficult to decompose the overall result, or may not fully appreciate the importance of each element of data or each assumption. That is, aggregation does not constitute a flaw in its own right; rather, the higher the degree of aggregation, the greater the chance of hiding various methodological flaws. For this reason, there has been considerable interest in developing reporting guidelines for economic evaluations in the interests of increasing transparency.[41–43]

Reporting Average Cost-Effectiveness Ratios

Many published economic evaluations report average cost-effectiveness ratios—that is, total costs divided by total effects for the 2 therapies being compared. Two problems arise here. First, for each therapy, the implied comparison is a mythical alternative with no costs and no effects. Second, one might be tempted to make a comparison between the 2 therapies based on these data.

Although many studies report average ratios, it is rare to find examples in which the ultimate sin (of implying a comparison) is committed. However, the following example was taken from a submission to the Transparency Commission in France (see Table 1). The details have been suppressed to protect the guilty.

Lack of Consideration of Generalizability Issues

This issue has 2 elements. First, results observed in clinical trials may not be achieved in regular clinical practice. Second, whereas clinical data may be transferable from place to place (eg, 1 geographic location to another), economic data may not.

The generalizability of data from an experimental setting to regular practice is widely discussed in economic evaluation. For example, there have been a number of papers discussing the problem in conducting economic evaluations alongside clinical trials (eg, protocol-driven costs). (See Glick et al[44] for a good review.) Also, it is common to see adjustments in decision analytic models to allow for lower compliance or lower efficacy in regular clinical practice. Whereas failure to make such adjustments would constitute a methodological flaw, economic analysts often have a better appreciation of this issue than their clinical colleagues.

The problems in transferring economic data from place to place has also been widely discussed. (See Sculpher et al[45] for a recent review.) However, the difficulties in interpreting economic data from another setting are not always adequately addressed. For example, the resources used during an inpatient stay in cardiology are likely to be much different between the United Kingdom, where clinical management is often conservative, and the United States, where there is a much greater use of investigational procedures. Of course, this could affect outcome as well as cost.

It would therefore constitute a methodological flaw not to address these points in the discussion and reporting of results, especially if they had not been addressed in the analysis. Many of the agencies requesting economic submissions require that the data and analyses reflect their own setting. This obviously applies to the unit costs (prices) used in the analysis, but also applies to the resource use estimates (because these reflect clinical practice patterns) and health state valuations for calculation of QALYs. Some agencies even request the economic models so that they can input their own data.

Selective Reporting and General Emphasis on Findings

Finally, in a situation of advocacy, manufacturers or sponsors of health technologies may be tempted to be selective in their reporting or to place an undue emphasis on particular findings. In general, this can only be overcome by decision-makers being more explicit about the analyses they would like to see, and doing their utmost to ensure transparency in methods and reporting. Another step forward would be to ban words like "substantial" and "minor" in the reporting of (say) the side effects of therapies.

GUARDING AGAINST THE MAJOR FLAWS

Although any methodological deficiency can, in principle, compromise the results of an economic evaluation, some flaws are clearly more fundamental than others.

First, failure to include all relevant alternatives can seriously bias study results. Therefore, it is important that the analyst demonstrates that he or she has considered all alternatives or can justify why some have been omitted. Also, an economic evaluation based on a clinical trial comparing only 2 therapies would not be suitable for decision-making purposes unless these were the only 2 therapies available.

Second, major errors can be incorporated through the use of inappropriate methods for identifying and synthesizing the effectiveness evidence. Therefore, it is important that a systematic review is conducted to locate all the relevant evidence and that data on the main parameters are synthe-

sized in a way that recognizes, and accounts for, the potential biases in making indirect comparisons.

Third, given the inevitable uncertainties in the data and methods of economic evaluation, it is important that uncertainty is characterized adequately. The preferred approach is to undertake a probabilistic sensitivity analysis (PSA), because this considers all the input parameters simultaneously, provides a complete picture of the joint parameter uncertainty, and gives a summary measure of the implications of uncertainty. This includes a careful consideration of the uncertainty associated with "structural" assumptions within a model such as those relating to extrapolation over time. In principle, this form of uncertainty could be included in a PSA, but to aid explicitness with decision-makers, reporting multiple scenarios may be preferable.

