USES OF MODELS IN ECONOMIC EVALUATIONS OF MEDICINES AND OTHER HEALTH TECHNOLOGIES

Brian Rittenhouse

Office of Health Economics
12 Whitehall London SW1A 2DY

USES OF MODELS IN ECONOMIC EVALUATIONS OF MEDICINES AND OTHER HEALTH TECHNOLOGIES

Brian Rittenhouse

Office of Health Economics
12 Whitehall London SW1A 2DY

Office of Health Economics

The Office of Health Economics was founded in 1962 by the Association of the British Pharmaceutical Industry. Its terms of reference are:

To undertake research on the economic aspects of medical care.

To investigate other health and social problems.

To collect data from other countries.

To publish results, data and conclusions relevant to the above.

The Office of Health Economics welcomes financial support and discussions on research problems with any persons or bodies interested in its work.

About the Author

Brian Rittenhouse has an AB in Economics from Oberlin College and MS and PhD degrees in Economics from the University of Wisconsin – Madison. He has worked in the field of economic evaluation of medical interventions as an employee of CIBA-Geigy, AG and the University of North Carolina. Currently he is an Associate Professor and Rhône-Poulenc Rorer Chair of Pharmacoeconomics at the Université de Montréal.
Faculté de médecine
Département de pharmacologie
Université de Montréal
CP 6128, Succ. Centre-ville
Montréal, PQ H3C 3J7
Canada

ISBN 1 899040 00X

Printed by BSC Print Ltd, London

Foreword

The use of models in economic studies evaluating medicines and other health technologies has become a controversial issue. This is because study results now matter – decision makers are increasingly acting on information about the cost effectiveness of treatments.

This publication is intended to introduce the key elements of modelling and to explain, with examples, the role that good modelling should play in an economic evaluation. It discusses some of the controversies around the use of models but is not intended to be a guide to this debate. Its purpose is to equip the reader with an understanding of the basic concepts of, and the main uses of, modelling in economic evaluation.

The author, Brian Rittenhouse, argues that we use the term 'model' in two quite different ways. In the first sense we take a model to be any artificial simplification of reality designed to enable us to better understand the world. A road map would fall into this category, as would a randomised controlled trial. It is the second meaning of the term model that is more controversial – where the simplification of reality includes the use of techniques to combine data from different sources, and, usually, the use of assumptions to enable extrapolation from the combined data or to fill gaps within the required data set.

After categorising types of model and introducing us to decision analysis, he then addresses the need to have information on effectiveness rather than efficacy. Most randomised controlled trials are designed to demonstrate efficacy, whereas decision makers need to know about effectiveness in clinical practice. This is particularly the case for pharmaceuticals, where pre launch trials are designed to meet regulatory requirements for safety and efficacy evidence, leading to study designs with high internal validity, but, often, limited external validity. Some of the deficiencies of these trials can, in principle, be dealt with by changes in trial design, others cannot. Modelling can provide a way of turning good efficacy and cost-efficacy studies into good cost-effectiveness analyses.

Of course, concerns about potential bias in data taken from non RCT based studies (see Sheldon (1994)) do have to be addressed, and Rittenhouse discusses sources of bias, hierarchies of evidence, and the extent to which sensitivity analysis and other ways of handling uncertainty can help decision makers understand the potential variation in outcome. He concludes that while sensitivity analysis is valuable it is no substitute for addressing concerns about bias prior to producing the central result of the study.

As Rittenhouse acknowledges, modelling is not without its drawbacks. Readers who wish to follow on from this publication to read more about some of the controversies surrounding the use of modelling, will find a well argued case for the appropriate use of modelling in Luce (1995) and in Gold et al (1996). A thoughtful review of the issues is set out in Buxton et al (forthcoming), and a note of scepticism, restating the case for society to invest on the collection of good evidence from randomised controlled trials, is contained in Sheldon (1996).

Rittenhouse is clear that modelling is a valuable, integral, and permanent feature of economic evaluations and can be of good quality. I hope you find his introduction to the concepts and role of modelling of interest, and that it will stimulate you to read more about modelling techniques and about the debates surrounding its role in economic evaluations and in decision making.

Adrian Towse
Director of the Office of Health Economics

References

Buxton M J et al (forthcoming). Modelling in Economic Evaluation: An Unavoidable Fact of Life. Submitted to the Journal of Clinical Epidemiology.
Gold M R et al (ed) (1996). Cost-Effectiveness in Health and Medicine. Oxford University Press.
Luce B R (1995). Policy Implications of Modelling the cost-effectiveness of health care technologies. Drug Information Journal. Vol 29 p1469-1475.
Sheldon T A (1994). Please Bypass the PORT. British Medical Journal. Vol 309 p142-143.
Sheldon T A (1996). Problems of using modelling in the economic evaluation of health care. Health Economics. Vol 5 p1-11.

Contents

1. Introduction

The word 'model' has many meanings to many people. Even within the policy analytic literature we will see that there are several meanings and types of models. For economic evaluation of medicines in particular modelling is currently in vogue. Why is this the case? What does the term mean? Why do we need models? Will we always need them? This monograph will provide one set of answers to these and other related questions.

There are at least two broad meanings of the term model that are used in this context. The first is very general. A model can be merely a simplified version of reality that is used to describe the essential elements of a real situation such as the economic implications of using a particular therapy. The point of such a model is to describe enough of the essential elements of reality to enable cost effective decision making. A more mundane example of this type of a model is a road map. It provides a simplified version of reality that describes essential elements to drivers. In this sense any economic evaluation will always use a model. Even a randomized controlled trial (RCT) is a model – measuring only particular aspects of a patient's experience with treatment. Such models will always be part of economic evaluations.

The second and more interesting meaning of the word model has to do with the type of data used in a model of the first type. We may model the outcome of an intervention in the sense that we do not measure it directly. Thus we may measure a given level of blood pressure reduction, but will be interested in aspects not measured, such as what resources they consumed, and what were the health or health-related quality of life implications? Many of these variables will not conventionally be measured because of cost and/or time concerns. However, for purposes of a complete economic analysis, some concept of final outcome and cost is essential. If these variables are not measured, they must be modelled. This can mean many things along a continuum of evidence from casual conjecture to rigorously precise epidemiological studies to extrapolate (model) from a clinical endpoint to a final outcome, such as years of life saved.

Traditionally, models have been used in economic evaluations often to compensate for lack of foresight (or interest) on the part of clinical investigators in pursuing economic goals. This is changing somewhat. As RCTs include more and more variables of interest to economists, it may appear that modelling, in the second sense of the usage here, may become needless. However, certain aspects of RCTs imply that this will not be the case. RCTs, even when they are designed to collect economic data, fall short of representing the reality of health outcomes and resource utilization to be expected in routine clinical practice in the same indication area. Models can be built to correct these problems. This paper will argue that this economic 'modelling', in the second sense of our meaning is currently, and will in the future, remain essential to the practice of the economic evaluation of medicines. That is, far from being a temporary fix for the lack of adequate planning for economic evaluations that it is often thought to be, modelling is a essential component of any accurate economic analysis.

2. Why are we interested in the economics of medical interventions?

'The health care industry is a $900+ billion endeavour that does not know how to measure its main product: health'
– D Fryback (1993)

The tremendous growth over the past century of the ability to influence disease prognosis has led to a revolution in thinking about medical resource allocation. In years past physicians were limited in the armamentarium at their disposal to fight disease. Their ability to intervene to any significant effect was restricted. This is (emphatically) no longer the case for many diseases. While cure is still often elusive, many things can be done for patients to increase the probability of successful prevention, diagnosis or cure. Old decision criteria must be reassessed in such a situation where medical care can now consume as much of society's resources as it chooses to devote to the task. There is no limit (other than financial) to the ability to spend to improve outcomes.

The traditional approach that dictated the allocation of medical resources (at least in spirit) was based on medical principle. Some have called this the 'rule of rescue' (Dworkin, 1994). The implications of this approach were that if something was likely to have a positive influence on outcome, even with low probability, then it was done. In an era where opportunities to intervene were practically limited by technology and knowledge, this provided relatively few financial difficulties. Now however, such a philosophy implies quite a different picture. Without any conceptual limit on ability to spend to positively influence health outcomes, how are individuals or society to make choices in the allocation of resources?

One approach is to evaluate health technologies in terms of their economic efficiency – their productivity in achieving health. This approach dictates that one should consider both the costs and consequences of alternative activities in allocating resources and choosing among them. What are the consequences/benefits of particular actions? What does it cost to bring about these particular consequences? Perhaps most often overlooked is the important concept of the *value* of the additional consequences brought about. The consequences themselves are only part of the story. Ultimately economics is interested in the *value* of consequences compared to the costs of achieving them. Setting priorities to squeeze more value out of a given set of resources essentially means achieving more health for the same budget.

How do economists (or others using the philosophy of economics) produce evidence on outcomes and costs of achieving them to guide economic decisions of individuals or policy makers? How do they measure outcomes and the relative value of those outcomes to inform decision makers so that priorities can be set to bring about the highest level of population health with a given set of resources? In a general sense, they build economic models in both our meanings of the term. Here we will describe some of the concepts used in such models.

3. What is a model?

'Don't ask what it means, but rather how it is used.'

– L. WITTGENSTEIN (Freedman et al, 1991, p. 475)

3.1 A simplified version of reality

As we noted the term model has two main meanings. In the first, more general meaning, a model is a simplified rendering of the real world used to capture the *essence* of reality while dispensing with much of the complicated excess baggage that accompanies reality, but makes analysis difficult. A model is intended to assist in decision making. As such, it should simplify the relevant issues to their minimal level of complexity (but no further!), retaining the essential components in order to remain useful as a tool for assisting in decision making. There can be an art to the simplification process that is essential to model development. Oversimplifying reality in the interest of making a problem more easily solved leads to inaccuracy. *Under*simplifying reality can lead to confusion or an inability to resolve the problem.

When results from an RCT are used for decision making, at least implicitly, decision makers have a model at work. In assessing relative performance over placebo or active control arm in a pharmaceutical trial, particular efficacy and safety endpoints are chosen and others are either given lower priority or are ignored. This is a simplification of reality necessary to enable decisions to be made in a timely, organized and cost-effective manner. Of course these simplifying assumptions can be wrong. Brody (1995) reports that in the early results of the TIMI-I trial of tPA v. streptokinase in thrombolytic therapy, patency of the infarct-related artery at 90 minutes after treatment was used as an intermediate endpoint, indicating marked superiority for tPA (see discussion of a model of these results in Appendix II, Box 7). Later results, using mortality data, indicated much more marginal superiority (The GUSTO Investigators, 1993). Re-analysis of the data from a Bayesian statistical perspective indicated that even those marginal (yet potentially important) differences were probably illusory (Brophy and Joseph, 1995).

3.2 Different types of models as versions of reality

An economic model uses simplification – the essential details are included, others are ignored or given lesser weight. What outcomes or costs receive primary attention will depend on the specific context of the evaluation. A model might leave out low probability and low risk events that might marginally influence outcomes. By definition these events are unlikely to have a major impact. Using them in the model will merely complicate the analysis without providing any corresponding benefit in terms of enhanced accuracy.

In an introductory text on policy analysis, Stokey and Zeckhauser (1978) describe several generic types of model, listed in Table 1. Descriptive models attempt to present a framework for prediction of what the results of some process will be (e.g. if one goes to the doctor, one will recover from an illness; if one studies effectively, one will learn). Prescriptive models on the other hand purport to indicate an optimal course of action rather than simply describe and predict action. Such models provide counsel (e.g. if one is ill, one should go to see a physician). Imbedded in a prescriptive model is always a descriptive model (e.g. one sees a physician because one is expected to recover more easily/rapidly because of such actions). Before one can make a prescriptive recommendation, one must know the consequences of alternative choices. Such consequences form the descriptive model within the context of the prescriptive model.

Table 1 General types of models

Descriptive:	Describe what will happen
Prescriptive:	Suggest optimal course of action
Deterministic:	Events happen with certainty
Stochastic:	Events happen with specified probabilities

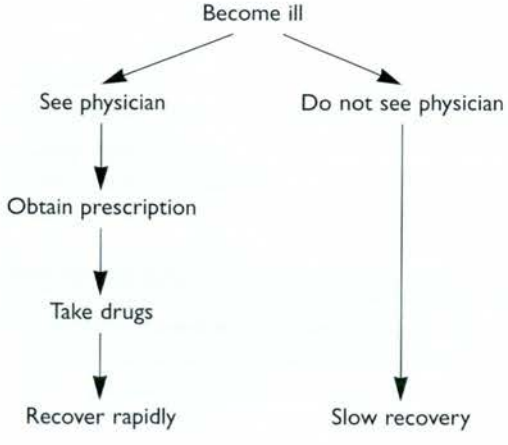
Source: Stokey and Zeckhauser, 1978

One must also have an objective in order to use a prescriptive model. In the above example, the objective may be to recover from illness. Some objectives are more clearly appropriate than others. This is particularly the case when other constraints enter the picture, such as the costs of doing trials. Primary endpoints are chosen for sample size, time and cost reasons sometimes. Often such endpoints are intermediate in nature (e.g. blood pressure reduction). Ultimately the objective is not the achievement of these intermediate endpoints, but some more valuable final outcome. Whether the final outcome is sufficiently associated with these measured intermediate endpoints is sometimes questionable.

Even were the link a strong one, there may still remain conflicts over what the appropriate final

outcome is. In chronic disease management, the probabilistic effect on health outcome must be weighted against the possibility of reduced quality of life due to side effects of long term drug therapy. In a cancer patient, recovery may not be the sole concern. Quality of life may be an additional important factor. Possibly conflicting objectives require a choice or a way (another model) of combining objectives (e.g. quality and length of life). Here one may see the potential conflict between chosen assumptions in a prescriptive model. Simplifying reality must occur in any model. The question is whether reality has been oversimplified or if the process of simplification is legitimate (we will see later a debate on this in the context of quality and length of life and whether particular concepts in efforts to combine length and quality of life are legitimate).

Figure 1 A deterministic, prescriptive model



Deterministic and stochastic models can each be embedded in either of the above types of models. The former makes statements with certainty, the latter with probability. Thus, deterministic models are implicit in the brief descriptions above, while stochastic models would be represented differently. Figure 1 shows a very simple prescriptive deterministic model to help determine whether one should see a physician when one is ill. There are two choices with deterministic outcomes associated with each of them. Depending on the perceived relative value of early and late recovery one chooses the optimal course of action. One might also associate costs with each of the alternative actions and tradeoff the added value versus the costs of early recovery associated with the physician visit.

A stochastic prescriptive model might state that if one studies, one is *likely* to learn (perhaps associating a probability with this outcome as well as with its complement – that one might not learn). Such a model might say that if one is ill, one should or should not visit a physician, the decision being based on the interplay of likely consequences if one either does or does not make the visit. In Figure 1 we could have incorporated stochastic elements at various points to make the model more realistic. For example, we could have made it probabilistic rather than deterministic that one would be given drugs by the physician or that they are taken properly or that recovery is actually achieved. The outcomes of such a model are probabilistic only, and one must entertain the possibility that one's a priori decision will be shown ex post to have been in error. One makes decisions based on relative probabilities, outcomes and the values of the outcomes. Box 1 (page 10) illustrates the making of decisions using the mathematical concept of 'expected value', an average outcome that is in fact 'expected' only in the sense of its being an average if the situation were played out many times. Nevertheless, the expected outcomes from various possible actions can be quite useful in decision making under conditions where there is uncertainty in consequences of different actions. These methods are often referred to as 'decision analytic' models.

What are some of the advantages of models? We have already mentioned the great advantage of simplifying reality to facilitate prediction or prescription. Dealing only with simplified versions of reality allow us to examine essential features of reality without much of the (hopefully) superfluous details so often associated with reality. Thus a map is a model of real topographical features of a landscape or of roads that cover an area. Maps are simplifications of reality, and depending on the use to which they are to be put, they may suffice to describe reality in ways that are useful without requiring all the details of reality.

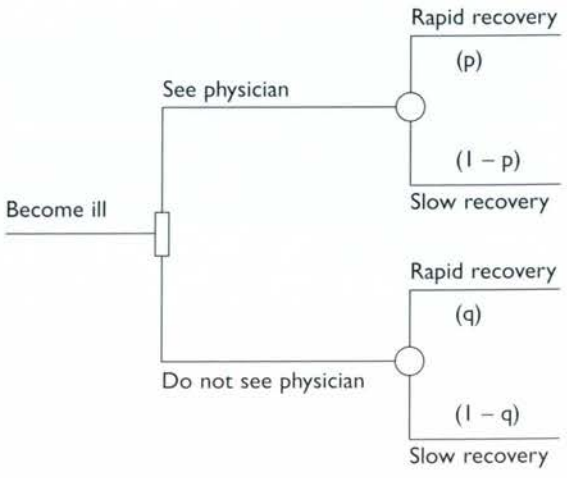
Another advantage provided by a model is that it serves what Stokey and Zeckhauser (1978) call a bookkeeping function. It describes all assumptions that the model uses and thereby helps an analyst organize and keep track of the various components of the analysis. This same property has another useful attribute. It facilitates a communication function at the same time by explicitly detailing these assumptions for those to whom the model is

presented. Box 3 on page 17 indicates the organising function of models. Decision analytic models in particular tend to be quite helpful in indicating the organisation and structure of the analysis as their structure is laid out diagrammatically. We will see several examples of such models in this monograph.

It is stochastic, prescriptive models that are most frequently of use in the economic evaluation of medicines, and it is these that will be the focus of all the examples provided in this discussion (sometimes with descriptive models embedded within them). The practice of medicine is uncertain; indeed that fact does much to explain why economists and epidemiologists and others are spending so much of their time on these issues. The inherent uncertainty in the field demands careful attention to detail in providing models to assist decision makers. Furthermore, it is often a course of optimal action that we seek to determine, not merely to describe what people might choose to do. Thus it is not descriptive but prescriptive models that are the goals of most of the economic evaluations of pharmaceuticals. The very lack of adequate information so inherent in the medical field makes the reliance on description somewhat unsatisfactory and incomplete. We are after answers to what we *should* do, not predicting what people *will* do – should a particular drug be used in a certain type of patient?; should the price be reimbursed by the insurance system (implicitly asking whether there is sufficient value obtainable from the product to justify the outlay)?

Figure 2 shows a decision analytic model in the form of a 'decision tree' as part of a prescriptive, stochastic model. The decision in this case is whether to see a physician or not when one becomes ill. Of course, such a decision will depend on the specifics of the illness, but a general model will serve to indicate the process. The model is prescriptive in that it will indicate which decision is optimal given the objective, in this case to recover from the illness. Decisions are represented in trees as squares with the possible decisions emanating from the squares, in this case there are only two possible decisions – see a physician or not. These squares are known as decision nodes. Circles denote chance nodes. When a circle appears, the outcome is not decided by a decision but by chance. At each chance node the branches emanating will have probabilities associated with them. The probabilities of all the branches at

Figure 2 A stochastic, prescriptive model



each chance node must sum to equal one; that is they must collectively exhaust all possibilities – all possibilities must be accounted for. Here one either recovers or does not. With two branches to the tree, if the probability of rapid recovery is denoted by 'p', the probability of slow recovery can be written as '1 - p'. Since there are two alternative actions, there should be two different sets of probabilities. We denote the probabilities of recovery from each of these actions as p and q.

