Evaluating health care technologies:
Certain methods for uncertain situations

Colofon

ISBN 90-9015098-6

Lay-out, graphics and cover: Remy R.A. Jon Ming, AODWontwerp, Rotterdam
Printed by: Ridderprint, Ridderkerk

EVALUATING HEALTH CARE TECHNOLOGIES

CERTAIN METHODS FOR UNCERTAIN SITUATIONS

Gezondheidszorg-technologieën evalueren:
zekere methoden voor onzekere situaties

Proefschrift
ter verkrijging van de graad van doctor aan de Erasmus Universiteit Rotterdam
op gezag van de Rector Magnificus Prof. dr. ir. J.H. van Bemmel
en volgens besluit van het College voor Promoties

De openbare verdediging zal plaatsvinden op
donderdag 11 oktober 2001 om 13.30 uur

door

Maiwenn Johanna Al
geboren te Vancouver, Canada

Promotiecommissie

Promotoren
Prof. dr. B.A. van Hout
Prof. dr. F.F.H. Rutten

Overige leden
Prof. dr. S.A. van de Geer
Prof. dr. M.G.M. Hunink
Prof. dr. Th. Stijnen

Financial support by Outcomes Research Group, Pfizer Inc.,
Groton, USA and Pharmacia BV, The Netherlands for the publication
of this thesis is gratefully acknowledged

"We have a habit in writing articles published in scientific journals to make the work as finished as possible, to cover up all the tracks, to not worry about the blind alleys or describe how you had the wrong idea first, and so on. So there isn't any place to publish, in a dignified manner, what you actually did in order to get to do the work."

Richard Feynman, Nobel Lecture, 1966

Publications

Chapters 2 to 7 are based on the following articles:

Chapter 2	Costs, effects and C/E-ratios alongside a clinical trial[1] B.A. van Hout, M.J. Al, GS Gordon, F.F.H. Rutten *Health Economics* 1994; 3: 309-319
Chapter 3	Sample size calculation in economic evaluations[1] M.J. Al, B.A. van Hout, B.C. Michel, F.F.H. Rutten *Health Economics* 1998; 7: 327-335
Chapter 4	A Bayesian approach to economic analyses of clinical trials: the case of stenting versus balloon angioplasty[1] M.J. Al, B.A. van Hout *Health Economics* 2000; 9: 599-609
Chapter 5	The cost effectiveness of diclofenac plus misoprostol compared with diclofenac monotherapy in patients with rheumatoid arthritis[2] M.J. Al, B.C. Michel, F.F.H. Rutten *PharmacoEconomics* 1996; 10: 141-151
Chapter 6	Cost-effectiveness of lung transplantation in the Netherlands: a scenario analysis[3] M.J. Al, M.A. Koopmanschap, P.J. van Enckevort, A. Geertsma, W. van der Bij W, W.J. de Boer, E.M. TenVergert *Chest* 1998; 113: 124-130
Chapter 7	Optimal allocation of resources over health care programs: dealing with decreasing marginal utility and uncertainty M.J. Al, T.L. Feenstra , B.A. van Hout submitted

Reprinted with kind permission of:
1. John Wiley & Sons, Ltd;
2. Adis International, Ltd;
3. Chest

Contents

CHAPTER 1

Introduction

In the past decades the demand for health care has increasingly surpassed the supply. This is due, in part, to continuing technological developments in the field of medicine, as well as to other factors such as the increasing number of elderly people in western societies. Consequently, health care budgets are currently under pressure, which raises the question as to whether the health benefits of a new technology can justify its costs. This question is typically addressed through an economic evaluation of the new technology, where it is compared with its best alternative. Unfortunately, such economic evaluation does not always provide a definite answer, partly because the outcome of the evaluation will always be surrounded by some degree of uncertainty.

Several sources of uncertainty can be distinguished (see Manning et al., 1996; Drummond et al., 1997; Briggs & Sculpher, 1995). Manning and colleagues differentiate between two main types of uncertainty: parameter uncertainty and modelling uncertainty.