Finally, errors can be introduced by using data that are not relevant to the setting that is under consideration. Therefore, issues relating to the generalizability of data need to be recognized and, if necessary, adjustments made. Issues of generalizability should also be taken into account when assessing whether the results of a given study apply to other settings.

CONCLUDING REMARKS

This article has attempted to identify the major flaws in the design, analysis, and reporting of economic evaluations and, when possible, to suggest improvements. No study is likely to be completely free of flaws, so it is important that study users know how to detect them. As an initial contribution to this process, a list of potential flaws and questions for users to ask is given in Table 2.

In exposing all the potential flaws in economic evaluations, the danger is that the perfect will be the enemy of the merely good. Therefore, it is important to recognize that

TABLE 2. Detecting Flaws in Economic Evaluation

Potential Flaw	Questions for Users to Ask
Omission of important costs or benefits	Given the chosen study perspective, is it likely that any of the omitted costs or benefits have a big impact on study results?
Selection of an inappropriate alternative for comparison	Are any relevant alternatives omitted?
	Is the alternative selected for comparison an inefficient treatment option?
Biases in synthesizing clinical data	Where effectiveness estimates are based on synthesis (eg, metaanalysis), have potential differences in the clinical studies been recognized and allowed for?
Inadequate representation of the effectiveness data	Are all the available clinical studies used as a basis for the cost-effectiveness study?
	If some studies are excluded, is this decision justified?
Inappropriate extrapolation beyond the period observed in clinical studies?	Is the time horizon for the economic study adequately justified?
	In any extrapolation, are both "optimistic" and "pessimistic" scenarios explored?
Excessive use of assumptions rather than data	Are any data available to substantiate the major assumptions in the study?
	Is the impact of major assumptions explored through sensitivity analysis?
Inadequate characterization of uncertainty	Is some measure of precision presented for the main study parameters (eg, a standard error)?
	Does the analyst go beyond simple, 1-way sensitivity analysis?
Inappropriate aggregation of results	Are the component parts of the cost-effectiveness ratio, or net benefit estimate, presented?
Reporting average cost-effectiveness ratios	Are all the comparisons between the alternatives expressed in incremental form?
Lack of consideration of generalizability issues	Does the analyst *assume* that the study results apply in other settings?
	What are the main differences, between the study setting and others, that would affect the cost-effectiveness results?
Selective reporting and general emphasis on findings	Are words like "substantial" and "minor" added as descriptors?
	Is the discussion of study results evenhanded, or is undue emphasis placed on the more positive results?

most, if not all, reimbursement decisions are made at a time when only imperfect information is available. Therefore, the appropriate way to judge economic evaluations is not whether they embody some ultimate "truth," but whether they lead to a better decision than would have been made in their absence.

ACKNOWLEDGMENTS

This paper was originally presented at a meeting of the Agency for Healthcare Research and Quality (AHRQ) Workgroup on Infrastructure and Systems for CEA review, Washington DC, May 21–23, 2003. The authors are grateful to the meeting participants for comments received. Finally, the authors are grateful to 3 anonymous referees, whose comments greatly improved the paper.