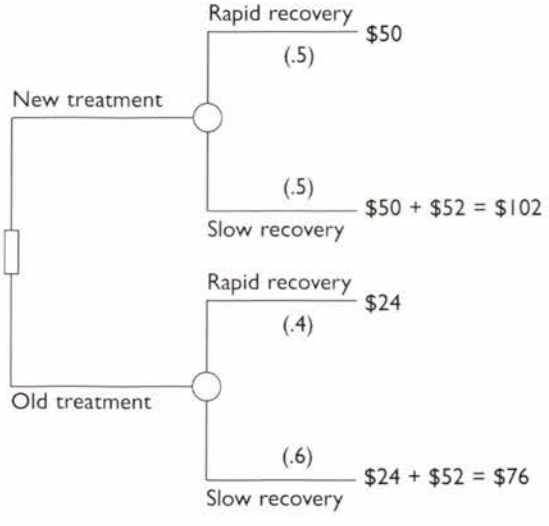
One may imagine chance nodes with several other possibilities. Each branch coming from a chance node must also be mutually exclusive, that is, there is no overlap between outcomes at chance nodes. These events (recover rapidly or slowly) are mutually exclusive and (assumed to be) collectively exhaustive. A chance node with only the recover option emanating from it would not be exhaustive (unless recovery occurred for all persons, in which case it is not properly a chance event, though sometimes for clarity of depiction certainties are presented with probabilities of one); a chance node with recover as one branch and experience an adverse event at another and not recover at a third would contain the possibility of overlap. One could recover with an adverse reaction for example. In this example, it is assumed that all patients will recover. If that is not the case the model would need to be modified.

If we are interested in maximizing the chance of rapid recovery in this example, we will choose the option with the highest probability of that outcome. This model can be expanded to include a choice of

Box 1 Decision making based on expected values of outcomes

In Figure 3 below, we represent a simple choice between treatments, new and old. The problem is represented as a decision tree. The new treatment has evidence supporting a 0.5 probability of producing a rapid recovery as compared to the old treatment probability of only 0.4. The only other possibility entertained in the model is that a slow recovery occurs

Figure 3 Simple decision tree



which is assumed to be the equivalent of not having received treatment. The new drug costs $50; the old only $24. If either drug does not work, a further $52 is spent in palliative care during the slow recovery. The model developers are solely interested in minimizing the costs of the treatments in this case. That is, they care nothing for the possibility of improved outcome outside of the possibility that such improved outcome reduces costs (it should be clear by now that this is not a proper economic analysis since it ignores outcomes and their valuation; nonetheless, this type of analysis is sometimes used, and it serves to illustrate the basic point of expected value decision making).

Expected costs are the outcome criteria for evaluating this model. 'Expected' in this sense is a very special use of the word that is somewhat counterintuitive in that, at least for any given patient, such costs as are 'expected' will in fact never occur. 'Expected' costs refer to a more global perspective of an average cost if many patients were to be treated under each of the two alternative actions. Thus it is expected in the sense of a population perspective. If we were to choose one particular treatment, its expected costs would be expected in the conventional sense to be the average of the patients exposed to this treatment.

The expectation is a mathematical operation that is simply a weighted average of the outcomes, where the weights are the probabilities associated with the outcomes. The cost outcomes are written into the decision tree above. Anyone receiving the new or old drug incurs a cost of $50 and $24 respectively. In addition, anyone who does not achieve rapid recovery receives the palliative care during the slower recovery they do experience. This will add a further $52 to the total treatment costs of those individuals.

The expected cost of the new treatment is the weighted average of the two possible cost outcomes weighted by their probabilities:

EC (new) = (0.5) ($50) + (0.5) ($50 + $52) = $25 + $51 = $76

Similarly, the expected cost of the old drug is:

EC (old) = (0.4) ($24) + (0.6) ($24 + $52) = $9.60 + $45.60 = $55.20

The old drug has a lower expected cost and apparently should therefore (by our criteria) be the drug of choice. However, in addition to ignoring the outcomes of these treatment alternatives, the analyst has ignored another factor that is more important because it affects the optimality of his or her decision, even based on the limited considerations of budget that is here assumed to be the relevant assessment criterion. Note that, by assumption, treatment failure implies slow recovery which is the same as the result if no treatment were applied at all. The costs associated with slow recovery are assumed to be $52. It would be cheaper to not treat anyone and pay this cost than to choose the 'cheapest treatment' that has an expected cost of $55.20.

This example indicates the necessity of examining all relevant alternatives, a suggestion often ignored (through ignorance or through conscious strategy). Canada's guidelines for the economic evaluation of pharmaceuticals (CCOHTA, 1994) reflect this approach stating that the relevant comparison is 'existing treatment' (which is carefully defined so that it is not just any existing treatment) and 'minimal treatment.'

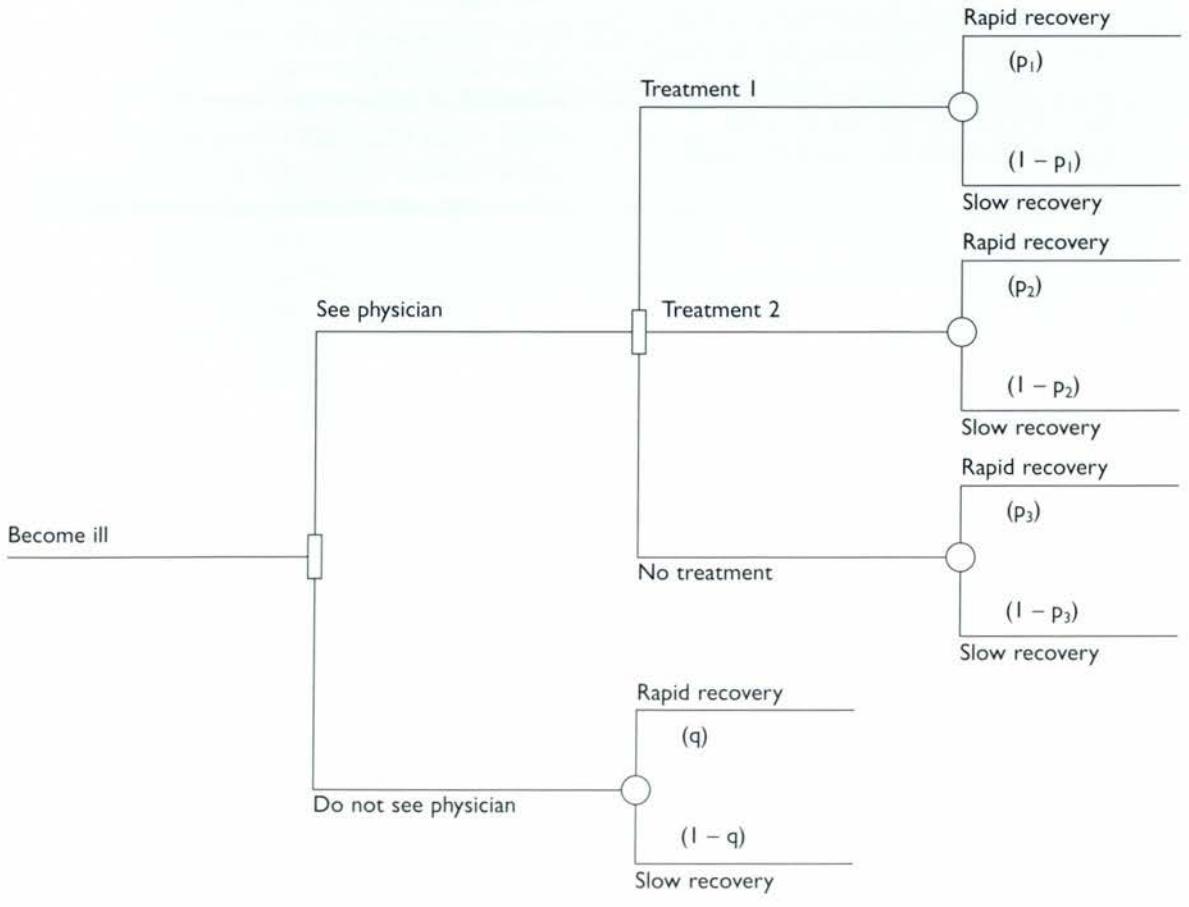
In general we will be interested in the outcomes as well as the costs, and we would calculate expected outcomes in the same manner as we calculated expected costs. There is an additional complication that is often of interest in the calculation of expected outcomes. That is that we are not always simply interested in the expected outcome, but in the expected valuation of that outcome. These valuations frequently rely on the assessment of 'utilities' of outcomes. Loosely speaking, these 'utilities' are the level of relative satisfaction obtained from various actions. We will address this issue further later.

drugs conditional on seeing the doctor. This choice of drugs may be either a descriptive or a prescriptive model depending on the purpose it is to serve. If it is to help physicians decide what treatment is appropriate then the model should be prescriptive. In Figure 4, Figure 2 has been modified to include alternative treatments with potentially different probabilities of achieving outcomes. If a decision is to be made as to which treatment is to be used, then there should be (as indicated) a square decision node at the branch where there is a choice of treatments.

If the model is designed to help someone to decide whether to see a doctor or not, and the drug treatment decisions were determined solely by the doctor, then the drug choice is 'exogenous' (determined outside of the model) to the decision and may be represented by a chance node with probabilities of each type of drug treatment entered into that model if they are known. At a societal level, if one knows that some given percentage of physicians use a particular treatment then these percentages could be converted into probabilities and used here. In this case the square node for treatment choice would be replaced by a circle chance node. Then the model could be used to determine whether someone should see a physician based on this set of information. Thus the model may be descriptive or prescriptive once one has seen the physician (or has not) within an overall prescriptive model concerning the decision whether to see the physician.

The discussion of the opportunity to choose within an overall prescriptive model whether to imbed another prescriptive model or descriptive model brings up another option – adding a deterministic element to the model at this stage. We could just assume that 'good clinical practice' is the mode of care provided. That is, optimal treatment (somehow defined) is provided. In some cases this could imply that all patients reaching a particular point in the model (having a particular outcome) will be treated in one way and only in one way. This would imply a deterministic aspect at that stage, reflecting the good clinical practice assumption.

Figure 4 Additional elements to stochastic prescriptive model in Figure 2



In other cases even adhering to good clinical practice may imply a variety of possibilities, the frequencies of which may be known (this either reflects physician responses to patient attributes not apparent in the model or simply other legitimate physician variation in treatment patterns). In this case we would have a stochastic model where good clinical practice of one sort or another is delivered. Any individual physician may know how a particular patient in his or her practice may be treated, but the analyst would not necessarily have this information. The outcomes of this level of care are then modelled as stochastic or deterministic events depending on what is called for by the situation.

Another possibility can be envisaged. In the last example, one supports implicitly the delivery of good clinical practice in the model; however, such practice may not always be delivered. This is particularly the case in general medical practice as opposed to the situation in RCTs. We will have more to say on this issue below. Recognizing that good clinical practice may not in fact be the norm, we can choose to represent reality in the model rather than some wished-for state of the world. Thus we will have a more predictive model of what the implications of any decisions are in our world rather than in some ideal one, and we can thus better evaluate the prescriptive model as a result – based on the knowledge of what will be the results of various actions contemplated.

As an example in a model of a pharmaceutical treatment, we may use a medicine in place of surgery, but if the therapy fails for some reason, surgery may be the back up option. Suppose that there are guidelines for what type of surgery is performed depending for example or patient type. Let us also assume that 30 per cent of the surgeries performed ignore these guidelines. We now have a choice to make in the model. We can assume that good clinical practice is followed (ignoring reality) or we can adopt more realistic assumptions (ignoring good clinical practice). Since the choice of surgery is exogeneous to the model (one may imagine that the evaluation is being done to decide the cost-effectiveness of a pharmaceutical), one may easily justify the use of descriptive elements regarding surgery procedures – after all we want to know the predictions of using the drug to decide whether we should use it. We need a descriptive model within our prescriptive model.

As another example, if one must take a drug within 48 hours of symptom onset in order for it to be effective, and most patients will not, then we have a similar choice to make. We might use a descriptive element (actual practice) or we might use good clinical practice. It may not be good clinical practice to administer the drug more than 48 hours after symptoms appear. Costs would be incurred (in terms of both drug cost and possible adverse reactions) with no benefit for patients. Cost-effectiveness in reality might be significantly reduced. If one knows that this is in fact the actual practice pattern, only by incorporating that descriptive element into the model will one end up with a predictive model of the implications of using that product, and only then can one prescribe whether the product should in fact be used (the goal of the prescriptive model). We use a descriptive model inside the prescriptive model.

In some ways the choice of modelling reality or good clinical practice can be defended either way, and it might be argued that it would be helpful to do the analysis both ways (ideal and realistic) to indicate the changes in recommendations depending on the level or standard of care actually delivered. This can help direct attention to aspects of care delivery that are crucial in making particular decisions optimal or not. There often will be a role for a model using some type of additional evidence. Figure 5 shows a very simple model where RCT data have been supplemented to increase the relevance of the model to final outcomes and economics. The success and failure data for the treatments in the RCT are used as further inputs to a model that incorporates extra-trial information on the percentage of those who fail pharmaceutical therapy who go on to surgery and who subsequently recover, die in surgery or continue with their chronic condition because no treatment was effective. Each of these outcomes can have associated costs and valuations added, combining economic data with a modified outcomes model. We note that the probabilities associated with recovery, surgical death, and continued chronic condition have been differentiated based on the preceding drug treatment. Thus here it is assumed that these probabilities will vary by initial drug treatment. This need not be the case, in fact it may be quite reasonable (and a useful simplifying assumption) to assume that these probabilities are independent of initial treatment. Thus the number of patients reaching the surgery nodes under each of the two treatments would presumably differ under the two

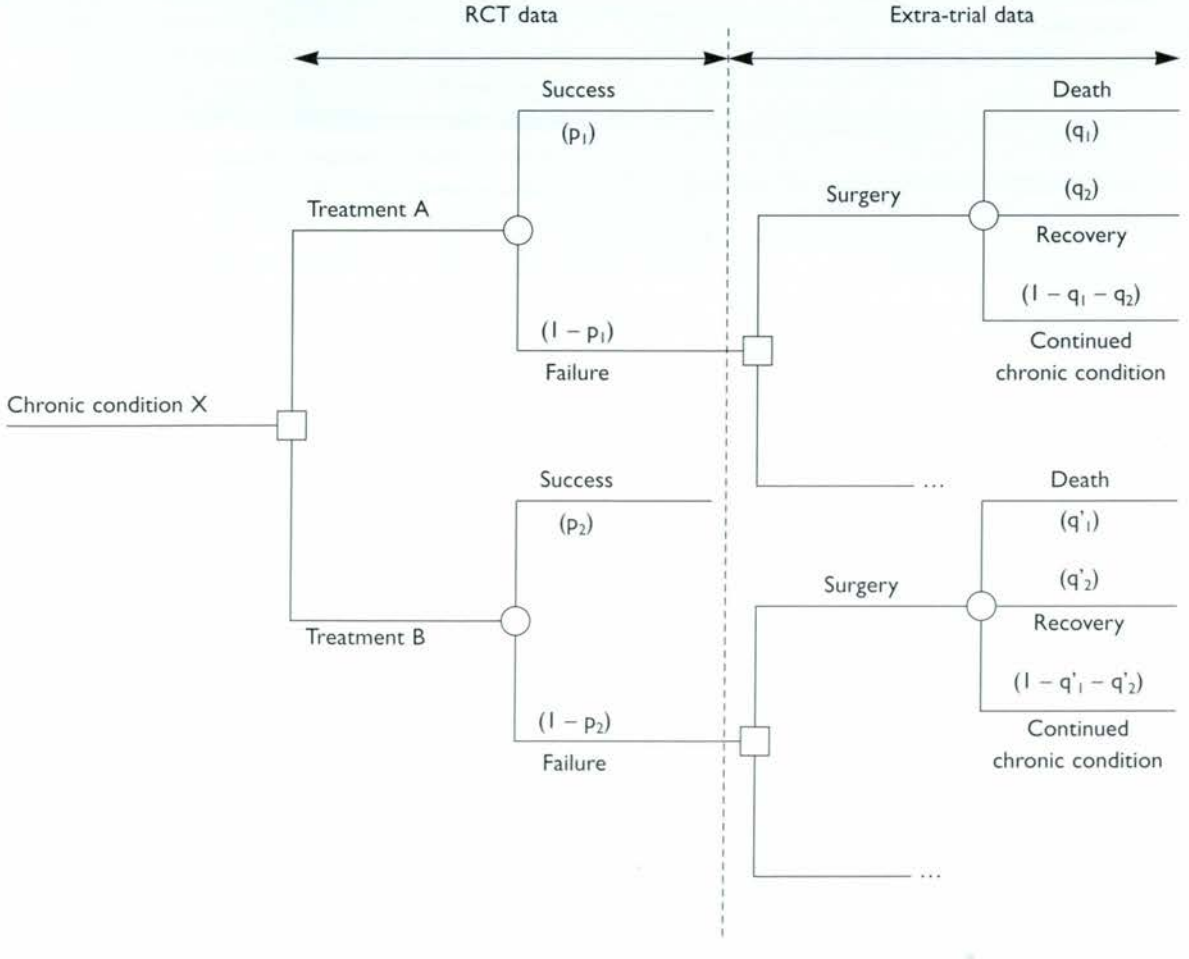
treatment regimes, but the success rates of subsequent activity may not (it would probably be the case that in a model incorporating some inappropriate surgery, probabilities would need to be shown separately according to whether surgical guidelines had been followed). Box 2, on page 15, and Box 7 in Appendix 2 examine some of these probability issues in more detail.

3.3 Modelling data

Here we introduce the other meaning of model. In all cases we have a model of the type described above. If we measure resource use, treatment failure rate, and what happens subsequently, we will have a model with observed evidence on all factors for the same group of patients (as discussed, this may or may not be the evidence one actually wants – good clinical practice or not for example). We might however still lack some of the evidence. In such a case we either cannot perform an economic evaluation or must 'model' the missing data. Modelling in this sense means that we conjecture what would be the case, so that we fill in the missing data that we did not measure. This conjecture can be taken from some measurement source outside the context of the other observations (such as databases for similar patients) or be based on more casual 'guesswork' as to what may be expected to happen to patients.

Indirectly we have come to the issue of what type of evidence we will use in the model. Here we expand

Figure 5



upon the second use of the term model with the introduction of elements of information from diverse sources into a model of the meaning already discussed. This second type of modelling effort corresponds to what might be thought of as linking of diverse sources of information into a (hopefully) coherent whole rather than making a model solely from a set of data based on a particular experiment for example a model based solely on trial information where one measures both outcomes and resource utilization.

Table 2 Evolution of economic evaluations of medicines

Sole concern for acquisition costs
↓
Consideration of some downstream implications of drug use (measured however possible, or just wishful thinking)
↓
Concern for careful measurement (adoption of RCT as a gold standard for economics)
↓
Realization of biases in RCT assessment
↓
Improving RCT design for economic evaluations
↓
Consideration of observational study designs
↓
Recognition of benefits of combining evidence from multiple study designs to reduce bias

Table 2 indicates some of the evolutionary steps in the development of economic evaluations in medicines. Early development of economic evaluation was based on models in the loosest sense of the term. Many of the early efforts (largely informal) at indicating the economic implications of various treatments tended to be rather casual 'marketing stories' with little basis in real data other than a modicum of attention paid to possible beneficial effects that could have economic implications (generally not analyzed formally and with a minimal attention to probabilistic events and full consideration of all possible associated events). These efforts were largely speculative and based more on wishful thinking and casual conjecture. While often not explicitly using a model, these efforts implicitly based their hopes on a simple model – usually extremely simple. Those types of efforts serve as a useful lesson as to the benefits of using formal modelling techniques.

A second level in the evolution of economic evaluation techniques is found in the retrospective examination of data generally limited to measurements taken as part of earlier clinical trials or as part of a culling through databases. This data was also generally limited in scope, typically excluding much in the way of resource utilization or even clinical events not of direct interest to the trial – ignoring failure rates or at least downstream events that follow treatment failure. Some of this data is collectable in databases, but other methodological/bias problems with databases have sometimes limited their value in these ventures.