Parameter uncertainty is the uncertainty about the true value of the parameters used as input and can arise in several ways. For instance, the value of a parameter may be estimated from a study sample or may be estimated on the basis of opinion, e.g. from an expert panel. Furthermore, parameter uncertainty may be introduced when fairly certain outcomes are generalised to a larger population or another setting, or extrapolated beyond the observation period or from an intermediate to a final endpoint (e.g. from bone mineral density to quality-adjusted life years). Parameter uncertainty also occurs when there is disagreement about the appropriate value, as might be the case with the discount rate or when a choice is made for the valuation of quality of life through either time trade-off or standard gamble.

Modelling uncertainty arises when the analyst is uncertain about the mathematical equations of the model, regarding both the inclusion of certain variables and the functional form by which parameters should be combined. For example, does coronary heart disease multiply the age-sex-specific mortality rates by a constant amount or does it add a constant amount to the mortality rate. Such choices are often made in the absence of clear evidence that one or other of the functions is the appropriate one to use.

In summary, there are many sources of uncertainty and it is unrealistic to assume that all uncertainty can be eliminated. Instead, methods are developed to deal with uncertainty in the best possible way. A first step in dealing with uncertainty is to establish how uncertain we are; for example, a well-known way of describing uncertainty is through use of a 95% confidence interval. Next, methods are sought to decrease uncertainty as much as is feasible. This may be achieved, for example, by specifying before a study is conducted which level of uncertainty is acceptable and then calculating the corresponding sample size. Or, after the results of a study become available, by combining these results with results of previous studies, for instance through meta-analysis. Finally, we should strive to help decision makers to deal with uncertainty in the decision-making process.

Overview of developments

In the early days of cost-effectiveness analysis in health care, most studies used a modelling approach, since data on costs and effects at the patient level were scarce. Decision trees, a frequently used type of model, had been used for years in operations research and medical decision making. Decision trees allow a systematic assessment of all possible outcomes of a process given a set of decisions on, for example, interventions and diagnostic procedures. A sensitivity analysis has always been part of such analysis: probabilities in the model are varied to see if and how the decision changes. It seems reasonable to assume that sensitivity analysis would also be an integrated part of modelling studies in economic evaluations of medical technologies. However, this proved to be less obvious than expected. In 1995, Briggs & Sculpher reviewed economic evaluations published in 1992. A total of 93 studies (modelling, prospective and retrospective) were examined. Of these, 24% failed to consider uncertainty at all and 38% of the studies had attempted a sensitivity analysis that was judged too limited to provide an adequate representation of uncertainty. Of the remaining studies, 14% was judged good and 25% adequate in representing uncertainty.

Agro et al. (1997) published a similar review; they studied 90 articles published between January 1989 and December 1993, equally divided among health economic journals, medical journals and pharmacy journals. They concluded that only 59% of the studies used sensitivity analysis. However, there was a large difference between the types of journals: 80% of the articles in the health economics journals and 77% of the articles in the medical journals had conducted sensitivity analysis, compared with only 20% of the articles in the pharmacy journals.

In those studies that did include a sensitivity analysis, these analyses were mainly one-way, multi-way or extreme case sensitivity analyses. In a one-way or multi-way analysis one or more (usually not more than three) parameters are varied between certain limits. However, no attention is given to the fact that the limits are often less likely to occur than the baseline estimate. Consequently, in the last few years some studies also include a probabilistic sensitivity analysis, in which all input variables are varied simultaneously, according to probability distributions (e.g. chapter 5; Michel et al., 1996; Doubilet et al., 1985). Such an analysis presents information on all possible outcomes, as well as on the likelihood of these outcomes.

Situations may occur where parameters themselves are not associated with uncertainty because the parameter is a decision parameter. However, when no final decision has yet been made, it may be unclear how these parameters influence the outcome. For instance, in a cervix cancer screening program, a decision has to be made about the screening frequency. This could be once every 3 years or once every 5 years and it may not be clear how this choice influences the cost-effectiveness ratio. In such a situation a scenario analysis can be used to calculate the outcome for several options of the decision parameter. The scenario analysis can also be used to see what happens to the outcome when decision makers influence certain parameters, i.e. when there are external influences on a parameter. This is, for instance, the case for transplantation programs: while it is possible to estimate the number of donor organs available from

observations, this number can be influenced by e.g. new laws or protocols.