REFERENCES

1. Commonwealth Department of Human Services and Health. *Guidelines for the Pharmaceutical Industry on Preparation of Submissions to the Pharmaceutical Benefits Advisory Committee including Major Submissions Involving Economic Analysis.* Canberra: Australian Government Publishing Services; 1995.
2. Ontario Ministry of Health. *Guidelines for Economic Analysis of Pharmaceutical Products.* Toronto: Ministry of Health; 1994.
3. National Institute for Clinical Excellence (NICE). *Guide to the Methods of Technology Appraisal.* London: NICE; April 2004.
4. Drummond MF. *The Relevance of Health Economic Evaluation to Health Care Decision Makers. Spectrum Report.* Waltham, MA: Decision Resources Inc; 2002.
5. Gerard K, Smoker L, Seymour J. Raising the quality of cost-utility analyses: lessons learnt and still to learn. *Health Policy*. 1999;46:219–238.
6. Gold MR, Siegel JE, Russell LB, et al. *Cost-Effectiveness in Health and Medicine.* New York: Oxford University Press; 1996.
7. Jefferson T, Demicheli V. Quality of economic evaluations in healthcare: is it time for action to ensure higher methodological quality? *BMJ*. 2002;324:313–314.
8. Jefferson T, Demicheli V, Vale L. Quality of systematic reviews of economic evaluations in health care. *JAMA*. 2002;287:2809–2812.
9. Neumann PJ, Stone PW, Chapman RH, et al. The quality of reporting in published cost-utility analyses, 1976–1997. *Ann Intern Med*. 2000;132:964–972.
10. Phillips KA, Chen JL. Impact of the US panel on cost-effectiveness in health and medicine. *Am J Prev Med*. 2002;22:98–105.
11. Kassirer JP, Angell M. The journal's policy on cost-effectiveness analysis. *N Engl J Med*. 1994;331:669–670.
12. Hillman AL, Eisenberg JM, Pauly MV, et al. Avoiding bias in the conduct of cost-effectiveness research sponsored by pharmaceutical companies. *N Engl J Med*. 1991;324:1362–1365.
13. Friedberg M, Saffron B, Stinson TJ, et al. Evaluation of conflict of interest in economic analyses of new drugs used in oncology. *JAMA*. 1999;282:1453–1457.
14. Hill SR, Mitchell AS, Henry DA. Problems with the interpretation of pharmacoeconomic analysis: a review of submissions to the Australian Pharmaceutical Benefits Scheme. *JAMA*. 2000;283:2116–2121.
15. Sheldon TA. Problems of using modelling in the economic evaluation of health care. *Health Econ*. 1996;5:1–11.
16. Claxton K, Sculpher M, Drummond MF. A rational framework for decision making by the National Institute for Clinical Excellence. *Lancet*. 2002;360:711–715.
17. Mark DB, Hlatky MA, Califf RM, et al. Cost effectiveness of thrombolytic therapy with tissue plasminogen activator as compared with streptokinase for acute myocardial infarction. *N Engl J Med*. 1995;332:1418–1424.
18. Johannesson M, O'Conor RM. Cost-utility analysis from a societal perspective. *Health Policy*. 1997;39:241–253.
19. Canadian Coordinating Office for Health Technology Assessment (CCOHTA). *Guidelines for Economic Evaluation of Pharmaceuticals*. Ottawa: CCOHTA; 1994.
20. O'Brien BJ, Anderson DR, Goeree R. Cost-effectiveness of enoxaparin versus warfarin prophylaxis against deep-vein thrombosis after total hip replacement. *Can Med Assoc J*. 1994;15:1083–1090.
21. Palmer S, Sculpher M, Philips Z, et al. A cost-effectiveness model comparing alternative management strategies for the use of glycoprotein IIb/IIIa antagonists in non-st-elevation acute coronary syndrome. Report to the National Institute for Clinical Excellence. Available at: http://www.nice.org.uk/Docref.asp?d=32030. 2002.
22. Song F, Altman DG, Glenny A-M, et al. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ*. 2003;326:472.
23. Wilby J, Kainth A, Hawkins N, et al. A rapid and systematic review of the clinical effectiveness, tolerability and cost effectiveness of newer drugs for epilepsy in adults. Technology Assessment report commissioned by the HTA programme on behalf of the National Institute for Clinical Excellence. In press.
24. Ades AE. A chain of evidence with mixed comparisons: models for multi-parameter evidence synthesis and consistency of evidence. *Stat Med*. 2003;22:2995–3016.
25. Jönsson B, Bebbington P. Economic studies of the treatment of depressive illness. In: Jonsson B, Rosenbaum J, eds. *Health Economics of Depression: Perspectives in Psychiatry*, vol 4. Chichester: Wiley; 1993:1–13.
26. Dunbar GC, Cohn JB, Fabre LF, et al. A comparison of paroxetine, imipramine and placebo in depressed patients. *Br J Psychiatry*. 1991;159:394–398.
27. Freemantle N, Maynard A. Something rotten in the state of clinical and economic evaluations? *Health Econ*. 1994;3:63–67.
28. Jönsson B. Economic evaluation and clinical uncertainty: response to Freemantle and Maynard. *Health Econ*. 1994;3:305–307.
29. Oster G, Epstein A. Cost-effectiveness of anti-hyperlipidemic therapy in the prevention of coronary heart disease: the case of cholestyramine. *JAMA*. 1987;258:2381–2387.