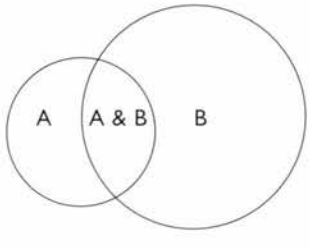
Many economic evaluations have been done based on data collected as part of an earlier effort with some other set of intentions. Frequently no resource use data have been collected in such efforts. To perform economic evaluations in such cases, we must use our second concept of a model – that corresponding to conjecture rather than actual measurement. In using the term conjecture, it is not meant that any casual approach is implied. Conjecture may be quite sophisticated and precise, but it is conjecture in the sense that the patients supplying the basis for the health outcomes data do not necessarily experience all of the events or use the resources that the model attributes to them. No direct measurements were taken. Rather, we conjecture that the patients having particular complications received particular services with particular frequencies and costs (and perhaps particular long term outcomes). Eddy (1992) has described this second use of model as a way of incorporating 'indirect evidence' with 'direct evidence'. Thus the trial would be direct evidence that would be supplemented by the indirect evidence of events not measured directly in the trial. We have 'modelled' the resource use, and perhaps certain of the health outcomes. With retrospective analysis of data, particularly of that collected for noneconomic purposes, this type of modelling is unavoidable.

More recent economic analyses have been done as part of data collection efforts that have explicitly incorporated economic variables into the study and data collection design in prospective trials. Some of these efforts are insufficiently informed or unavoidably constrained in the extent of the economic variables on which data are collected. In some cases it is simply practically impossible to collect all that is needed for economic evaluation in

Box 2 Conditional probabilities

It needs to be emphasized that the probabilities in decision trees, as the branches within them proliferate, are *conditional* probabilities. That is, if event A occurs in a decision tree, then the probability of event B occurring (or not) in the next branch of the decision tree is the probability of B *occurring given that A has occurred*. This is typically quite different from the unconditional probability of B occurring. This unconditional probability, the marginal probability of B, is denoted P(B). The conditional probability of B occurring given that A has occurred is denoted P(B | A). Figure 6 below will assist in this discussion. We may think of the area described by the circle marked A in the figure as representing the probability of A occurring – its marginal or unconditional probability. The probability of A not occurring may be thought of as the area within the rectangle, but outside the circle A. Thus the rectangle area will be assigned the value one as an area. The area of circle A will then correspond to the probability of A which will be a fraction. The probability of A plus the probability of A not occurring is thus equal to one, or P(A) + P(not A) = 1.0. The same holds for event B and the circle describing it.

Figure 6 Conditional probabilities



Conditional probabilities are the probability of one event occurring given that another event has occurred. Just as P(A) is the fraction of the entire probability of 'something' occurring (A or not A) which equals 1.0, so too the conditional probability is the fraction of a different probability – that of the event that one is 'conditioning' on. Thus the probability of B conditional on A having occurred is the fraction of A for which B also occurs. In conditional probabilities, the circle A is analogous to the entire square in marginal probabilities. A fraction of the entire square represents the probability of event A; a fraction of circle A represents the probability of B happening *conditional* on A. This fraction is the area represented by the area A & B as a proportion of the conditioning event, area A. In mathematical notation this is represented as the area A & B divided by the area B: P(A & B) / P(A).

The probability of A & B is the joint probability of these two events and is generally denoted P(A, B). The order in joint probability notation does not matter, so this is equivalent to P(B, A). Putting all these probabilities together gives us P(B | A) = P(A, B) / P(A). We could perfectly analogously write a formula for a different conditional probability, P(A | B) = P(A, B) / P(B). The numerators of these expressions are equal, but the denominators are not. We note that it should be clear that the two conditional probabilities will not generally be equal to each other unless in some fortuitous coincidence P(A) = P(B). Examining the figure indicates that in this case these two marginal probabilities and therefore the two conditional probabilities are not equal (see Box 4, on page 23, for an important example of confusion caused by such misconceptions).

It is also that case that (at least in many examples we will examine in the medical field) there is another set of conditional probabilities that will also not generally be equal to each other. The probability of A *conditional on the occurrence of B* will not generally be equal to the probability of A *conditional on B not occurring*.

(1) P(A | B) = P(A, B) / P(B)
(2) P(A | not B) = P(A, not B) / P(not B)

The only way for these probabilities to be equal is for the events to be independent of each other. In that case P(A, B) = P(A) P(B) and P(A, not B) = P(A) P(not B) and both conditional probabilities (1) and (2) are equal to P(A). As a practical example, the probability of death conditional on having cancer would not generally be equal to the probability of death conditional on no cancer. This is because the probabilities of cancer and death are not independent.

What this may mean as a practical matter in decision analysis models is that we will need more data than we might originally think unless we are willing to make an independence assumption as one of the simplifications in the model building (see Tugwell et al 1985). For example, one may consider compliance rate as either a fixed rate or one that is conditional on the drug one is considering (or the type of patient, etc.). If compliance rate is considered to be a constant, one can use one estimate for all drugs in the model. If it is variable conditional on the drug or patient type then this information needs to be in the model and more data will be required to build the model – conditional compliance probabilities will be needed.

This is an important point, and it is important to recognize that either assumption can be correct

Box 2 Conditional probabilities (continued)

depending on the case at issue. In a model predicting the full costs of using birth control devices, one must consider what happens when the device fails to prevent pregnancy. When a pregnancy occurs, any of several events may happen. These may not depend at all on the device used and could then be considered independent events from the device (e.g. whether one used condoms or occasional abstinence as methods, the frequency of spontaneous abortion *conditional* on pregnancy might be expected to be the same (though perhaps not the frequency of pregnancy itself! – a different issue).

On the other hand, when representing the frequency of expected *elective* abortions in a model, it might be quite important to know other details about the women being studied (religion, for example). Presumably the expected frequency of elective abortion will not be independent of religious background. Making the assumption that it is independent by virtue of using the same probability for all women may lead to significant errors in prediction. It is doubtful that anyone would explicitly make such an error in a model where religion was explicitly incorporated, but one could *implicitly* make the error by transporting a model across cultures with insufficient care (for example, a model with probability of elective abortion based on the religious attitudes of a representative sample of American women will probably not transfer without modification to Ireland).

Decision trees are valuable tools for analysis, but are only as good as the assumptions going into them. The trees are constructed as conditional events – each event to the right of another is conditional on the event to its left. Decision trees (like computers) will generally do what analysts ask them to do without knowing whether it is correct or incorrect. The simplicity of the decision tree is beguiling, but it is only as accurate as its designer. Careful attention to detail is essential.

conventional RCTs, most of whose primary purpose is supplying information for registration approvals. Thus here too we will need to supplement the data collection effort with modelling efforts to supply necessary information. We will see below that there are good reasons to employ such conjectural modelling. Even when economic variables are routinely collected as part of RCTs, solving one set of economic evaluation problems, there will remain a need for this type of modelling to accurately portray the real implications of using particular therapies.

Section four discusses some potential problems with data from RCTs as the sole basis for economic evaluations, emphasizing the distinction between the 'efficacy' (as measured by an RCT) and 'effectiveness' – the counterpart when measured under actual practice conditions. Section five discusses modelling as an approach to attempting to solve problems of bias identified in section four. It also outlines the problems of bias with other sources of data. Section six presents ways of combining data from multiple sources of identical or complementary design. Section seven addresses the issue of the acceptability of modelling studies.

Box 3 The virtues of systematic thinking through modelling

One of the important virtues of formal modelling techniques is in keeping track of data and organizing thoughts in complex situations. One of the more striking examples of this comes from Eddy (1982) in which physicians were asked to answer a question about the use of screening in identifying breast cancer. The problem is presented below.

The prevalence of breast cancer in a particular population of women is one in one hundred (.01). The accuracy of diagnostic mammography is assumed to be represented by sensitivity and specificity values of .792 and .904 respectively. These values are the proportion of similar women who would test positive if they actually had breast cancer and the proportion who would test negative if they truly do not have breast cancer.

Given this information, the physicians were asked what the probability of cancer would be if a particular woman's test was positive. Ninety-five per cent of the physicians in the sample answered that the probability was about 0.75. In fact the correct answer is an order of magnitude lower (0.078). Such errors can imply significant overuse of further diagnostics, leading to higher expenditures (largely useless) and high levels of unwarranted stress on many of the affected individuals. What can explain this error rate? More importantly, perhaps, how can we avoid this type of error?

Eddy suggests that the error may be due to a confusion of basic probability concepts by the physicians. It is clearly not the result of careful systematic thinking about the problem, but of casual approaches to decision making. Kahneman et al (1982) imply that physicians would not be unique in their confusion. However, in this particular case their errors can have significant consequences. Specifically, the physicians may be confusing the probability of obtaining a positive test conditional on having the disease (also equal to the sensitivity of the test) with the probability of having the disease given that one has obtained a positive test. These two conditional probabilities are not the same (as we saw in Box 2), and the implications of assuming that they are may be rather significant if errors of the magnitude mentioned above are made.

Table 3 indicates the data needed to answer this question with asterisks next to the numbers given in the problem. All other numbers are calculated solely from those initial numbers. Figure 7 indicates a model of the calculations showing that they are quite logical (and border on at least relative intuition), something that the basic formula for obtaining the correct answer presented below does not share.

$$P(C \mid +) = \frac{P(+ \mid C)\, P(C)}{P(+ \mid C)\, P(C) + P(+ \mid NC)\, P(NC)}$$

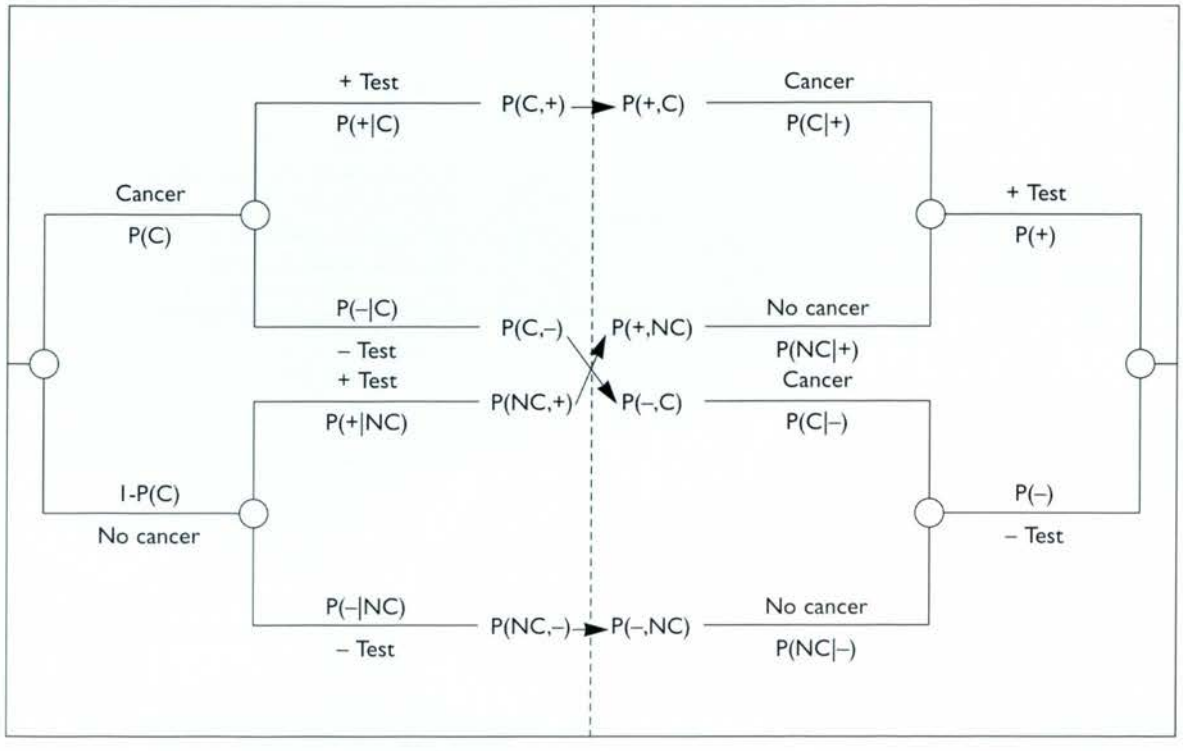
Table 3 Probability notation and numerical values

Notation	*Meaning*	*Numerical value*
P (C)	Probability of cancer (prevalence; prior)	(0.01)*
P (NC)	Probability of no cancer [=1–P(C)]	(0.99)
P (+|C)	Sensitivity (=probability of a positive test given cancer is present	(0.792)*
P (–|C)	(1 – sensitivity) =probability of negative test given that cancer is present = false negative	(0.208)
P (–|NC)	Specificity (=probability of negative test given no cancer is present)	(0.904)*
P (+|NC)	(1 – specificity) =probability of positive test given that cancer is not present = false positive	(0.096)
P (+,NC)	Joint probability of having no cancer and a positive test [=P(NC,+)]	(0.095)
P (+,C)	Joint probability of having both cancer and a positive test [=P(C,+)]	(0.008)
P (+)	Probability of a positive test result =P(C,+) + P(NC,+)	(0.103)
P (C|+)	= Posterior probability (probability of cancer given a positive test)	(0.078)

Source: Rittenhouse 1994

This formula will yield the correct answer, but lacks intuition. The model provides at least some intuition once it is explained as well as a systematic way of reconstructing and communicating the process and results (for more detail see Rittenhouse 1994). From the left of the figure we see one decision tree, beginning with prevalence of disease P(C), one of the initially provided pieces of information. From this prevalence we can also calculate the prevalence of no disease P(NC) since all probabilities at chance nodes must sum to equal one. Figure 7 provides the general notation while the table provides the specific numbers. Given the presence or absence of cancer we have data on the results of tests (sensitivity, P(+ | C) and specificity, P(– | NC). These are conditional probabilities, and since chance node probabilities must sum to equal one, we can calculate the complementary probabilities at each of the chance nodes [P(– | C) and P(+ | NC)]. All the probabilities in the left side decision tree are thus known and one can calculate (as shown in Box 2) the joint probabilities of each of the branches [eg P(C,+)] as the multiplication of the constituent path probabilities (the appropriate marginal and conditional probabilities) leading to each of the four terminal nodes (outcomes) [eg P(C,+)=P(C) P(+ | C)].

Figure 7 Calculating updated probabilities of cancer after a disgnostic text



The right side of Figure 7 shows a 'flipped' decision tree based on the same data. We will proceed to calculate and fill in the probabilities. First we can start with the aforementioned joint probabilities and move them from the left side tree to the right side 'flipped' tree (the leftmost part), switching the positions of the joint probabilities as indicated by the arrows to account for the alteration in their placement given the different structure of the trees. We can now fill in the 'marginal' probabilities at the rightmost side of the right side tree.

There are two ways and only two ways of having a positive test – with and without cancer accompanying it. These two joint probabilities thus summed will give the probability of a positive test [P(+)]. One may perform a similar calculation for the probability of a negative test or simply use the knowledge of P(+) and the fact that the chance node probabilities must sum to equal one, implying that P(–) = [1 – P(+)]. Now all probabilities in the right tree are known except for the conditional ones, one of which (P(C | +)) is of interest in that it is the answer to the question posed, the probability of disease given a positive test was obtained. We can use the knowledge from Box 2 again to calculate these probabilities from the joint and marginal probabilities [P(C | +) = P(+,C)/P(+)]. The answer is the same as that which would be given by the equation above, but (at least in retrospect) the process is simpler (and easy to reconstruct upon reflection). Such a model can assist in keeping the concepts clear and avoiding the mistakes so common in casual thinking about such problems.

4. The efficacy/effectiveness distinction

'Very true' said the Duchess: 'flamingoes and mustard both bite. And the moral of that is – 'Birds of a feather flock together."
'Only mustard isn't a bird,' Alice remarked.
'Right, as usual,' said the Duchess: 'what a clear way you have of putting things!'
– ALICE IN WONDERLAND (Freedman et al, 1991, p.133)

Various types of evidence can be brought to bear on an issue by incorporating them into a model. We have seen that many early models of the economics of pharmaceuticals had to rely on indirect evidence for particular data since none had been collected as part of the trial upon which most of the economic analysis is based. In later efforts where one is intending to collect economically relevant data as part of the trial, one remains limited sometimes by other development concerns e.g. not overly burdening trial investigators. The primary purpose of the trial is registration; this goal will not be sacrificed to make a trial more relevant to economics. When data requirements for economics appear to be overly burdensome for trial investigators, they are left out. In part one might argue that these problems are in the past. As it is acknowledged that economics is important both to the sponsoring company and to others, these problems may solve themselves. It may be argued that we no longer have to use indirect evidence in models – we can measure directly and not "model" reality by using indirect evidence. It is interesting to ask the question, if trial developers handed over the reins tomorrow to economists, would the need for this incorporation of indirect evidence into a model disappear? To a degree the answer must be 'it depends', specifically on how much true latitude the economist would have. But let us say that the goals of registration must still be met in these same trials thus limiting to a degree the latitude to radically alter conventional trials. Are the problems now absent? Here I will present the case that we would still need models, at least if we want to have reliable predictions of the economics of products once they are on the market. Much of this discussion follows that of Rittenhouse (1995, 1996) and Rittenhouse and O'Brien (1996) to which interested readers are referred for additional details.

The randomized, blinded controlled clinical trial design (RCT) is the workhorse of the evaluation process in pharmaceuticals. This design answers the question of *efficacy:* 'can this product be superior to (usually) placebo?' It answers this question very well. Perhaps its chief strength is that it eliminates (or at least does better than any other possible design) potentially confounding influences on the observations of treatment effect. It achieves a high level of what is called 'internal validity,' or what Senn (1990) has called 'proof within the trial.' Dispensing with randomized allocation of patients to treatment arms in a trial or reducing the control will weaken the inference regarding treatment effect observed in the study. In a well-designed RCT, the inferential basis from cause to effect is a strong one – *for the populations studied under the conditions under which they were studied.* When one infers from such a trial that one product is superior, one can have a high degree of confidence that the experiment in fact truly indicates this and the result is not due to some other influences on outcome. That is the chief goal for registration purposes.

What is often ignored in discussions of whether RCTs are the appropriate vehicle for answering other questions is that there are two types of validity. The other type is 'external validity' (Cook and Campbell, 1979) and is more or less synonymous with 'generalizability' – does the result translate to other contexts? It is with other contexts that economists are concerned – how will the product work in practice, not can it work. This is the question of *effectiveness:* 'how does the product work in the practical environment in which it will actually be used?' Critics of nonexperimental methods (non RCT methods) often seem to de-emphasize the importance of this type of validity, sometimes implying that there is only a general type of validity to research (see for example Sheldon, 1994). Medicines are not administered under randomized blinded or controlled conditions in the real world of clinical practice. They are not administered by reference to strict protocol. The individual patients who receive them are not selected according to particular traits (e.g. the lack of comorbidities and the willingness to be randomized in an experimental trial). Numerous issues associated with the RCT imply that it may be quite poor in terms of its external validity. Table 4 lists some of the potential problems with the RCT when it is employed in the service of economic evaluations. Each of these topics will be discussed briefly below.

Table 4 Potential problems with RCTs used as the basis for economic evaluations

Choice of Comparison Therapy
Protocol-driven Costs and Outcomes
Artificial Environment
Intermediate v Final Outcomes
Inadequate Patient Follow-up
Selected Patient and Provider Populations

Choice of comparison therapy

One of the more fundamental problems associated with the RCT-based pharmacoeconomic study is that the chosen comparison therapy is often not that which is relevant or most relevant for policy questions of interest. The most obvious case of this problem is the placebo comparison which has no relevance at all to practical policy as a placebo will never be a practical alternative. Showing relative superiority (and appropriate pharmacoeconomic studies must always be comparative) to placebo is simply not relevant. When the policy-relevant alternative is an active treatment, the placebo is clearly not going to represent that treatment. Even when the relevant alternative is no treatment, placebo will not represent that either unless placebo effects, the very reason for including placebos, are wholly absent.