In recent years, data on costs and effectiveness at the patient level are often collected alongside clinical trials. The analysis of such data requires statistical methods describing the uncertainty due to sampling. For a large part, the same methods used for the analysis of clinical data were used for the economic and outcomes data. The analysis of cost data, however, raised some imported issues, due to the skewness of the cost distribution. This skewness is easily explained by the fact that, for many diseases and their treatments, most patients will incur small to moderate costs, while a small number of patients will incur very high costs due to medical complications or extended hospital stay. In biostatistics, skewed data is usually log transformed and all statistical inference is then performed on this transformed data. Testing the mean of log transformed data is equivalent to testing the geometric mean on the original (cost) scale. For skewed data, the geometric mean costs will always be smaller than the average costs. However, for economic evaluations the measure of interest is the average costs per patient, as the purpose of the analysis is to inform decision makers on the budget needed to treat patients, which can be obtained by multiplying average costs by the total number of patients. Thus, inference based on the geometric mean will provide irrelevant conclusions. Clearly, the analysis of skewed cost data is not straightforward and consequently this subject has recently received much attention (see e.g. Briggs & Gray, 1998a; Thompson & Barber, 2000; O'Hagan & Stevens, 2001b).

Another feature of economic evaluations that attracted attention is the incremental cost-effectiveness ratio (ICER), as a ratio of two stochastic variables. The ratio of two normally distributed variables will rarely have a well-behaved distribution. The distribution may be skewed, and may also have heavy tails, which implies that, theoretically, mean and variance do not exist. In 1994, O'Brien et al. proposed an analytical method for the calculation of a confidence interval around the ICER. This report was the first of numerous papers on this issue, one of which is presented in chapter 2. Both analytical and bootstrap (i.e. simulation) methods have been proposed and compared (see e.g. O'Brien et al., 1994; van Hout et al., 1994; Wakker & Klaassen, 1995; Sacristan et al., 1995; Chaudhary & Stearns, 1996; Willan & O'Brien, 1996; Briggs et al., 1997; Polsky et al., 1997).

When confidence intervals for the ICERs are calculated, it becomes clear that these intervals are usually very wide, as a direct result of the fact that the sample size of clinical trials is usually based on the primary clinical endpoint of the study. As the variation in costs is often much larger than in the primary (clinical) endpoint, cost-effectiveness studies usually have little power to detect whether a new intervention is cost-effective compared to the alternative. In cost-effectiveness studies, we study the ratio of two variables, which may be highly correlated. This means that the methods for sample size calculation used for the primary endpoint cannot easily be translated to methods for the ICER. Consequently, specific methods for sample size calculation in economic evaluation have been developed. At first a simulation method was proposed (see chapter 3) and recently several papers have described analytical methods (see e.g. Briggs & Gray, 1998b; Laska et al., 1999; Willan & O'Brien, 1999).

Theoretically, it would be possible to include economic parameters in a sample size

calculation for randomised clinical trials. However, in studies where the primary goal is to show improved efficacy or effectiveness, with proof of cost-effectiveness as secondary goal, the question arises whether it would be ethical to include economic parameters in the sample size calculation. Surely an important purpose of sample size calculation is to assure that enough patients enter a randomised clinical trial to allow for statistically significant results while ensuring that not too many patients are exposed to the risks that usually accompany such a trial. When the primary goal of a study is to prove cost-effectiveness, while superior efficacy has already been demonstrated, it may be considered unethical to assign patients to a control group with the lower efficacy. Besides ethical issues, practical reasons may preclude basing sample size on economic parameters; for example, the period of enrolment may become too long, or the costs of the clinical trial might become too high.

Clearly it is unreasonable to expect that the sample size of future studies should be increased in order to reduce uncertainty around the ICER. Therefore it is important to consider other available options. One important option is to use Bayesian statistics. In 1763 Bayes' paper "Essay Towards Solving a Problem in the Doctrine of Chances" was published posthumously and this paper became the basis of what is now called Bayesian statistics. In Bayesian statistics, a prior probability distribution for the parameter of interest is specified first. Sample information is then obtained and combined using Bayes' theorem, resulting in a posterior probability distribution for the parameter. This posterior distribution can then be used for statistical inferences about that parameter. The key, and somewhat controversial, feature of Bayesian methods is the notion of a probability distribution for a population parameter. According to frequentist statistics, parameters are constants and cannot be represented as random variables. Bayesian proponents argue that, if a parameter value is unknown, it is logical to specify a probability distribution that describes the possible values for the parameter as well as their likelihood. In general, we can state that Bayesians deal with the probabilities of hypotheses, given a data set, whereas frequentists deal with the probability of data sets, given a hypothesis.