29a. Chancellor JV, Hill AM, Sabin CA, et al. Modelling the cost effectiveness of lamivudine/zidovudine combination therapy in HIV infection. *PharmacoEconomics*. 1997;12:1–13.

30. Schulman KA, Lynn LA, Glick HA, et al. Cost effectiveness of low-dose zidovudine therapy for asymptomatic patients with human immunodeficiency virus (HIV) infection. *Ann Intern Med*. 1991:798–802.
31. Sharma S, Brown GC, Brown M, et al. The cost-effectiveness of photodynamic therapy for fellow eyes with subfoveal choroidal neovascularization secondary to age-related macular degeneration. *Ophthalmology*. 2001;108:2051–2059.
32. O'Brien BJ, Goeree R, Hux M, et al. Economic evaluation of donepezil for the treatment of Alzheimer's disease in Canada. *J Am Geriatr Soc*. 1999;47:570–578.
33. Neumann PJ, Hermann RC, Kuntz KM, et al. Cost-effectiveness of donepezil in the treatment of mild or moderate Alzheimer's disease. *Neurology*. 1999;52:1138–1145.
34. Kendrick M, Johnson KI. Long-term treatment of multiple sclerosis with interferon-beta may be cost-effective. *Pharmacoeconomics*. 2000;18:45–53.
35. Chilcott J, McCabe C, Tappenden P, et al. Modelling the cost effectiveness of interferon beta and glatiramer acete in the management of multiple sclerosis. *BMJ*. 2003;326:522.
36. Caro J, Klittich W, McGuire A, et al. The West of Scotland coronary prevention study: economic benefit analysis of primary prevention with pravastatin. *BMJ*. 1997;315:1577–1582.
37. Kristiansen IS, Eggen AE, Thelle DS. Cost-effectiveness of incremental programmes for lowering serum cholesterol concentration: is individual intervention worthwhile? *BMJ*. 1991;302:1119–1122.
38. Briggs A, Sculpher M. Sensitivity analysis in economic evaluation: a review of published studies. *Health Econ*. 1995;4:355–371.
39. Claxton K, Sculpher M, McCabe C, et al. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. *Health Econ*. 2005;14:339–347.

40. Briggs A. Handling uncertainty in economic evaluation and presenting the results. In: Drummond MF, McGuire A, eds. *Economic Evaluation in Health Care: Merging Theory With Practice*. Oxford: Oxford University Press; 2001:172–214.
41. Weinstein MC, O'Brien BJ, Hornberger J, et al. Principles of good practice for decision analytic modelling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices—modelling studies. *Value in Health*. 2003;6:9–17.
42. Drummond MF, Jefferson TO, for the BMJ Economic Evaluation Working Party. Guidelines to authors and peer-reviewers of economic submissions to the *British Medical Journal.* 1996;313:275–283.
43. Drummond MF, Brown R, Fendrick AM, et al. Use of pharmacoeconomics information—report of the ISPOR task force on use of pharmacoeconomic/health economic information in health care decision-making. *Value in Health*. 2003;6:407–416.
44. Glick HA, Polsky DP, Schulman KA. Trial-based economic evaluations: an overview of design and analysis. In: Drummond MF, McGuire A, eds. *Economic Evaluation in Health Care: Merging Theory With Practice*. Oxford: Oxford University Press; 113–140.
45. Sculpher MJ, Pang FS, Manca A, et al. Generalisability in economic evaluation studies in health care: a review and case studies. *Health Technol Assess*. 2004;8:1–206.