Less clear is the frequent irrelevance of chosen controls even when the trial design incorporates active controls. The most relevant alternative may be nondrug therapy. The early economic evaluation of cimetidine for use in duodenal ulcer used a comparison of surgery although the trials for product approval were done against placebo (Culyer and Maynard, 1981). Most studies choose comparison therapy without regard to practical policy questions concerning the relevant comparator for economics. It is not always clear what this alternative would be even were development departments specifically sensitive to the needs of economic evaluation. What is relevant to one policy context is not necessarily so in another. This is evidenced by recent publication of guidelines for economic evaluations by both Australian and Canadian authorities. In the Canadian guidelines the relevant comparator is stipulated to be both 'existing practice and minimum practice'. Existing practice (which explicitly acknowledges nondrug possibilities) would either be the single most prevalent clinical practice (if there is one that is dominant), or it could be current practice weighted by market share. Minimum practice would normally be either the lowest cost comparator that is more effective than placebo, or the do-nothing alternative, as appropriate (CCOHTA, 1994). In the original version of the guidelines for Australia, the relevant alternative was the most widely used alternative (Drummond, 1992). The revised guidelines, while defining the concept more precisely, appear to appeal to the same basic notion (Commonwealth Department of Human Services and Health, 1995).

Drummond et al (1993) make a further point – proving that a therapy is cost-effective incrementally against a trial comparator does not mean that it is cost-effective incrementally against the relevant alternative. If society is using a product, assuming incorrectly that it is cost-effective, and a new product is shown to be incrementally cost-effective against it, we may be missing the relevant comparison. That comparison may be against doing nothing or at least some alternative not considered in a trial.

With the frequent global development plans of many pharmaceutical companies and the multicenter, multicountry trials used to fulfill those plans, it is clear that any particular RCT may have a difficult time in satisfying the sometimes competing needs of varying purchasing authorities. The problem may become more acute in decentralized environments such as the US where any payer may develop its own criteria as to relevant comparator depending on what is currently on its formulary. While the Canadian guidelines may be the most theoretically appropriate in stipulating the relevant comparator (they were developed by consensus building and based on state-of-the-art methods) the relevant comparator in practice is not going to be the same in all environments. Consequently, a trial using the right comparator for one customer of economic evaluation will be using the wrong one for another. The solution will need to be one where modelling is done to predict the performance of the relevant alternative, perhaps incorporating results of RCTs using those alternatives. A multiplicity of clinical trials to answer the question in an RCT framework is simply not practical even if it were desirable.

Protocol-driven costs and outcomes

Particular procedures to ensure the compliance, safety or optimal care of patients participating in medical trials may not be done in general practice, thus the costs observed in a trial will be overestimated as

compared to real clinical practice. It is also possible that costs in a trial may be underestimated. It may be that appropriate levels of care are less than those supplied in practice (or at least in some practices). This may be simply due to lack of information, or outdated information, or to the practice of 'defensive medicine' in countries like the US with particularly active malpractice attorneys. Regardless of the cause, the state-of-the-art medicine practiced in many trials may in fact be quite different than in general practice. This may have implications for both health outcomes and the costs of achieving them.

A particularly good hypothetical example is provided in Eisenberg et al (1989) where the authors speculate on the effects of routine testing within a trial for the presence of subclinical disease when the same level of testing would not be performed in general clinical practice. The example shows the ambiguous effects on number of cases identified and the average cost of a treated case (see Box 4 on page 22). The earlier detection with the protocol-mandated procedures may identify more cases (some of which would never manifest symptoms), but the earlier detection may imply more straightforward and simpler (less costly) treatment of those cases. Untangling these effects of protocol-mandated procedures is not trivial. What is clear is that the trial results may bear little resemblance to those to be expected in general practice – in terms of health outcomes or costs of achieving them.

The protocol that identifies patients for the RCT can inject biases. If a screen is necessary to identify patients eligible for the trial, a screen that is more accurate than that used in practice will identify more true positive cases. As such drugs will have inflated efficacy over that to be expected in the more general population. Russell (1994) has observed that labs involved in clinical trials of cholesterol-lowering agents use a difficult, but accurate method for measuring blood cholesterol – the modified Abel-Kendall method – whereas most general labs use a less precise method and are less diligent about accuracy. A 1985 study by the American College of Pathologists found extremely large variations in lab reports on a sample with a known cholesterol value of 262.6 mg/dl ranging from 101 to 524 (Garber et al, 1989). Similar results were obtained in subsequent surveys. Since LDL levels of cholesterol are the basis for final diagnosis and treatment and these levels are determined from results from total and HDL levels (the latter being even less accurate than the total levels), there is room for significant concern about the level of accuracy in diagnosis and treatment in general practice. There is likely to be significant misclassification in the real world relative to that of trials. This will reduce the possible benefits from cholesterol lowering therapy which are already in doubt for other reasons.

The Canadian guidelines include a suggestion that 'protocol-driven costs should be excluded if they would not occur as part of the intervention on a regular basis' (CCOHTA, p.31). However, there is no mention of the mere exclusion of such costs being an incomplete correction for all protocol-induced effects on a trial. The ambiguity of protocol-driven effects on both costs and outcomes is neatly illustrated in Eisenberg et al (1989) referred to above and discussed in Box 4. The subclinical disease is treated in the trial; however, such treatment would not typically be done in practice. Thus it is clear that protocol-mandated testing can affect not only costs (by including the costs of the tests) but also outcomes and thereby the downstream costs of the trial. Furthermore, the effect on cost is ambiguous. The early identification of subclinical disease, some of which would never manifest itself as clinical disease (false positives of a sort) may imply a greater cost in the trial as more of these cases would be observed in that environment. Countering this effect might be the earlier identification of subclinical disease, perhaps enabling quicker and cheaper treatment or prevention, arguing for lower costs in the trial. The effect on costs is ambiguous; however, it would appear that the effect of such testing would have on outcomes measures is unambiguous. Outcomes will clearly be better than those to be expected in practice. Untangling these various influences on costs and effects is not a simple matter. Often the problem is not even acknowledged.

Artificial environment

In a trial patients are often reminded explicitly or implicitly of the importance of compliance with medication taking directives. Patient diaries of medication taking behaviour are common. As a consequence, even for drugs for which compliance may differ in general practice, there may be little indication in the trial of the difference. To the extent that this compliance level is important for treatment outcome, treatment outcomes may be more similar in the trial than those to be expected in general practice.

Box 4 The ambiguous effect of protocol-induced costs on costs and outcomes

Eisenberg et al (1989) have provided a hypothetical example of the ambiguous effect that subclinical detection in a trial can have on the economic evaluation of treatments. The case is one in which two prophylactic treatments are compared in an RCT where active case-finding (testing for subclinical levels of disease) and passive case-finding (reacting to clinical symptoms) are both pursued but where in routine clinical practice only passive case-finding would be the norm. Figure 8 below represents the case findings for 100 patients under routine care (left) and the RCT (right) assumptions. Table 5 presents the cost assumptions used in this example. In the RCT 12 subclinical cases are detected. In the routine practice, none of these would be detected at that level, though 10 exist (they are assumed to never develop to clinical cases). These figures imply that two of the subclinical cases detected in the trial would have developed clinical symptoms had they been investigated under passive case-finding. In general we can see that the RCT enables earlier detection, with fewer cases being in the more severe ranges.

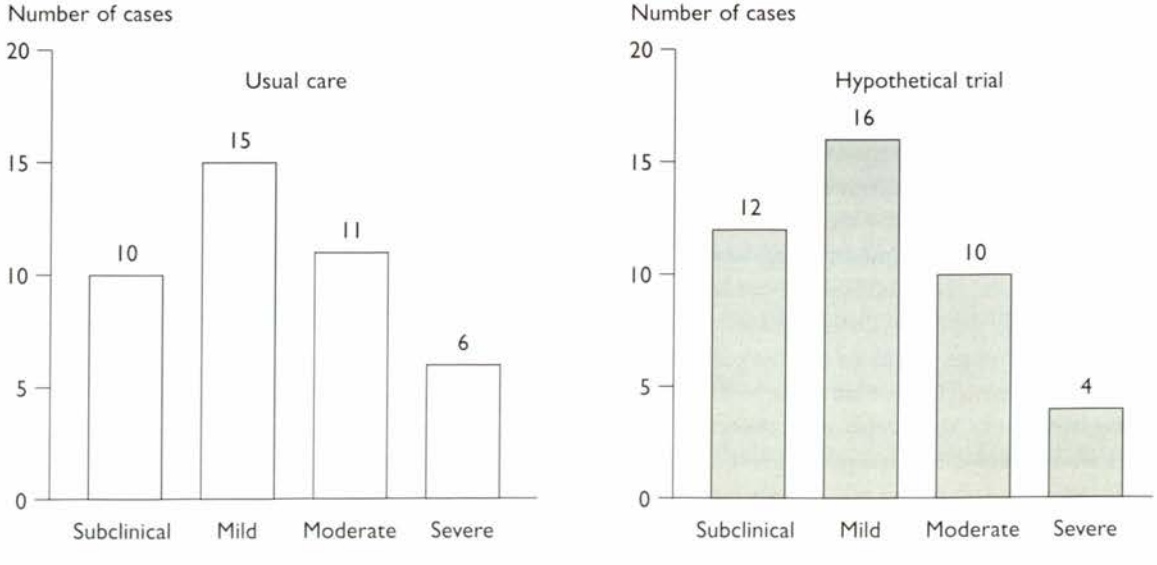
Multiplying the numbers in each of the figures by the respective additional treatment costs at each level of detected disease, as set out in Table 5, yields the total costs of treating these 100 hypothetical patients under the two prophylactic regimes (note that in the routine care case, no expenditure is made on subclinical cases since they remain undetected). For the routine care case there are 32 cases detected and treated at a cost of $5500, or an average of $172 per case. For the RCT, the number of cases is 42 (it includes the 10 subclinical cases that will never develop symptoms) and the total costs are $5880, yielding an average cost of treatment of failed prophylaxis of $140. We can see that the total costs of treating the failed cases is higher in the trial, but lower in an average per case basis. The authors continue their example to show that under further hypothetical assumptions, the RCT results indicate a cost-saving treatment while the routine care results indicate a modest increase in cost for the derived benefit. Thus the economics of the treatments can be affected by the effects of protocol induced procedures. They could either increase of decrease average per case. It is not generally predictable.

Table 5 Cost per case in hypothetical clinical trial

Type of disease	*Average cost*
Undetected subclinical disease	$0
Detected subclinical disease	$90
Mild disease	$100
Moderate disease	$200
Severe disease	$300

(From Eisenberg, et al., 1989)

Figure 8 Number of cases in usual care versus a hypothetical trial



(from Eisenberg, et al., 1989)

Intermediate v final outcomes

It is often necessary in order to obtain results in a timely or less costly manner to use intermediate endpoints in a trial. In some cases final outcomes are not as frequent (at least within the time period of study). Thus a study may be designed and the sample size chosen to assess an intermediate endpoint that will occur more frequently within the study period. These results are of limited value to economic evaluation. The achievement of an intermediate endpoint means little in and of itself for economic evaluation (or for health outcomes in general for that matter). Economic assessment must depend on final outcomes and their valuations and costs of achieving. Intermediate outcomes are useless unless they are linked via a model to these other outcomes that are of ultimate interest. Intermediate outcomes may wholly misrepresent the reality of the situation by emphasizing differences that may not be apparent in final version. Brody (1995) reports that in the TIMI-I trial of tPA versus streptokinase for example, the primary endpoint was 'patency of the infarct-related artery' within 90 minutes of infusion of the study drug. In part this endpoint was chosen because it enabled a considerable reduction in sample size as compared to that necessary had mortality been chosen as a primary endpoint. This trial was in fact terminated early because of the substantial differences between the study drugs in achieving the endpoint (70 per cent v 43 per cent). The implications of that endpoint appeared to be dramatic; however, later results indicated that such an endpoint did not translate into final outcomes of nearly as marked significance. It is clear, therefore, that intermediate endpoints can be highly flawed in indicating health outcomes. They are also unable alone to indicate the resource utilization ultimately traceable to the therapy unless supplementary data are incorporated. Thus from a resource utilization and a health outcomes perspective, intermediate outcomes are insufficient and must be supplemented in some manner to answer the questions of ultimate interest.

Inadequate patient follow-up

Frequently an economic analyst is faced with a data set that abruptly ends just when the data become interesting. In many cases data collection in a trial ends when a patient reaches certain predefined endpoints. These are not always endpoints of most interest to the economist. For example, patient withdrawals from therapy and drop outs from a trial are frequently related to treatment. When this is the case, alternative therapies must be used and, if the causal event for the treatment discontinuation was an adverse reaction, treatment of the reaction may imply resource consumption. In some cases the withdrawal of the treatment could cause a worsening of the underlying disease at least until alternative therapy can be introduced. In cases where therapy is used in prevention of acute events, the withdrawal of treatment could induce such a manifestation (e.g. antiepileptic drugs and seizures). These effects and resource utilization are attributable costs to the failed drug therapy, but often, beyond documenting that a withdrawal occurred and whether it appeared to be treatment related, traditional RCT data collection ceases at this point. Data on what happens next can be extremely important to the economist.

Selected patient and provider populations

In order to reduce the potential for confounding influences on treatment effect, RCTs frequently restrict the population in studies to a relatively homogeneous set of individuals by, for example, excluding patients with comorbidities or certain age groups. Such exclusions may imply a biased response compared to what is to be expected when a broader range of patients are treated. Other patients may simply react differently to the treatments (efficacy or adverse reactions). For example the efficacy of blood pressure reduction on severe hypertensives may differ from that on mild hypertensives (rarely is a drug never cost-effective, or, indeed, always cost-effective, but rather it is or is not only in particular populations). Unless the drug is to be used in exactly the same types of persons as in the trial this is always a hazard – it is not necessarily the case that the drug would not work, but that it would work less well, and the economic implications of reduced benefit may be that the product would not be cost-effective in the expanded population (Russell, 1994). Also there is a more subtle difference. Patients agreeing to be randomized into a trial to test medications tend to be healthier than the average patient. They also tend to be more compliant. These effects can interact with drug attributes to produce absolute and relative differences in outcomes. Senn (1990) has observed that 'such patients are not a random sample from any

useful population to which we might wish to generalize results, nor could they ever be, since by definition of being involved in a trial they have all given consent.'

The differential abilities of providers are also potentially leading to bias in results. The provider population in most trials is far from representative of the general provider population. Rather one often sees some of the best investigators at the best medical centers involved in trials. Outcomes are likely not representative of general practice. Rittenhouse (1995) has shown the potential for bias traceable to the differential diagnostic abilities of physicians in a pivotal trial versus those in a sample of general practice. When the diagnostic abilities are less accurate and many of the patients receiving the drug of interest are unable to benefit from it, the economic picture can become rapidly unappealing.

Some of the problems discussed above can be limited by more careful design of RCTs (if those in control of the design are receptive to the needs of economists). Thus, the problem of inadequate follow-up can be at least partially addressed by extending follow-up to all patients in the trial, including those who discontinue treatment prematurely. Other problems will remain however. Certain problems are inherent in the designs of RCTs – randomization for example. It is clear also that trials will of necessity continue to use exclusion/inclusion criteria to homogenize the trial population. The use of randomization and specified entry criteria for RCTs are strengths for showing efficacy, but can be weaknesses for economic evaluation. All of this implies that the RCT will continue to be a flawed device for measuring what economists want to measure. Two alternatives exist. One is to wait and do economic analysis after marketing approval is obtained when less controlled trials can be done or when other study designs are useable. While it may be interesting to do these studies, it is clear that economic results are quite useful earlier in the drug diffusion process for both manufacturers and purchasers. Even were this option a realistic one, the inherent biases of other (e.g. observational) study designs may make the reliance on them equally problematic. The other solution is to incorporate modelling to a greater extent than is currently practiced, admitting the biases of all study designs, measuring them and attempting to correct for them.

Of course, these are not mutually exclusive alternatives. Modelling (drawing on evidence from pre launch RCTs) can be used to hypothesize likely real world cost-effectiveness, with subsequent "naturalistic" trial data or other study results being generated and used to validate or revise the model.

5. Non-RCT sources of evidence, the need to address bias, and the use of sensitivity analysis

'The fact that such studies seem difficult to justify when judged by the standards of clinical trials is not sufficient to condemn them. We do not call geology malformed because it is not physics.'

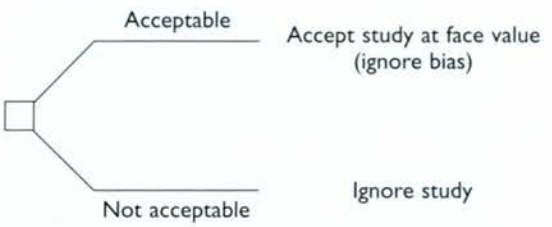
– S. SENN (1990)

While Senn's comment quoted above was directed at the defense of case-control studies specifically (a particular type of epidemiologic study), it is applicable to the general area of observational or nonexperimental studies and modelling efforts. Nonexperimental designs often offer advantages in terms of cost, time, appropriate research design for the problem at hand and ethics (Eddy, 1990). While I will focus here on the issue of appropriate research design, the cost and time involved in RCTs is a major practical disadvantage. If they were the only legitimate method, these criticisms might carry less weight; however, they are not the only method available. Moreover, they may not always be the best method for the task at hand in economic evaluation.

The Canadian guidelines (CCOHTA, 1994) for the economic evaluation of pharmaceuticals emphasized that 'pharmacoeconomic studies should use effectiveness data,' but offer little guidance as to where such effectiveness data are to come from at a time when products are not yet marketed. Prelaunch studies will only have access to efficacy data from trials. 'Thus, prelaunch... studies must extrapolate as best the can from trial efficacy to utilization effectiveness. The assumptions used in this extrapolation (for example, patient compliance rates) must be explicit and must be tested thoroughly with sensitivity analysis.' (CCOHTA, pp. 24-5). It is primarily through modelling efforts that such 'effectiveness' data can be extrapolated from estimates based on efficacy.

Eddy (1990) has suggested that the typical approach to evaluation is to label a particular method or study as "biased" and thereafter ignore it or to accept it as containing acceptable levels of bias and thereafter accept it at face value, ignoring the bias altogether (Figure 9). Typically RCTs are accepted as evidence and the biases discussed above are largely ignored – their results are usually accepted at face value. Other standard epidemiologic study designs that may be much less expensive to conduct are dismissed as 'biased' as compared to the presumed gold standard

Figure 9 Typical approach to evaluation of bias in data *(Eddy, 1990)*



of the RCT. In practice, those with a strong preference for RCT based evidence do often recognise the relevance of other sources of evidence, providing the potential for bias is explored and taken into account. The NHS Centre for Reviews & Dissemination (CRD) (1996) gives an example of a hierarchy of evidence, reproduced as Table 6 below.

Table 6 An example of a hierarchy of evidence

Experimental	
I	Well-designed randomised controlled trials
	Other types of trial:
II-1a	Well-designed controlled trial with pseudo-randomisation
II-1b	Well-designed controlled trials with no randomisation
Observational studies	
II-2a	Well-designed cohort (prospective study) with concurrent controls
II-2b	Well-designed cohort (Prospective study) with historical controls
II-2c	Well-designed cohort (retrospective study) with concurrent controls
II-3	Well-designed epidemiological case-control (retrospective) study
III	Large differences from comparisons between times and/or places with and without intervention (In some circumstances these may be equivalent to level II or I)
Expert opinion	
IV	Opinions of respected authorities based on clinical experience; descriptive studies and reports of expert committees

Source: NHS Centre for Reviews & Dissemination (1996)

The CRD paper includes sources of checklists to help assess the quality and likely validity of data from both experimental and observational studies. It notes that 'validity depends not only on the type of study but how well it was designed, carried out and analysed' (NHS Centre for Reviews & Dissemination 1996 p32).