The use of Bayesian methods in health care has been proposed in an attempt to make full use of other information already available outside the clinical trial (see for instance Eddy et al., 1990; Jones, 1996). Since the Bayesian approach deals with uncertainty by treating data collection as an ongoing process whereby current beliefs are modified in the light of new information, the uncertainty about the parameter values is reduced but is never completely removed. One of the main challenges of using a Bayesian approach in health economics is describing these current beliefs through a prior distribution (see also chapter 4).

Besides being able to reduce uncertainty by combining sources of information, the use of Bayesian methods may offer other benefits (Briggs, 1999; Heitjan et al., 1999). One such benefit is that Bayesian inferences are easier to understand in the sense that, for instance, Bayesian confidence intervals have a far more intuitive interpretation than classical confidence intervals. A frequentist 95% confidence interval tells us that if the trial was repeated often enough and if a confidence interval was calculated each time, 95% of those intervals would contain the true difference. Most people, however,

are inclined to interpret the interval as having a 95% chance that the true difference lies within the interval. Within the frequentist framework this interpretation is not allowed, whereas it is within the Bayesian framework. Another benefit is that some problems that cannot be easily handled with classical methods are rather straightforward using Bayesian analysis. A well-known example is the testing for equality of means with unequal variances (see Box & Tiao, 1973).

The recent interest in the use of Bayesian analysis in economic evaluations has led to the "Bayesian Initiative in Health Economics & Outcomes Research". This initiative was established in 1997 by MEDTAP International to explore the value of the Bayesian approach to health economics and outcomes research. Their premise is that the unique decision-oriented perspective of the Bayesian approach will produce information that is more interesting and relevant to decision makers (see also *www.bayesian-initiative.com*). This idea is, for instance, also supported by Claxton, who describes a Bayesian approach to the value of information and its implication for the regulation of new pharmaceuticals (Claxton, 1999a) and optimal trial design (Claxton, 1999b).

Thus, there is a growing trend in health economic literature towards discussions on how the outcomes of health economic evaluations are used and may be used in aiding decision making. Results of these evaluations may be used for decisions on reimbursement for one specific technology, or for the allocation of a budget over a large number of technologies. Furthermore, outcomes may be used to decide on further data collection. Some methodological issues regarding the possible role of uncertainty in the decision-making process are discussed in this thesis.

Outline of this thesis

Chapter 2 presents the 'C/E acceptability curve' as a method for statistical inference regarding the ICER. This method can be used in economic evaluations that are performed alongside randomised clinical trials, in which economic data, such as safety and efficacy data, are collected at the patient level. The acceptability curve measures the probability that the ICER is acceptable in comparison to a predefined limit R (which reflects the critical cost-effectiveness ratio society has decided on) for all positive values of R. The problems and proposed methodology are illustrated by using, as an example, data from a phase II trial in which the costs and effects of anakinra (a potential therapeutic agent for the treatment of sepsis syndrome) were investigated.

In chapter 3 uncertainty in the design phase of a cost-effectiveness study is addressed. A simulation method is presented for sample size calculation in such studies and an assessment made as to how sensitive the estimated sample size is to changes in assumptions about the expected costs, effectiveness and the correlation between them. Data from two trials are used to illustrate this; one comparing primary coronary angioplasty and streptokinase in the treatment of acute myocardial infarction and the other comparing the efficacy and safety of lansoprazole with omeprazole in the treatment of reflux oesophagitis.

Chapter 4 summarises some first experiences with a Bayesian economic analysis based on data from a randomised clinical trial using informative priors. First, a

univariate approach is presented for the analysis of the differences in costs and effects. Second, some initial ideas about a multivariate approach for the analysis of the ICER are presented. Subsequently, the methodology is applied using data from two trials. For both trials, the balance was assessed between the costs and effects of stent implantation compared with balloon angioplasty in patients with angina. Several prior distributions are used for the analysis of the second trial to see how the conclusions differ after combining data with these priors.