In practice an element of pragmatism is therefore required. Some forms of bias are more problematic than others, and the optimal source of information or mode of study may depend on the specific context. The best method of attacking the analytic problem may be through combining information from a mixture of study designs, adjusting results for biases, in a modelling exercise. This is, in effect, the strategy that Eddy proposes in his 'flexible, but firm' approach to technology assessment (see Table 7). The flexibility is implied by the willingness to incorporate data that

Table 7 'The flexible, but firm' approach to evaluation *(Eddy, 1990)*

1. Drop preconceived notions of acceptability of various research designs
2. Gather whatever evidence exists
3. Identify biases
4. Estimate the magnitudes of bias
5. Adjust results for bias
6. Use the adjusted results as the best estimate available for decision making

does not come from RCTs. The firmness is based on the utilization of formal statistical models of biases 'to incorporate focused subjective judgements (not global clinical impressions)' and to adjust results for bias.

A suggestion often proposed to 'solve' the problems associated with measurement under conditions of potential bias (or uncertainty in general) is the use of sensitivity analysis. Sensitivity analysis refers to the variation of a particular parameter value in a model, for example the rate of a particular side effect's occurrence, in order to examine the result of such variation on the model conclusions. The idea is typically invoked because of a lack of certainty about the value in question. If the results are relatively insensitive to variation in the value, the results are considered to be more robust than if this is not the case. Sensitivity analyses are no panacea however (O'Brien et al, 1994a). Typically, a given value is varied from its baseline value in the model by both increasing and decreasing the value. There is no rule as to how much to vary the parameter other than what appears to be reasonable. The choices are entirely subjective. When 19 out of 20 sensitivity analyses indicate no change in conclusions, can we claim that we have produced something like a p-value of .05? The answer is clearly no, because there is no theory of random (or any other kind of) sampling underlying sensitivity analyses. The fact that variation did or did not produce some particular level of change in the model's conclusions has no such interpretation.

Many analyses use univariate sensitivity analysis – varying only one parameter at a time – which only captures the effect of minor change in the overall model, particularly if there are many biases infecting numerous parameter values. Some have attempted to solve this difficulty by performing multidimensional sensitivity analyses – varying several parameters at once. Sometimes this is quite beneficial; however, it rapidly can turn a report into pages of numerical results from the sensitivity analyses. Again, with no theory behind these sensitivity analyses, there is little to guide one in their interpretation unless the results are completely insensitive to any changes, an unlikely situation.

There have been several attempts to improve upon the scientific and theoretical basis of sensitivity analysis. Doubilet et al (1985) explores 'probabilistic sensitivity analysis' which bases the variation in values of any particular variable on a distribution of those values (hypothetical or estimated). O'Brien et al (1994a) have developed an approach to developing a confidence interval for cost-effectiveness analyses. A combination of these approaches may offer hope for the future, but presently, these methods are still being developed. Briggs (1995) provides a helpful summary of these concepts.

Sensitivity analyses do have a role to play, despite their subjectivity. They are themselves a modelling technique. However, they are best put to use after attempts have been made to reduce the potential biases in the underlying model (Rittenhouse, 1995). Thus the bias in study results is measured and applied as a 'corrective' to develop a baseline set of parameters for the model. These parameters are then varied in a sensitivity analysis. We start from a more realistic set of baseline values, and can have greater confidence that our sensitivity analyses in total have a reasonable foundation.

6. Methods of combining evidence for models

'Data! data! data! he cried impatiently. 'I can't make bricks without clay.'

– SHERLOCK HOLMES (Freedman, 1991, p.305)

In recognition of the frequent inability of any one study to shed sufficient light on many of these complicated medical issues, methods have been developed to combine evidence from multiple sources to more definitively describe treatment influences. Here we will very briefly describe two of them: meta-analysis and cross-design synthesis. These methods are themselves models and their results can be useful inputs into other economic or health outcomes models.

Meta-analysis

Meta-analysis is a term describing techniques of combining evidence from different studies (usually of similar design). While the term itself is relatively new (Glass, 1976), the methods have been used since the 1930s (Petitti, 1994). Meta-analysis is a systematic and formal mathematical alternative to a literature review. It attempts to quantify and combine evidence from multiple sources in a more objective manner than the less formal literature review. Frequently this effort is motivated by the inability of any individual study to pronounce on the question at hand. Combined, the studies may reach statistical significance where individually they do not.

Meta-analysis requires identification of studies as well as their integration. Because of potential publication biases, identification of studies may not be a simple task. Casual approaches to identification of studies may lead to biases in the meta-analysis (Sheldon, 1996). The first step in a meta-analysis is to identify *all* studies available in the area of interest. Some of these may never have been published. The lack of publication may be the result of poor study design, unattractive results or negative result publication bias. With the latter two reasons, studies should still be examined for inclusion in the meta-analysis. Such studies may be crucial in their indications that run counter to the published literature.

Once studies are collected, they must be reviewed according to explicit criteria for determining quality of the study and ultimate suitability for the meta-analysis. This is often done by a 'blinded' reviewer (authors and journal references deleted for the reviewer) to avoid bias. The criteria for inclusion often include the relative comparability of studies (experimental, nonexperimental, blinded, randomized, etc.). The use of explicit criteria is to add to the reproducibility of the meta-analysis – one attempts to eliminate or reduce the subjectivity of reviewer decisions on the suitability of the studies.

Once studies are selected, the data must be abstracted from them in order to be combined in the meta-analysis. This too should be systematized to ensure reliability (specifying how one will deal with missing data for example). Lastly, statistical analysis of the combined data is done according to any of several methods (Petitti, 1994). Part of the statistical procedure is to check for 'homogeneity.' As L'Abbe et al (1987) have noted,

> *'an underlying assumption in combining individual study results to arrive at a summary measure is that their differences are due to chance alone (sampling variation), and therefore all study results are homogeneous, that is they reflect the same 'true' effect. In other words, when results are combined, random error cancels out and 'n' [study] results are better than one.'*

The check for homogeneity involves formal statistical techniques; however, the power of these tests tends to be low because of the typically small number of studies involved in a meta-analysis. L'Abbe et al (1987) suggest therefore when formal tests fail to reject the homogeneity assumption (the likely result with a low power) that 'informed subjective judgement' be utilized. When the common measure of outcome in the studies within the meta-analysis is judged to be homogeneous, a summary measure can be derived from pooling the results. L'Abbe et al (1987) provide a checklist for rating meta-analyses that is reproduced as Table 8.

Table 8 Checklist to evaluate quality of meta-analysis

Is there evidence of a working protocol?
Are literature search strategies explicitly described?
Are inclusion and exclusion criteria specified, papers included and excluded listed, and reasons given for exclusions?
Are visual displays and tests of homogeneity done?
Are appropriate statistics and sensitivity analyses employed?
If the pooled analysis shows significant differences, is the issue of publication bias addressed?
Are conclusions drawn for treatment recommendations (beneficial, equivocal, harmful) and for future research?

(L'Abbe, et al., 1987)

The lack of homogeneity of results between studies is considered to be a problem in traditional meta-analysis. In a new type of study that integrates results across other studies, lack of homogeneity is not a problem. Indeed, it is almost a goal in itself. More accurately, the goal is heterogeneity of study *design;* the homogeneity of the different study *results* is merely an interesting outcome not a goal in itself.

Cross design synthesis

Meta-analytic methods frequently combine evidence from one type of study design. Most meta-analysis in the medical area have been based on the RCT. However, as argued earlier, this design may not be ideal for the purposes of economic evaluations. It follows that a meta-analysis based solely on RCT studies may be inadequate as well. Cross design synthesis (Droitcour et al, 1993; US GAO, 1992) is designed to solve this problem. Cross design synthesis is a new form of meta-analysis that explicitly sets out to exploit the differences in study designs and their strengths with the intention of eliminating the bias inherent in any one particular study design. It is another type of model that can serve as input to economic evaluations. Unlike most meta-analyses, cross design synthesis studies are chosen for complementarity of research design so that one type of bias may be eliminated by exploiting the strengths of another.

Since RCTs have a great strength in their internal validity, but potentially poor external validity, and databases, for example, may be subject to selection biases, but do make observations under more naturalistic or routine practice conditions, these types of studies are considered complementary in design. One's strength is the other's weakness. Table 9 summarizes this concept.

There are two steps in cross design synthesis. The first is to assess the overall quality of the studies identified and the second is to incorporate a 'focused assessment' and choose studies based on their bias and the possibility of its elimination through use of complementary design. This second step – the goal of elimination of bias – is the hallmark of cross design synthesis that distinguishes it from the general field of meta-analysis. As an example, to pronounce on the generalizability of the results from an RCT, cross design synthesis would examine the way in which patient recruitment was accomplished and compare

Table 9 Two complementary study designs

Design	*Primary strength**	*Primary potential weakness**
Randomised controlled trials (RCTs)	Patients randomly assigned to treatments; controlled comparison	Not representative of full range of patients and treatment implementations
Database analyses	Coverage of medical practice (full patient population; all implementations)	Uncontrolled and imbalanced comparison (treatment assignment bias)

(from Droitcour, et al., 1993)

*Strengths and weaknesses according to the research question: Does the treatment 'work' in medical practice?

the representativeness of that sample with the population of interest. It would examine the inclusion/exclusion criteria, the investigators' choices of eligible patients, the patients willingness to participate once selected, etc. If the sample differed greatly from the target population, this would supply a initial basis for further investigation for potential bias. In terms of adjusting results for bias, age or sex linked results could be examined and a trial sample's results reweighted according to population representativeness if it differed in that regard.

Cross design synthesis is a relatively new field with interesting potential, but with considerable development still required. It requires investigators to exercises judgment in many situations, rather than take and apply an objective method off the shelf. This appears as a weakness, but one shared by the entire modelling discipline. Whether one can still be said to be engaged in scientific enterprise is an issue for debate. Policy makers and other customers for economic analyses and modelled results may cringe at the potential for subjectivity. It is to this issue that we now turn.

7. Modelling or meddling?

'...discovery commences with the awareness of anomaly.'

– T. KUHN (Thaler, 1992, p. 5)

'...now heaven knows, anything goes.'

– COLE PORTER (1934)

It should be clear that current practice in the evaluation of medicines for purposes of registration is less than ideal when viewed from the perspective of an economic analyst. What are the options available for solving this problem? One can redesign trials so that they are measuring what economists and those interested in economic results would like to see measured. Current and (realistically) future regulations of the clinical development process will prevent this if financial concerns do not. While RCTs can be modified to *better* measure what is of interest for economics, their structure (particularly in a pre-marketing environment) effectively prevents many of the aforementioned problems from being completely addressed. Alternatively, one can wait to perform economic evaluations until products are marketed, and either conduct less rigidly controlled trials or observational studies as a substitute. This solution suffers from a timing problem – the economic information is needed by all interested parties at the time of marketing, not several years afterwards. Furthermore, such trials or observational studies do not eliminate bias, they merely reduce it or alter the type of bias.

This monograph has described the concepts of outcomes and economic modelling to attempt to correct for the basic deficiencies of evidence in pre-marketing RCTs. It has presented the case for not only a continuing but an expanded role for modelling in the economic analysis of pharmaceuticals. It would appear that Eddy's 'flexible, but firm' approach (Eddy, 1990) to technology assessment has much appeal. It will enable economic evaluations to proceed on the basis of whatever evidence exists at a given time – with the credibility that results from the types of evidence available and the way in which it is used. The only way to obtain accurate answers to these questions is for modelling to continue and expand in its use.

Of course nonexperimental methods of outcomes research and modelling efforts which can be as casual as educated guesses or as formal as a sophisticated formal epidemiology study are subject to criticism. One researcher, Peto, indicated his adherence to the RCT paradigm in stating his opinion of the substantial investment in recent years in nonexperimental methods of 'outcomes research' as 'worse than just destroying the money because it gives the illusion of information (Anderson, 1994).' The RCT has served and continues to serve in one particular capacity quite well. That it does not serve as well in other capacities should not surprise anyone. The apparent lack of assailability in deployment to answer one set of problems should not blind its champions to its flaws when applied elsewhere. The RCT too is a model – with numerous assumptions that may or may not reflect the reality of medicine taking in the population eventually destined to receive the product. These assumptions may affect the validity of the clinical conclusions, but also the economic ones (even if the clinical conclusions are relatively accurate).

Modelling is not a panacea. The different results obtainable by using different intermediate endpoints in a model, for example, are due to some being good predictors of final outcomes and others failing in that regard. To the extent that the chosen endpoints are faulty, any model built upon them will be flawed. Despite the potential strengths of modelling, we have to expect this to occur from time to time. In Chapter 3 we discussed the TIMI-I trial and its choice of intermediate endpoints that led to remarkable overestimates of the superiority of tPA over streptokinase. Later analysis of mortality data indicated either a much smaller superiority for tPA or in fact none at all, depending on the statistical procedures used. A more serious case involving intermediate endpoints occurred with drugs used for cardiac arrhythmia. The endpoint chosen for early study was suppression of premature beats, an indicator of electrical instability in the heart. The strength of faith in the endpoint and the drugs themselves was so great as to almost ethically rule out the trial that ultimately tested them. This trial was in fact prematurely terminated when the treated group of patients was found to have an excess mortality rate over placebo (Moore, 1995). It is clear that models that translate intermediate endpoints to final ones must be based on solid foundations and, even then, may occasionally err. Models will, of course, not always be correct. They purport to provide (when done honestly) a 'best estimate' given

the currently available information. That they may later be shown to have been in error, is entirely possible. Models buys results in a timely and relatively inexpensive manner – the price can sometimes be inaccuracy.

Models have a great potential for being relatively subjective. One of the reasons for using the RCT is to avoid the dangers inherent in casual observation and the errors of subjectivity. When we leave behind the objective RCT for the less controlled techniques of modelling, we also create a great potential for encouraging pseudo-science. While some see the anomaly of health economic research using tools ill-suited to its purpose as a challenge for methods development and discovery, the same flexibility demanded in legitimate modelling efforts can be interpreted as exploratory license by others – 'anything goes.' The RCT has many advantages, one of which is the relatively secure paper trail of a protocol detailing primary endpoints, clear records of all variables collected on patients, etc. Modelling exercises appear to toss aside these strengths and rely at some level on speculation as to 'reasonable' values for costs or frequency of events not measured in the trial. There is no paper trail for the choices so made. There is no clear commitment up front to a particular assumption or set of assumptions. Rather one may try many assumptions and choose on whatever basis is desired the one which will be presented. Data mining (or assumption mining, its less limiting relative) offers considerable potential return. Thus while modelling offers the greatest potential for obtaining accurate results for economic evaluations when done sensibly, it also offers the potential for subjectivity and/or abuse or just plain old inaccuracy. See Sheldon (1996), Luce (1995) and Buxton et al (forthcoming) for further discussion of the problems with modelling.

Does 'anything go' in modelling? There is a potential for wildly inaccurate results based on the legitimate or illegitimate assumptions of modelling studies. Illegitimate assumptions can be generated through active attempts to mislead in a market where such studies may determine the movement of significant funds toward or away from particular products. They may also be generated through incompetence on the part of analysts. This may be particularly dangerous in the current environment where the casual observer hears many opinions offered as to the necessary (usually described as minimal) talent to do 'pharmacoeconomics'. Some analysts who would never presume to pretend to qualification in other fields of scientific inquiry apparently have few qualms at pretending to this throne (one does not observe similar casual claims to expertise in biochemistry or physics). Total (or near) ignorance of the history of development of analytic method in economics leads some of these people to 'make it up as they go' with a finger in the wind of methodology to judge the wisdom of their chosen method. It seems fine, and they adopt it, ignoring the (unknown) fact that they proceed in the direct contradiction to accepted method that may have been acknowledged in the economics or decision sciences literature half a century ago. It is clear that modelling is only as good as the person doing it; Schecter (1993) has indicated several examples of where authors implemented some form of decision analysis based on what may have appeared to be reasonable assumptions to the intelligent layperson, but which amounted to nonsense from a scientific and analytic perspective. This type of result is quite likely when one has persons untrained formally in analytic methods attempting to perform tasks beyond their talents. Moreover, many of these practitioners appear to be painfully unaware of their ineptitude – modesty is not a common trait. To paraphrase the American comic pundit Will Rogers, 'the problem isn't what they don't know, but what they know for sure that just ain't so.'

The current undersupply of high quality talent in the economic evaluation field appears to imply that this situation will not change rapidly. This situation lends credence to those who might conceive of the activities proposed herein more as meddling than modelling. The casual approach to the field exhibited by some has raised significant concern among many parties and may be at least a partial explanation for the movement toward the 'standardization' of methods (Hillman et al, 1991; CCOHTA, 1994). In the quest for a better vehicle to assist in answering the questions posed in economic evaluations of medicines can the modelling solution offer any guarantees that it will not be consciously or unconsciously abused? If not, pricing/reimbursement authorities and other 'customers' for economic evaluations will be slow to accept the modelling efforts and will perhaps be encouraged only by complete transparency of method and perhaps free availability of the data utilized.

In a recent draft document (intended, in its current form, for discussion rather than policy), the Division of Drug Marketing, Advertising and Promotion at the US FDA has suggested certain 'principles' for its review of economic evaluations used as promotional material by pharmaceutical companies (DDMAC, 1995). It is clear from that document that there is some level of awareness of many of the issues presented in this monograph. The US FDA intends there to be constraints put on economic analyses used for promotional activity. It is not at all clear at this point that any constraints will be fully transparent.

The draft document is quite vague (perhaps appropriately so at this stage). It is unclear whether 'modelling' is acceptable or not. Rather, it is clear that it might or might not be acceptable. Little guidance is currently proposed as to what will define the acceptability. The document states that 'models to provide estimates of PE [pharmacoeconomic] parameters should be used only when it is impractical or impossible to gather data using adequate and well-controlled studies (DDMAC, 1995, p.4). This appears to cast doubts on the future of modelling, at least for the FDA. However, other areas of the document appear to rely (without saying so explicitly) on modelling in their calling attention to issues of external validity. Depending on one's definition of 'impractical' or 'impossible', this suggested constraint on modelling may or not be limiting. If it is acknowledged that RCT results are insufficient, then it may be a mere corollary that it is 'impossible' to gather the *necessary* data using *only* adequate and well-controlled studies. The DDMAC document clearly recognizes the distinction between efficacy and effectiveness and the dual need for both external as well as internal validity in studies. What is necessary is that the document be clearer on what is or is not acceptable so that analysts have a known set of expectations before them. In many ways the current document stands as a wish list for external validity and internal validity and scientific rigor. In a way, who could object? Implementation is where it will matter. How can one satisfy the wish list? Will the attempt be acceptable?

The predominant general sentiments offered are that particular activities 'may be' acceptable or necessary, and that 'scientific rigor' is important. Hopefully the draft document will be extensively modified prior to its becoming policy so that what constraints remain will be valid and clear for both FDA personnel responsible for implementation and those they will be regulating.

Many analysts and reviewers are inexperienced in this area. There will be continuing pressure not to allow such modelling. This would be a mistake. That said, the concerns of those expressing doubt are not groundless. The well-known phrase, 'garbage in; garbage out' applies to models as to many other endeavors. One solution is a high degree of transparency in method and assumptions and perhaps data availability (at least to peer reviewers for journals, expert users, or DDMAC reviewers at the FDA). Rather than make a blanket policy against such models, perhaps a more rigorous and transparent review policy (with commissioned responses by qualified reviewers?) would be appropriate.

Ultimately, good (and bad) analyses will show themselves to be so either via the scrutiny of qualified reviewers or users, particularly larger sophisticated users with their own in house expertise, or through ensuing battles in the marketplace between analysts attacking each other's methods and/or assumptions. Meanwhile a good dose of scepticism is probably a good accompaniment to a review of *any* economic model for they will all of necessity be built upon a foundation of assumptions, some more heroic than others, depending on your perspective. A critical attitude can go a long way in the quest for accurate assessments of economic analyses. Transparency on the part of analysts would appear to be a necessary (though perhaps not sufficient) condition for signaling quality (or its absence). Models that lack transparency might *de facto* be suspect, if not from an analytical perspective, at least from a marketing one – why should you believe what you do not understand? Of course, one person's transparency is another person's superfluous detail – the mission of publication in scientific journals may be difficult to reconcile with some interpretations of transparency. *Caveat emptor* is a phrase that has stood the test of time. There is no reason to consider it an outmoded aid. Ultimately, education of users to the potential invalidating subjectivity of some models is the answer. Understanding how to review and critique a model contributes immeasurably to the goal of not being misled by them. Armed with that capability allows one to relax when confronted by a model, comfortable that it no longer is an uninterpretable 'black box'. Of course, this level of

knowledge and understanding does not come cheaply, but to invoke another time-honored phrase, 'if you think education is expensive, try ignorance'. The stakes are high and it must be recognized that models can both illuminate or obscure. True understanding requires an effort on the parts of both the modellers and the users of models – transparency and education.

The opposition to deploying other methods to answer the very practical questions of economic evaluations of medicines is disintegrating. While we should heed warnings of critics regarding potential validity problems, particularly the potential for subjective (and selected) assumptions to creep into modelling efforts (see Sheldon, 1996), we should not prevent the development of the field by adhering to methods unsuited for the issues at hand. That there will continue to be legitimate (and not) objections to these changes is as true as it is irrelevant. Serious issues and opinions should not be changeable overnight – they could not have been very seriously or reasonably held if they could be. The developments in this field must move onward, acknowledging and addressing, but not wallowing in, the warnings of critics. Quality of analysis will ultimately decide the issue. The exchange of debate and the continuing education of analysts and customers for evaluation results is very important in furthering the field as rapidly as possible to meet current and future challenges in health care priority setting. The Nobel laureate Max Planck once stated, 'a new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents finally die, and a new generation grows up with it.' Hopefully the future of economic evaluation method need not be this pessimistic.

References

Anderson C (1994) Measuring what works in health care, *Science;* vol. **263** p1080-1082

Bell D E, Raiffa H, Tversky (1988) A Descriptive, normative and prescriptive interactions in decision making in Bell D E et al (eds.) *Decision Making: Descriptive, Normative, and Prescriptive Interactions,* Cambridge University Press: Cambridge

Bone R C, Fisher C J Jr., Clemmer T P et al (1987) Controlled clinical trial of high-dose methylprednisolone in the treatment of severe sepsis and septic shock, *N England J of Med* **317** p653-658

Briggs A (1995) Handling Uncertainty in the Results of Economic Evaluation, *OHE Briefing,* London: Office of Health Economics, No. 32, September

Brody B (1995) *Ethical Issues in Drug Testing, Approval, and Pricing: The Clot Dissolving Drugs,* New York: Oxford University Press

Brophy J M and Joseph L (1995) 'Placing Trials in Context Using Bayesian Analysis: GUSTO Revisited by Reverend Bayes, *JAMA,* 273(ii); 871-75

Buxton M J et al (forthcoming) Modelling in Economic Evaluation: An Unavoidable Fact of Life. Submitted to Journal of Clinical Epidemiology

CCOHTA (November 1994) Canadian Coordinating Office for Health Technology Assessment, *Guidelines for the Economic Evaluation of Pharmaceuticals: Canada – 1st ed.* Ottawa: CCOHTA

Commonwealth Department of Human Services and Health (November 1995), Guidelines for the Pharmaceutical Industry on Preparation of Submissions to the Pharmaceutical Benefits Advisory Committee, Canberra: Australia Government Publishing Service

Cook T D and Campbell D T (1979) *Quasi-experimentation: Design and Analysis Issues for Field Settings* Boston: Houghton Mifflin Co

Culyer A J and Maynard A K (1981) Cost-effectiveness of Duodenal Ulcer *Treatment Social Science and Medicine* 15c p3-11

Culyer A J and Wagstaff A (1993) QALYs versus HYEs, *Journal of Health Economics,* **11** p311-323

DDMAC (1995) *Principles for the Review of Pharmaceutical Promotion* (DRAFT), March 20, 1995, FDA: Washington DC

Doubilet P, Begg C B, Weinstein M C, et al (1985) Probabilistic Sensitivity Analysis Using Monte Carlo Simulation: A Practical Approach, Medical Decision Making, 5:157-77

Droitcour J, Silberman G, Chelimsky E (1993) Cross-Design Synthesis: A new form of meta-analysis for combining results from randomized clinical trials and medical-practice databases, *International Journal of Technology Assessment in Health Care* **9:3** p440-449

Drummond M F, Stoddart G L, Torrance G W (1987) *Methods for the Economic Evaluation of Health Care Programmes,* Oxford: Oxford University Press

Drummond M F (1992) Basing Prescription Drug Payment on Economic Analysis: The Case of Australia *Health Affairs* (Winter) 11; **150** p191-196

Drummond M F, Torrance G W, Mason J (1993) Cost-effectiveness League Tables: More Harm than Good? *Social Science and Medicine* **37**(1) p33-40

Dworkin R (1994) Will Clinton's Plan Be Fair? *New York Review of Books,* Vol. **XLI**, No. 1 & 2, January 13

Eddy D (1990) Should We Change the Rules for Evaluating Medical Technologies? in Gelijns A (ed.) *Modern Methods of Clinical Investigation* National Academy Press, Washington, DC

Eddy D (1992) *Assessing Health Practices & Designing Practice Policies: The Explicit Approach* (American College of Physicians, Philadelphia)

Eddy D (1982) Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities in Kahneman et al (eds) *Judgment Under Uncertainty: Heuristics and Biases* Cambridge: Cambridge University Press

Eisenberg J M, Glick H, Koffer H (1989) Pharmacoeconomics: Economic Evaluation of Pharmaceuticals in Strom B (ed) *Pharmacoepidemiology,* Churchill Livingstone, New York

Finkler S (1982) The Distinction between Costs and Charges *Annals of Internal Medicine,* **96** p102-109

Freedman D, Pisani R, Purves R, Adhikari A (1991) *Statistics* (2nd Ed.) New York: W W Norton & Co

Fryback D (1993) QALYs, HYEs, and the Loss of Innocence *Med Decis Making* **13** p281-286

Gafni A and Birch S and Mehrez A (1993) Economics, health and health economics: HYEs versus *QALYs Journal of Health Economics* **11** p325-339

Garber A M, Littenberg B, Sox H C, et al (1989) *Costs and Effectiveness of Cholesterol Screening in the Elderly* Washington DC Office of Technology Assessment, US Congress, April

Glass G V (1976) Primary, secondary and meta-analysis of research, *Educ Res* **5** p3-8

Gold M. R. et al (ed) (1996). Cost-Effectiveness in Health and Medicine. Oxford University Press

GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. N Engl J Med 1993; 329:673-82

Hillman A L, Eisenberg J M, Pauly M V et al (1991) 'Avoiding Bias in the Conduct and Reporting of Cost-Effectiveness Research Sponsored by Pharmaceutical Companies' *New England Journal of Medicine*, May 9, Vol 324, No 19 p1362-1365

Johannesson M, Plishkin J S and Weinstein M C (1993) Are healthy-years equivalents an improvement over quality-adjusted life years? *Med Decis Making*, **13** p281-286

Kahneman D, Slovic P, Tversky A (1982) *Judgment Under Uncertainty: Heuristics and Biases* Cambridge: Cambridge University Press

Kassirer J P, Angell, M (1994) The *Journal's* Policy on Cost-Effectiveness *Analyses New England Journal of Medicine*, Vol. **331**, No. 10, p669-670, September 8

L'Abbe K, Detsky A S, O'Rourke K (1987) Meta-analysis in clinical research, *Annals of Internal Medicine* **107** p224-233

Luce B R (1995). 'Policy Implications of Modelling the Cost -effectiveness of health care technologies', Drug Information Journal Vol 29 pp1469-1475

McNeil B J, Pauker S G, Sox H C, Tversky A (1982) On the Elicitation of Preferences for Alternative Therapies *New England Journal of Medicine* **306** p1259-1262

McNeil B J, Pauker S G, Tversky A (1988) On the Framing of Medical Decisions in Bell DE et al (eds) *Decision Making: Descriptive, Normative , and Prescriptive Interactions*, Cambridge University Press: Cambridge

Mehrez A and Gafni A (1993) Health-years equivalents versus quality-adjusted life years, *Med Decis Making*, **13** p287-292

Mehrez A and Gafni A (1989) Quality-adjusted life years, utility theory and healthy-years equivalents, *Med Decision Making*, **9** p142-149

Meinart C, Tonascia S (1986) *Clinical Trials: Design, Conduct and Analysis*, Oxford University Press, New York

Mencken H L (1978) *Washingtonian* **14** (November):155

Mooney G (1994) *Key Issues in Health Economics*, Harvester Wheatsheaf: New York

Moore, T J *Deadly Medicine*, New York: Simon and Schuster, 1995

NHS Centre for Reviews & Dissemination (1996). Undertaking Systematic Reviews of Research on Effectiveness. CRD Guidelines for Those Carrying Out or Commissioning Reviews. CRD Report Number 4, NHS Centre for Reviews & Dissemination, University of York

O'Brien B, Drummond M F, LaBelle R J, Willan A (1994a) In Search of power and Significance: Issues in the Design and Analysis of Stochastic Cost-Effectiveness Studies in Health Care, *Medical Care;* **32**(2) p150-163

O'Brien B, Anderson D R, Goree R (1994b) Cost-effectiveness of enoxaparin versus warfarin prophylaxis against deep-vein thrombosis after total hip replacement, *Canadian Medical Assoc Journal* **150** (7) p1083-1090

Patrick D L and Erickson P (1993). *Health Status and Health Policy* Oxford: Oxford University Press

Petitti D B (1994) *Meta-analysis, decision analysis, and cost-effectiveness analysis*, Oxford University Press: Oxford

Porter C (1934) *Anything Goes*

Poses R M, Anthony M (1991) Availability, wishful thinking and physicians' judgments for patients with suspected bacteremia *Medical Decision Making* **11** p159-168

Rittenhouse B and O'Brien B (1996) Threats to the Validity of Pharmacoeconomic Analyses Based on Clinical Trial Data in Spilker B (ed) *Quality of Life and Pharmacoeconomics in Clinical Trials* Raven Press, New York

Rittenhouse B (1994)Teaching the Tools of Pharmaceutical Care Decision Analysis, *American Journal of Pharmaceutical Education* Vol. **58** (Winter)

Rittenhouse B (1995) The Relevance of Searching for Effects under a Clinical Trial Lamp Post: A Key Issue, *Medical Decision Making*, 15:4, pp. 348-357

Rittenhouse B (1996) Another Deficit Problem: The Deficit of Relevant Information when Clinical Trials are the Basis for Pharmacoeconomic Research, *Journal of Research in Pharmaceutical Economics* (Vol. 7, No. 3)

Russell L (1994) *Educated Guesses: Making Policy about Medical Screening Tests*, University of California Press, Berkeley

Schecter C B (1993) Decision Analysis in Formulary Decision Making, *PharmacoEconomics*, **3** p454-461

Schelling T C (1984) Economic reasoning and the ethics of policy in Schelling TC (ed) *Choice and Consequence: Perspectives of an Errant Economist*, Cambridge, MA: Harvard University Press, p1-26

Senn S (1990) Clinical Trials and Epidemiology, *J Clinical Epidemiology* Vol **43** p628-632

Sheldon T A (1994) Please Bypass the PORT *British Medical Journal* Vol. **309** p142-143

Sheldon T A (1996) Problems of using modelling in the economic evaluation of health care. Health Economics, Vol 5. pp1-11

Slovic P, Fischoff B, Lichtenstein S (1988) Response Mode, Framing, and Information-Processing Effects in Risk Assessment in Bell DE et al (eds) *Decision Making: Descriptive, Normative, and Prescriptive Interactions,* Cambridge University Press: Cambridge

Stokey E, Zeckhauser R (1978) *A Primer for Policy Analysis,* W W Norton & Co. New York

Strom B (1989) *Pharmacoepidemiology,* Churchill Livingstone, New York

Thaler R H (1992) *The Winner's Curse: Paradoxes and Anomalies of Economic Life,* Princeton University Press: Princeton

Torrance G W, Thomas W H, Sackett D L (1972) A Utility maximization model for evaluation of health care programs, *Health Services Research* **7** (2) p118-133

Tugwell P, Bennett K J, Sackett D L, Haynes R B (1985) The Measurement iterative Loop: A Framework for the Critical Appraisal of Need, Benefits and Costs of Health Interventions *Journal of Chronic Diseases,* Vol 38, No.4 p339-351

US General Accounting Office (1992) *Cross-Design Synthesis: A New Strategy for Medical Effectiveness Research* US GAO/PEMD-92-18, March 1992

Ziegler E J, Fisher C J, Sprung C L et al (1991) Treatment of gram-negative bacteremia and septic shock with HA-1A human monoclonal antibody against endotoxin New England Journal of Medicine, **324** p429-436

APPENDIX 1
Models of valuation: patient preferences and utility

'...it is pure myth that latently these probabilities and utilities existed deep down and that the analyst merely has to away the fat in order to display the pre-existing structure.'
– BELL et al (1988)

Early on in the economic assessment of medicines and other treatment programs, it was recognized that all treatment outcomes resulting in survival were not equivalent. The saving of life did not equate across all cases. Complete recovery was obviously different from recovery at a level less than that of full functioning, but how should these differences be accounted for? The pure measurement of quality of outcome is captured in the idea of quality of life (QoL) measures. However, many QoL measures lack the ability to tradeoff quality and quantity of life, making them less suitable to use in economic evaluations. Here we will concentrate on so-called generic measures of quality of life that are preference-based. We will examine methods that yield a single numerical measure of quality and are usable across disease categories. Readers interested in more broadly defined QoL measures as well as further discussion of those below are referred to Patrick and Erickson (1993).

Of all the ways in which models may be used in the economic evaluation of pharmaceuticals, modelling outcome valuations is perhaps the least controversial in terms of the consensus on the need for obtaining information outside the trial environment if the primary evidence for the model is an RCT. If the primary evidence is from retrospective analysis of trials or databases, there is simply no choice – one must model the valuation as no data would otherwise be available. When the primary data come from a prospective RCT, it is not as clear why the valuations should not come from the same source. Indeed it is possible to obtain valuations from those recruited to be in RCTs; it is simply that the sample of opinions may be too limited in such a sample and that the added requirements for clinical investigators may be problematic in a trial. If there is no necessity of obtaining this information from the same RCT patients, the additional burdens to both patients and investigators can be avoided.

One reason why the patients in RCTs need not be asked (or should not be asked) to make these valuations is that if it is public policy decisions (e.g. pricing, reimbursement, etc) that are to be informed by the analysis, then it is the views of the public that matter in terms of resource allocation and priority setting. More accurately, it is the views of the *informed* public that matter (Drummond et al, 1987). It is imperative that respondents to valuation questions be informed about the true nature of the health states they are valuing. It has been suggested that patients and health care professionals will be most knowledgeable about diseases and health states associated with them. Asking these individuals to value these states may be easier because these people are more quickly able to understand the important aspects of the outcomes, having experienced them or witnessed them in numerous cases. However, it may be the case that such individuals are biased in their views of the health states because of their experiences. Strategic interests may imply that responses may not be true as patients feel that their treatment or that of others like them may be influenced by their responses. Health care professionals may suffer from similar biases. Other nonstrategic biases may also be detected in the sample of respondents used. The experience of having a particular health state may influence the responses elicited as well. The interaction of many confounding influences on patient valuations of health states may imply that valuations are biased. Sociodemographic characteristics of health care professionals in particular may imply less generalizability to the general population (Drummond et al, 1987). If the valuations are being used to compare several treatment interventions as to their cost-effectiveness, it may be quite important to value outcomes in a uniform set of individuals, a criterion which patients chosen from across disease categories will not meet (Drummond et al, 1993).

There are disadvantages to valuations done with subjects who are less familiar with the states being valued. The descriptions of these states must be more detailed. There is significant evidence that the ways in which health states are described can have significant influences on responses. Researchers must take care to provide sufficient detail without overwhelming respondents with information. As essential as outcome valuation is to economic evaluations of medicines, the field treads on thin methodological ice. The process of preference elicitation – including not only rankings of health states according to value, but eliciting a *degree of preference* associated with each relative to the other – is a path strewn with pitfalls. Researchers must realize the potential biases in such preference assessment interviews (see Box 5 on page 37).

There are three ways in which a model can incorporate the valuations of health outcomes. One can use judgment, values from the literature or measurement on a sample of respondents (Drummond et al, 1987). Judgments can be reasonably accurate and may suffice in some cases to answer questions that are not particularly sensitive to small changes in valuations. Sometimes the literature will already contain valuations of health states from previous studies. The most appropriate way to obtain health valuations is to measure them directly. In this way one can tailor the valuations to the exact needs of the study. This is the more resource intensive alternative however.

Box 5 Framing bias in preference measurement studies

There are numerous pitfalls in the elicitation of patient preferences for health outcomes. One of the more interesting is referred to as 'framing bias'. McNeil and colleagues (1982) supplied some initial evidence of this problem in the health care arena though it was already established in other contexts in the psychology literature (Kahneman et al, 1982).

When patients were asked questions about their preferences for particular treatments described by the probability of treatment *success* (the success frame) they gave different responses than when asked the identical questions framed in terms of probability of treatment *failure* (the failure frame). It is important to note that this was identical conceptual information as the probability of success was simply one minus the probability of failure; yet, the responses differed. Interestingly, while the answers differed across the three categories of respondents questioned: patients, physicians and graduate students, each group exhibited the same direction of bias due to the framing of the questions. The latter groups had been considered to be less susceptible to this framing influence because of their training. A similar experiment was conducted in which a 'mixed frame' was also included where both survival and mortality frames were used (McNeil et al 1988, see Table 10). This survey was done in the US and in Israel with results confirming the earlier work reported below. Interestingly the mixed frame was somewhat similar to the mortality frame, perhaps indicating the dominance of mortality frame influences even in a balanced presentation.

Table 10 Percentage of respondents favouring radiation (over surgery) for lung cancer under three different framing formats

Framing format	*US sample*	*Israeli sample*	*total*
1. Survival	16%	20%	18%
2. Mortality	50%	45%	47%
3. Mixed	44%	34%	40%

Source: McNeil et al 1988

It should be clear that framing differences can yield different results potentially when the true preferences are not in fact different. This provides an argument for standardizing approaches to preference assessments so that differences attributable to these biases are not mistakenly attributed to treatments. Of course, deciding which method to standardize can be problematic as it is not clear what the correct method is. Also, commonality of method need not necessarily imply neutrality of method so that using the same method may imply a different ranking of treatments when using one method than another. More research needs to be one in this area.

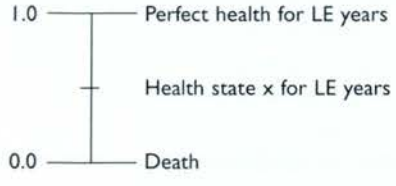
The problem is less acute within a single study where presumably the methods for assessing one treatment would be the same as another than between studies where common methods would be less likely. Such problems are particularly acute in the formulation of so-called League Tables where the cost-effectiveness results of a series of treatments are ranked from low to high. The value of these tables when a wide variety of methods has been used is questionable as it may be as much the variety of method that determines ranking as treatment effect and cost (Drummond et al, 1993).

Other biases in preference assessment also exist, and this should be considered in constructing models of patient preferences (McNeil et al, 1982). There is significant evidence that great care must be taken in performing and interpreting preference assessment studies. Slovic et al (1988) warn that the method of preference assessment may in fact be a major force in *shaping* the expression of respondents' values. While the economic evaluation field desperately needs such measurements, it needs to remain aware of the potentially fragile foundation many of them may be built upon.

There are three main techniques for assessing the preferences of respondents for health outcomes. These are the rating scale, the time-tradeoff (TTO) and the standard gamble (SG). Each is illustrated in Figures 10-12. In the rating scale method, respondents are instructed to make a mark on the indicated scale to describe where they feel the given health state ranks on a scale of death to perfect health. The distance between the mark and these anchor states is to represent the degree of relative preference between the health states being valued. Several health states can be ranked on one rating scale. The distances between these marks is translated into a 0-1 scale (death – perfect health) and 'utility' values are computed. Thus a health state corresponding to a mark halfway between death and perfect health would be assigned a utility value of 0.5.

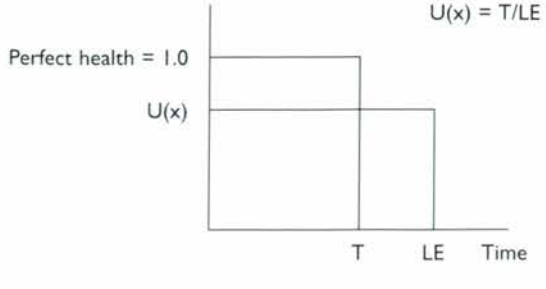
The TTO method was introduced as an approximation to the SG, in hopes that it would be easier to use as respondents sometimes have difficulty with the notions of probability inherent in the SG. It should be noted

Figure 10 Rating scale for measuring preferences for health states



however that the TTO method measures preferences (as does the rating scale) under conditions of certainty. This has been pointed out as a defect in these methods as there are theoretical arguments that medical outcomes (themselves frequently uncertain) should be assessed in terms of preferences for them under conditions of

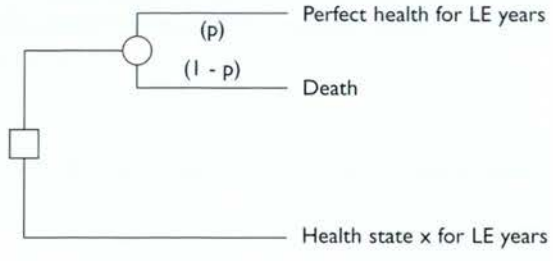
Figure 11 Time-tradeoff method for measuring preferences for health states



uncertainty. The SG method does this. Standard theory of economics suggests that preferences under certainty and uncertainty will not generally be the same, thus supporting the notion that the various 'utility' assessment methods are in fact measuring different things. Empirical studies of these methods do yield different answers to valuations of the same health states, lending some credence to this notion.

The TTO assesses the value of some health state x for the remainder of one's life. It accomplishes this by asking a hypothetical question of willingness to trade that health state and its duration for the state of perfect health for a lesser number of years. The more debilitating the health state being valued, the more years of life one would be willing to give up in order to attain perfect health for those years. If life expectancy is LE years in health state x and someone will trade this for T (<LE) years of life in perfect health, then the 'utility' of health states x is assumed to be representable by the fraction T/LE.

Figure 12 Standard gamble method for measuring preferences for health states



The standard gamble asks respondents to compare the health state being valued with a lottery or gamble between instant (painless) death and perfect health for the same number of years as health state x. The gamble has probabilities associated with it that make it relatively attractive or unattractive compared to health state x. The worse is health state x, the more acceptable will be the prospect of a higher probability of death to avoid the health state. The patient essentially chooses a probability for the perfect health outcome so that they are indifferent between the gamble and heath state x. The SG is built upon economic theory that claims that the satisfaction associated with a lottery can be represented by a so-called certainty equivalent to the gamble. When a respondent is indifferent between the certain event and the gamble, the utility can be represented by the probability of the perfect health outmode. The details are beyond the scope of this account, but can be found in Torrance (1972). As the SG is the only one of these

three methods that has any grounding in theory to support it, it is considered by many economists to be the gold standard.

Regardless of the method used to measure utilities, the conventional way of using these utilities is to convert them into quality-adjusted life years (QALYs). This is done by multiplying the utility value by the years spent in that health state. Thus 10 years in a health state valued at 0.5 would result in 5 QALYs and be assumed to be equivalent to 5 years of perfect health. Figure 13 shows the calculation of the QALYs saved as part of an intervention. Here it assumed that the relevant alternative is no treatment and that U(NT) corresponds to the utility and LE_{NT} to the length of life under that alternative. The QALYs for the no treatment alternative are therefore U(NT) x LE_{NT} or the area of the box with sides equal to U(NT) and LE_{NT}. U(T) and LE_T correspond to the utility and length of life under treatment with the QALYs, corresponding to U(T) x LE_T or the area of the box with sides equal to U(T) and LE_T. We can see that the difference is the QALYs gained by the treatment and is represented by the cross-hatched area. Mathematically this is equal to

$$\text{QALYs gained} = \{U(T) \times LE_T\} - \{U(NT) \times LE_{NT}\}.$$

The cost of achieving this outcome would need to be compared to the costs of no treatment to determine the incremental cost-effectiveness ratio used to determine whether the intervention is worthwhile (Drummond et al 1987). This particular type of cost-effectiveness analysis goes by the name of cost-utility analysis because of the incorporation of valuations of outcome in the form of utilities. Table 11 shows an example of various utility values for different disease areas.

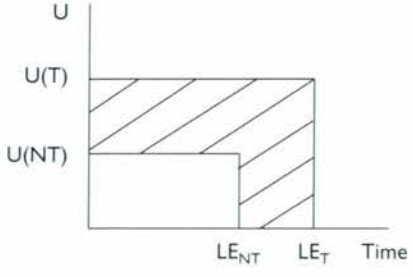
While the SG is accepted (in the field of economics) as a theoretically superior alternative to other methods of utility assessment (in fact 'utility' is a concept from economics and is by definition only valid if derived from the SG), the QALY construction stands on weaker foundation. An alternative, the Healthy-years equivalents (HYE) has been suggested to address some of the problems (Mehrez and Gafni, 1989). This alternative has in turn been questioned (Culyer and Wagstaff, 1993; Johanesson et al, 1993), but the literature controversy appears to be unresolved at the time of this writing (Gafni et al, 1993; Mehrez and Gafni, 1993).

Table 11 Two complementary study designs

Duration	*Health state*	*Mean daily health state utility*
	Reference state: perfect health	1.00
3 months	Home confinement for tuberculosis	0.68
3 months	Home confinement for an unnamed contagious disease	0.65
3 months	Hospital dialysis	0.62
3 months	Hospital confinement for tuberculosis	0.60
3 months	Hospital confinement for an unnamed contagious disease	0.56
3 months	Depression	0.44
8 years	Home dialysis	0.65
8 years	Mastectomy for injury	0.63
8 years	Kidney transplant	0.58
8 years	Hospital dialysis	0.56
8 years	Mastectomy for breast cancer	0.48
8 years	Hospital confinement for an unnamed contagious disease	0.33
Life	Home dialysis	0.40
Life	Hospital dialysis	0.32
Life	Hospital confinement for an unnamed contagious disease	0.16
	Reference state: Dead	0.00

(from Drummond, Stoddart & Torrance, 1987)

Figure 13 QALYs gained by treatment over no treatment alternative



APPENDIX 2
Examples of modelling to adjust results

'Investigators seem to have settled for what is measurable instead of measuring what they would really like to know'
– E Pelligrino (Meinart, 1986)

In many cases the economic researcher is faced with a previously collected set of data or at best the prospect of being allowed to 'piggyback' onto an RCT which is primarily intended to serve other purposes. Measurements of essential variables may or may not be made. If made, they may be made in a way unsuited for the economic analyst's purpose. Models can be used to correct for some of these problems. Two examples are provided here to illustrate these ideas.

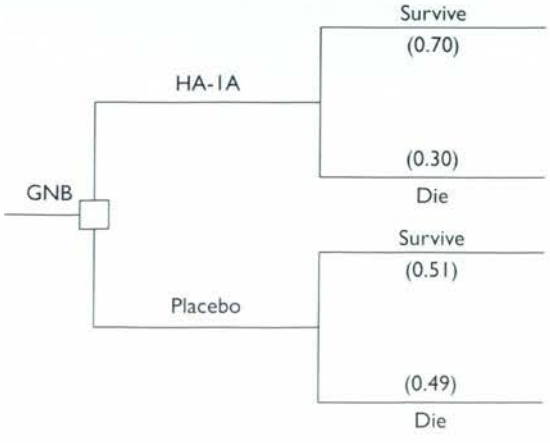
HA-1A in suspected sepsis

HA-1A was a product designed to be effective in sepsis (specifically in gram-negative bacteremia cases – GNB). An early trial showed it to be effective in this subgroup, and Rittenhouse (1995) indicated the potential effects of modifying the assumptions of the analysis to incorporate the likely differential abilities of trial investigators and physicians in less controlled environments.

The drug must generally be administered prior to diagnostic confirmation of GNB. This implies that in many cases it is likely to be given when the patient could not benefit since the prevalence of GNB in cases of suspected sepsis is relatively low as compared to gram-positive or non-bacteremic causes (nGNB). In an early trial of HA-1A plus conventional antibiotics compared to placebo and conventional antibiotics, Ziegler et al (1991) showed an impressive GNB subgroup analysis summarized in Figure 14 below.

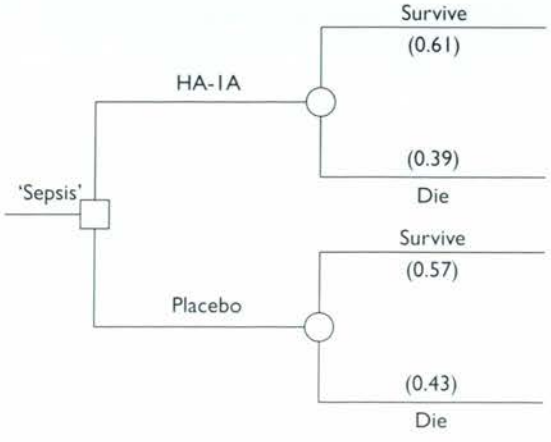
The subgroup analysis indicates that there is a 19 per cent increase in survival in the active arm. This was statistically significant and certainly appears to be clinically significant. To base a cost-effectiveness analysis on such results would however be highly misleading since the drug must be given to many more than just the GNB subgroup. However, another aspect of the trial also could result in highly misleading economic analysis. The trial used fairly restrictive inclusion/exclusion criteria, partially to inflate the percentage of GNB cases in its study sample. This is perfectly legitimate in a trial as it will reduce the sample necessary to show statistically significant effect. It also reduces exposure of patients who are unlikely to benefit from the inconvenience of participating in a trial. However, for economic evaluation purposes it is potentially highly misleading to use these results unless the drug would be used in as highly restrictive manner in the non-trial population once it were marketed. This is generally not likely. In the specific case of HA-1A, it is even less likely because of the potential legal liability (in the US) if the restrictive use patterns deprived a patient, eventually shown to be GNB, of access to the drug. More importantly however, might be the general inability of physicians to diagnose bacteremia or GNB. This may lead to significant errors in use of the product – many more than in the trial will not be able to benefit, but will incur costs. We note that this does not necessarily mean that the product is not effective for a subgroup of patients who are in fact GNB; it does mean that the costs per unit of desired effect will be higher however. This may imply that the product, though effective in a subgroup, is not cost-effective. This is largely because of the inability to identify that subgroup in a timely manner. The analysis below indicates the value of a rapid diagnostic ability paired

Figure 14 Trial results (subgroup)



Source: Rittenhouse, 1994

Figure 15 Trial results (entire sample)



Source: Rittenhouse, 1994

with such a drug – its cost-effectiveness would be vastly improved.

In fact in this case, there is even more reason to develop the diagnostic test. The costs are not the only variable affected by the improper use of the drug in non-GNB patients. Unfortunately, it appears that nonGNB cases are adversely affected by HA-1A. They are not merely not helped, they appear to have a higher mortality rate than the placebo arm. Thus a rapid diagnostic ability that could in a timely manner differentiate between those likely to be helped versus those likely to be harmed by the product would assist in the cost-effectiveness of the product by affecting both the cost and the effect.

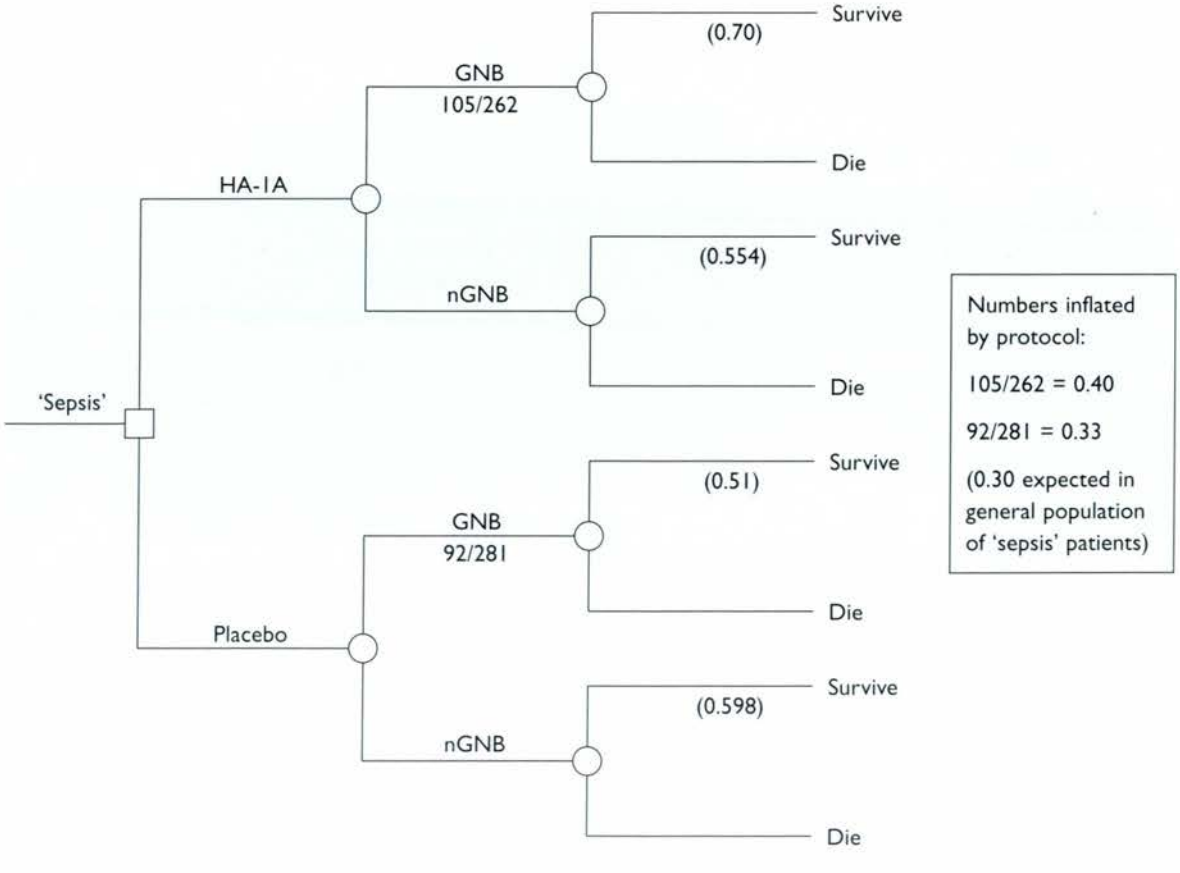
Figure 15 indicates the results for the entire sample. The beneficial effect for HA-1A is significantly altered from the sub-group case, with a four percent (not statistically significant) improvement over placebo.

Figure 16 provides more detail about these groups. Specifically it provides details on the frequencies of GNB in each treatment arm (.40 in the HA-1A arm and .33 in the placebo arm). Both proportions are higher than that which would be expected in the general 'sepsis' population (Bone et al, 1987). The overall survival rates for each treatment arm (second figure) are weighted averages of the survival rates of each sub-group, GNB and nonGNB:

P(S | HA-1A) =
P(S | GNB) P(GNB) + P(S | nGNB) P(nGNB) =
(0.70) (0.40) + (0.554) (0.6) = **0.61**
and
P(S | placebo) =
P(S | GNB) P(GNB) + P(S | nGNB) P(nGNB) =
(0.51) (0.33) + (0.598) (0.67) = **0.57**

These data represent conditional probabilities of survival based on using HA-1A or placebo and having GNB or

Figure 16 Trial results in more detail



Source: Rittenhouse, 1994

not. For example the first equation uses the conditional survival rates for GNB and nGNB for HA-1A times the probabilities of GNB and nGNB to achieve a weighted average of the probability of survival when HA-1A was administered in the trial.

In the nontrial world the proportion of GNB is expected to be around 0.30. This would cause the overall survival rate to fall in the HA-1A group as we can see if we substitute this proportion in for the weighted survival rate for HA-1A (we need to change the nGNB proportion as well). Both changes imply a decrease in the overall survival rate since the group with the higher survival rate (GNB) is being reduced in proportion to the group with the lower survival rate (nGNB).

Interestingly, the same changes in the placebo arm will have the opposite effect. This follows from the fact that the higher survival rate in the placebo arm occurs in the under-represented arm of nGNB. Thus this arm should in the nontrial world have a greater representation and would therefore imply an increased survival for the placebo arm as a whole. The combination of these changes would cause the overall survival rate differentiation between treatment and placebo (0.61-0.57) to shrink.

It is instructive to consider two extreme examples. What would happen if only GNB patients were to receive the product? In this case the subgroup analysis is valid and an extremely rudimentary economic analysis (looking only at drug costs) would indicate approximately a $21,000 cost per saved life. This is based on the estimated cost of the drug of $4,000 per administration (Rittenhouse, 1995), implying a treatment cost of $400,000 to 100 patients of which there would be an increased survival of 19 lives through using HA-1A.

Under a less optimistic view, if the probability of GNB fell to 0.20, the differential survival probability would fall to zero and the expenditure would result in no benefit at all – an unimpressive infinite incremental cost-effectiveness ratio! While it appears counterintuitive that the frequency of GNB would fall below its population expected value of 0.30, Poses (1991) provides data that indicate a fairly poor diagnostic ability of physicians in this indication area. Combined with other incentives (e.g. liability induced) such inappropriate levels of use can easily be imagined.

This example is instructive in several respects. Most importantly, it indicates the important effect that relative diagnostic abilities can have on the cost-effectiveness of products. The abilities in a trial with highly competent investigators and strict protocols for inclusion/exclusion criteria versus those physicians in the nontrial environment are potentially quite different. This difference can have a dramatic impact on the economics of the drug. Another important point brought out by

Box 6 The value of information

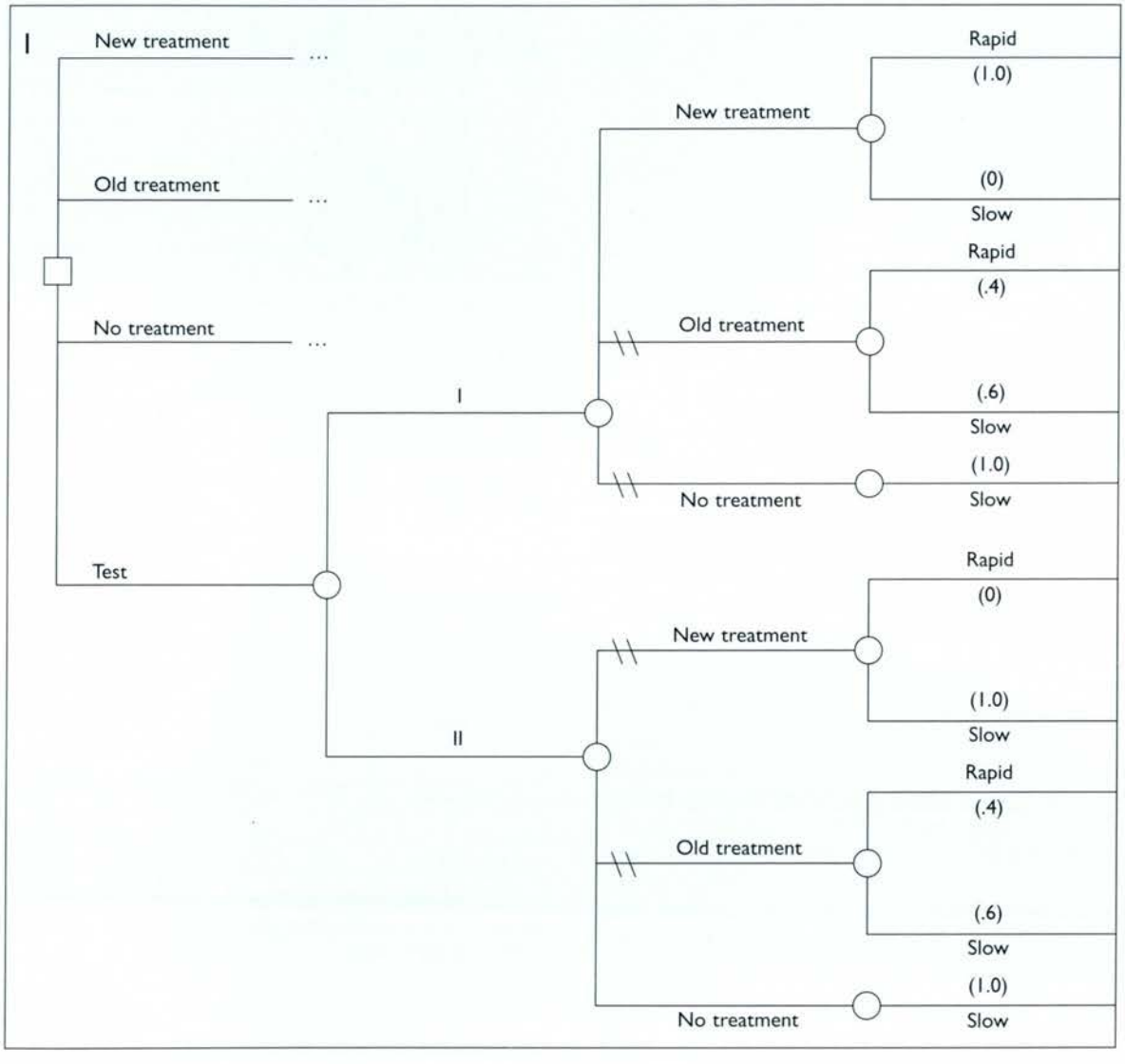
In Box 1 on page 10 we presented a very simple decision analytic model of treatment choice. Sometimes we can take our decision-analytic model and reduce the uncertainty in it – usually at a cost. Instead of just using a treatment without knowing what the effects will be except probabilistically, we can seek information that will reveal the optimal course of action. But is that information worth seeking? We would like to know the value of such information in order to determine whether its acquisition is 'worth it'. Here the example expands upon that of Box 1.

Recall from the earlier example that the choice of therapy was between new and old treatments, with the outcomes as rapid or slow recovery. We saw in that example that the therapy choices should be expanded to include a 'no treatment' option. We add now to that example. If a patient is of type I, then she will achieve rapid recovery with one hundred percent certainty under the new treatment. However, if she is of type II, she will (with certainty) have a slow recovery even with the new treatment. The old treatment works at the same probabilistic level as before in all types of patients. One can model the decision with this uncertainty (if we have some idea of the probabilities of being type I or type II) and determine the best course of action. Alternatively, one might have access to a diagnostic test that identifies the individual as type I or II. In the simple case outlined here, the decision to use such a test will depend on the cost of the test and the prevalence of type I and II individuals. Generally such a test may be worth using (or worth developing), depending also on the cost of making an error in giving the wrong drugs to the wrong people.

We can evaluate the testing strategy itself in a decision analytic framework by adding a choice to the existing tree. We now have an alternative choice consisting of testing the type of the patient and treating according to the results of the test. We can determine whether testing is ever warranted and we can determine how the cost of the test should influence our decision. Along the way we can explicitly determine the value of the information that the test supplies. We found in the earlier example in Box 1 that no treatment was the optimal result (with our sole concern being monetary costs). Will this remain the case?

Box 6 The value of information (continued)

Figure 17 Expansion of Figure 3 Simple decision tree



First note that once the test is performed, we know the optimal choice. That is, once we know the individual's type, we analyze a smaller decision problem – optimal treatment conditional on the type of individual. We can see this by analyzing the two 'sub-trees' coming from the branches marked 'I' and 'II' in Figure 17 above. These are analyzed as before except that they have a few different probabilities associated with the branches. If the individual is type I then the optimal choice of treatment will be the new treatment:

New treatment expected cost: (1.0) ($50) + (0) ($100) = **$50**
Old treatment expected cost: (.4) ($24) + (.6) ($76 = **$55.20**
No treatment expected cost: **$52**

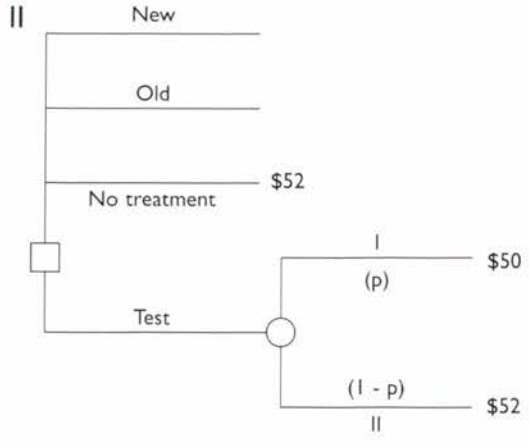
If the patient is type II, then no treatment is optimal:

New treatment expected cost: (0) ($50) + (1.0) ($100) = **$100**
Old treatment expected cost: (.4) ($24) + (.6) ($76) = **$55.20**
No treatment expected cost: **$52**

Since we know what we would choose once we find out the type of patient, we can eliminate all extraneous information from the decision tree, reducing it to that below in Figure 18. If the type of patient is type I, we will treat with the new drug at a cost of $50; if she is type II, we will not treat and it will cost $52. The

expected cost of the testing strategy (except for the cost of the test) is a weighted average of these two possible outcomes where the weights are the probabilities of type I and II patients.

Figure 18 Reduced decision tree



At this point we can see that if the test were free, we would use it, but if it were not free, we might need more information. This is because we want to compare the test strategy with the no treatment strategy (found to be optimal previously before we considered the possibility of testing). The no treatment strategy costs $52. We know that the test strategy (assuming the test is free) will have an expected costs that will be a weighted average of the two possible outcomes from the test ($50 and $52). Unless all patients are of type II, the weighted average of these two numbers must be less than $52, so we would use the test.

We know the value for the prevalence of type I and II patients. It must be .5, since the new drug was given an efficacy rate of .5 previously, and since it achieves either 100% success or failure, this implies a prevalence of .5 for each type of patient to be consistent with that earlier data. Knowing the prevalence, we can determine whether we should use the test for any given cost of the test. Alternatively (and more easily), we can determine at what price the test becomes cost-reducing in the model.

If we denote the cost of the test as T, then the cost of the testing strategy now becomes the weighted average (with .5 as each of the two constituent weights) of $50 + T and $52 + T, the outcomes of the test indicating a type I or II patient. The no treatment option is the best alternative in the absence of a test (as previously determined). The two strategies are equally optimal if their expected costs are identical. To find when this is the case, we simply set the two expected costs equal to each other and solve the resulting equation for the cost T that satisfies the equation.

$$\begin{aligned} EC\ (\text{No Treatment}) &= \$52 \\ &= (\$52 + T)\ (.5) + (\$50 + T)\ (.5) \\ &= EC\ (\text{Test Strategy}) \end{aligned}$$

This implies that when T is equal to $1.00, the two strategies are equally attractive. Therefore any value for T less than that would imply that the test strategy is superior.

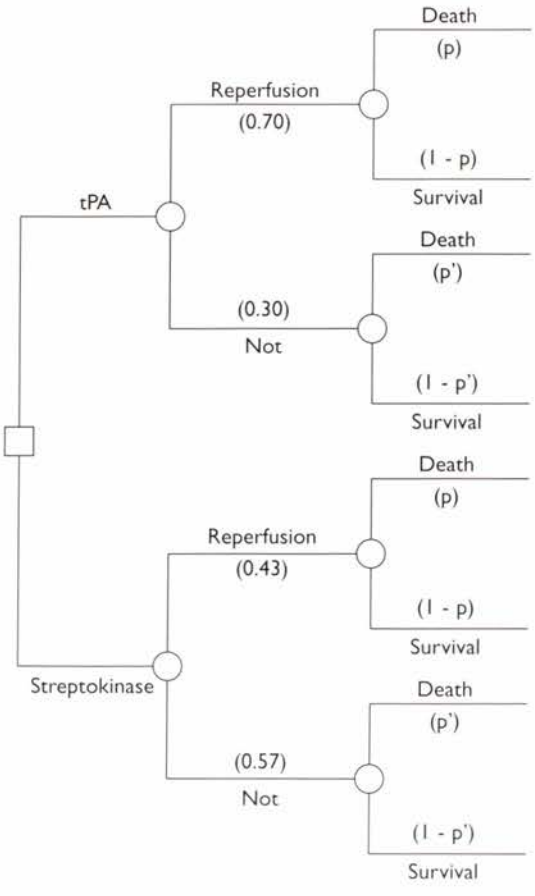
Such models can be used to determine when additional information should be sought. One can use such a model to determine whether one should seek more information in developing a cost-effectiveness model. How much is the information worth? Could the additional information make a difference of any important magnitude in the model, or would it be a waste of time and effort to get the added information? Effort might be better expended elsewhere. In the above case, we examined the case where the information provided was perfect. That is, it was definitive once the test was run. Many tests of course do not give a perfect indication of the situation. Imperfect tests would be analyzed in a similar manner, except that additional errors would need to be incorporated along with their expected costs.

Box 7 Modelling based on intermediate end points: tPA and streptokinase revisited

In the earlier discussion of tPA and streptokinase it was noted that intermediate endpoints could be highly misleading of both final health outcomes and costs of achieving them. Implicitly those who would use an intermediate endpoint have a model in mind where that endpoint translates into a measure of final outcome. In using the intermediate endpoint of the patency of the infarct -related artery, this 'model' was of dubious validity. The example shows that an economic model in such cases is critically tied to the quality of the linkages expressed between intermediate and final outcomes and the resource utilization assumptions employed in that linkage.

One could have developed a model of the study drugs that used this intermediate endpoint and some assumption of constant percentage of mortality reduction given that the reperfusion was or was not achieved. The model could have assigned resource utilization based on some conjectural evidence about what achieving or not achieving reperfusion implies for resource utilization and what might or might not happen afterwards. Such a model is illustrated in the figure below. It illustrates the importance of good outcomes assessment for economic evaluations (even independent of resources use modelling).

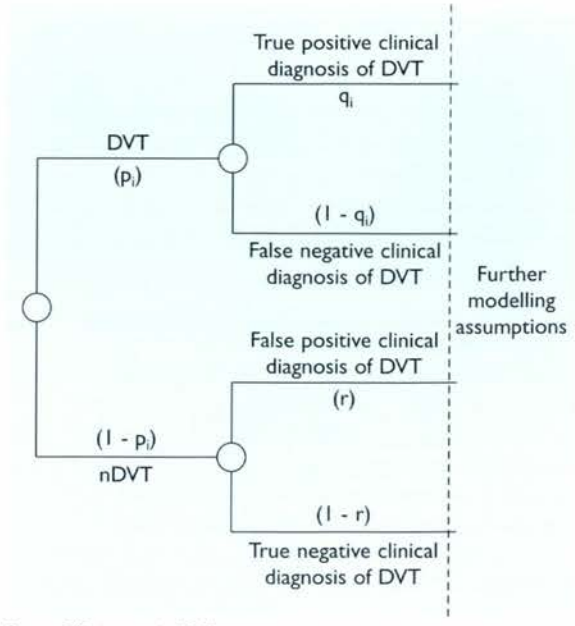
Figure 19 Simple (and incorrect) model linking tPA/streptokinase reperfusion with survival



This model reminds us of the choices one often has in modelling. It might be expected that the mortality rates under the two treatments would be the same conditional on reperfusion success. Presumably those who would have us believe in a model using that as an intermediate endpoint would subscribe to such an assumption. Alternatively, we could (if we had data) model the problem with different conditional probabilities: mortality probability conditional on reperfusion under tPA and conditional on reperfusion under streptokinase treatment (the choice of modelling assumptions was mentioned in the discussion of Box 2). In choosing to emphasize the intermediate outcome analysts were implicitly endorsing the first option (data were not probably available to base a model on the second). Thus, the mortality probabilities conditional on achieving reperfusion (or not) would be the same for each treatment (denoted p and p' in the figure). Recalling that the joint probability (the probability of reaching each terminal node for the decision tree) is the multiplication of the path probabilities along the branches leading to that terminal node, the difference in reperfusion rates will drive the model to the conclusion that tPA is highly superior. While not apparently completely unreasonable assumptions (a priori), results on final outcomes of mortality were significantly less impressive. The equality of the conditional probabilities was simply not true, despite its apparent sense.

Brody (1995) suggests that another choice of primary endpoint (again an intermediate outcome), left ventricular ejection fraction, was acknowledged at the time as 'a powerful predictor of survival after myocardial infarction.' For those patients for whom data were available on this endpoint, while tPA still showed superiority in terms of reperfusion, it did not in left ventricular functioning. Subsequent outcomes of later trials have shown a modest superiority for tPA, though nothing approaching that implied in the model above based on reperfusion. Intermediate endpoints are not all created equal. In addition to an indictment of careless use of intermediate endpoints, this should also serve as a warning to those who would extend them via modelling to final outcomes – the result can be only as good as the constituent assumptions.

Figure 20 Model of DVT prophylaxis in hip surgery



Source: O'Brien, et al, 1994b

this analysis is the importance of diagnostic tests. If we can determine the cost of using the product, then we can also determine the value of the test that would perfectly (or otherwise) discriminate between gram negative bacteremia and not. Box 6 indicates through a simple example the process by which such calculations may be made.

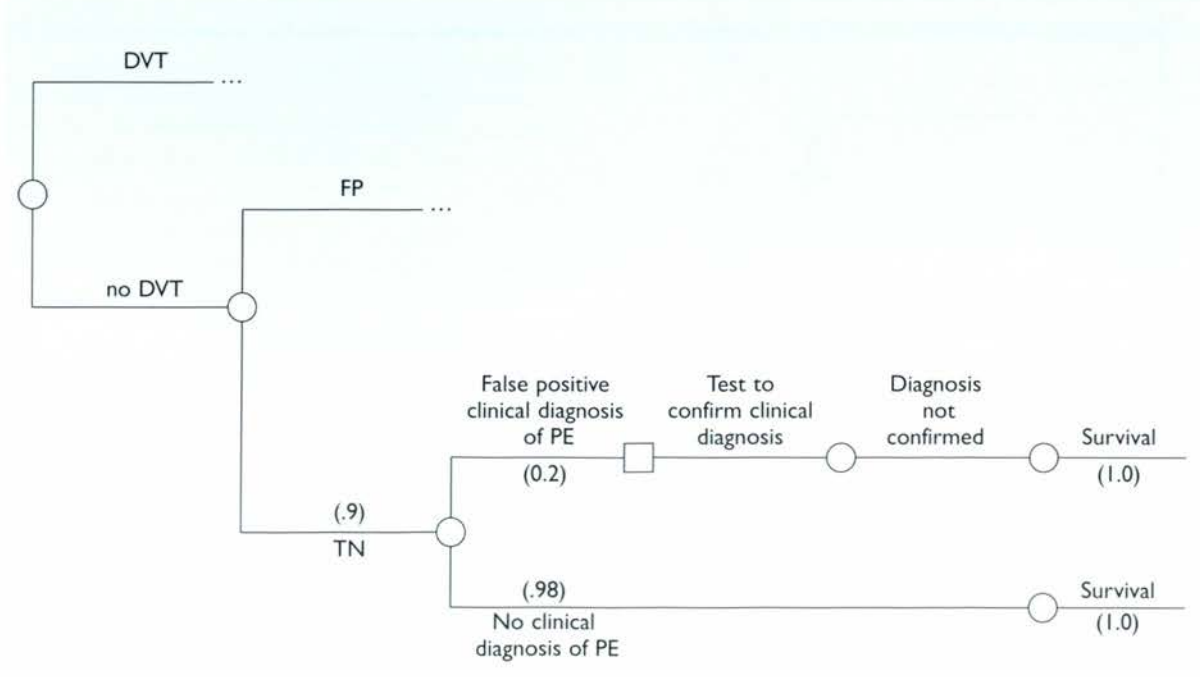
Deep Vein Thrombosis Prophylaxis

O'Brien et al (1994b) have extensively modelled the data in an evaluation of enoxaparin v warfarin in prevention of Deep Vein Thrombosis (DVT) after hip surgery. Trial data were not available on the direct comparison of these two products as head-to-head comparison were done only between enoxaparin and standard heparin. However, the most commonly used anticoagulant in Canada for DVT prophylaxis in hip surgery was warfarin. Therefore this was deemed the relevant comparison for economic purposes (the recently published Canadian guidelines for economic evaluation of pharmaceuticals has condoned exactly such a strategy – CCOHTA, 1994)). Without incorporating modelling techniques, the relevant comparison could not have been made.

This part of the model relied on meta-analysis of existing trials of each of the study drugs against whatever alternatives were used in those trials. Relevant trials were selected according to pre-defined protocol of what was to be admitted into the meta-analysis (see section on meta-analysis in Chapter 6). Thus a 'head-to-head' comparison between enoxaparin and warfarin was modelled from data on their comparisons with other drugs.

In this model the actual DVT rates in trials were used,

Figure 21 Detail of Figure 20 showing additional model assumptions



Source: O'Brien, et al, 1994b

but the resource use was modelled via assumptions on the number of true cases and false cases that would have been identified through traditional observation methods rather than those that were available to the trial. That is, the trial itself implemented nonstandard protocol to identify subclinical cases of DVT. Not all of such cases would go on to develop DVT; not all would have been detected in a normal practice setting; the costs of the detection would have been entirely different. In the model in O'Brien et al they use the DVT rates from the trial but modify them to determine how many of the actual cases (and noncases) would have been identified correctly and incorrectly in routine practice where such diagnostics are not conventionally used. Then subsequent confirmatory testing was assumed to confirm or deny the disease in the true and false positive cases respectively. Resources used in such modelled cases provided the estimate for the cost side of the economics effort; the resource use actually observed in the trial was not used in the analysis.

Figure 20 indicates a section of the model described above. From the meta-analysis was obtained frequency estimates for the probability of DVT for each of the two drugs (p_i, where the subscript refers to either enoxaparin or warfarin). The tree depicted thus stands for two trees, one for each drug, identical except for the probabilities of events. A detected case is assumed in the model to be a real case that, in practice, would either be detected or not clinically (there is no allowance in the model for the subclinical detection to be wrong itself). Thus a real case in practice can be a true positive detection or a false negative detection. The rates of true positive detection in the model were represented in the figure as q_i, again different probabilities for the two different drugs (these, the results of another modelling effort indicating different rates based on the estimated different frequencies of types of DVT likely to result – and the relative abilities to detect them – from each type of prophylactic treatment). In the complete model (not represented here for ease of exposition) true positive detection is assumed to go on for further confirmatory testing in a routine practice setting. This confirmatory testing would have resource use implications, and in the O'Brien model had further error rates built into it.

The other possibility noted in the tree is that the lack of DVT might be correctly detected or that it might be incorrectly interpreted clinically as a real DVT. The latter possibility is denoted by the probability r which is assumed *not* to vary by treatment. In the terms used throughout this document, r is a conditional probability that is independent of treatment, whereas q_i, the rate of true positives is not independent of treatment.

Figure 21, expanding on the bottom branch of Figure 20, shows one aspect of the more detailed model to indicate the type of assumptions used. Even in cases where no DVT is detected and it is assumed that clinical diagnosis would succeed in estimating this truth, it is possible that a subsequent clinical diagnosis of pulmonary embolism would be made (in error, by the assumptions of this model). Such an error would have resource use implications as tests were used to determine whether pulmonary embolism was indeed present. While the detection would be in error, it would still imply that further resources would be used and should appear in a realistic model. Incorporating the likely errors of the routine practice world will reflect the likely resources to be used under the drug treatment regimes. To ignore these errors is to misrepresent the reality of the treatment programs.