In chapter 5 an example is given of a modelling study that uses sensitivity analysis to address uncertainty. The objective of this study was to compare the costs and effects of a fixed dose combination diclofenac/misoprostol and diclofenac alone in the treatment of patients with rheumatoid arthritis. The sensitivity of the results for variation in the input variables was studied both univariately and multivariately (i.e. probabilistic). The univariate analysis shows which variables markedly influence the outcome, whereas the multivariate analysis yields the most probable outcome, the 'worst' and the 'best' possible outcome and also the distribution of possible outcomes.

Chapter 6 presents an example of a scenario analysis, as it was performed in the economic evaluation of the Dutch lung transplantation program. The scenario analysis describes future transplantation programs, operational for 15 years. This simulation period is necessary to reach a stable number of patients on the waiting list, which has not yet been established during the observation period. Societal costs, survival and quality of life are followed for up to 40 years, comparing the situation with and without a lung transplantation program. Besides a baseline scenario (i.e. a prolongation of the current program), two alternative scenarios will be presented: a policy scenario, restricting the inflow of patients on the waiting list and a donor scenario assuming a larger supply of donor lungs. These scenarios address the uncertainty arising from possible external influences on a parameter.

In chapter 7, models for budget allocation are explored. In most textbooks and articles on the foundation of cost-effectiveness analysis, one of the basic premises is that the goal is to maximise health gains given a fixed budget. However, this optimisation problem does not allow for decreasing marginal value and uncertainty. An alternative model is proposed with decision makers maximizing a value function with decreasing marginal returns over health effects and the budget available for purposes competing with health care spendings. The question is raised whether different models lead to differences in the decision as to whether a new program should be implemented, and in the effect of the introduction of a new program on current programs. Next, attention is given to the fact that, in practice, the estimates of costs and effects will be surrounded by uncertainty. Clearly, if costs are uncertain, it is impossible to stay within the fixed budget with certainty. In this chapter an optimisation problem is formulated that allows for uncertainty, based on expected utility. The question may now be asked whether accounting for uncertainty has any influence on the decision which programs should be funded.

Finally, chapter 8 presents the conclusions of the previous chapters and discusses recent developments and areas for future research.

CHAPTER 2

Costs, effects and c/e-ratios alongside a clinical trial

Summary

A general approach is discussed to assess the uncertainty surrounding the incremental cost-effectiveness ratio (ICER) estimated on the basis of data from a randomised clinical trial. The approach includes the calculation of a 95% probability ellipse and introduces the concept of a so called C/E-acceptability curve. This latter curve defines for each predefined ICER the probability that the ICER found in the study is acceptable. The approach is illustrated by estimates of costs per life saved and costs per patient discharged alive on the basis of data from a phase II trial addressing the value of anakinra in treating sepsis syndrome.

Introduction

Before new medical procedures are introduced, they are traditionally assessed in terms of safety and efficacy. Such assessments are obligatory in the case of the registration of pharmaceutical products and there are strict rules on how such assessments should take place.

While safety and efficacy may be the primary parameters for the registration of pharmaceutical products, it is increasingly recognised that additional parameters, such as costs, need to be considered when decisions have to be made about reimbursement. Since policy makers have encountered large increases in health expenditures because of the introduction of new medical procedures, there is a growing interest in information on the additional costs and benefits of these procedures. This need for information has been formalised in the guidelines concerning pharmaceutical products in Australia (see Commonwealth of Australia, 1992), in the proposed guidelines in Canada (see Schubert, 1993), and in the requirement for information about costs and benefits when financing new health care programmes in The Netherlands (see Rutten & Bonsel, 1992).

Among the techniques to assess safety and efficacy, the randomised clinical trial (RCT) is, despite its limitations, respected as the golden standard. The results of an RCT may not always reflect common practice and the comparator may not always be the appropriate one. However, the experimental design minimises potential biases in the measurement of differences (e.g. Pocock, 1983). Additionally, the RCT offers the possibility to test, by use of statistical procedures, whether the observed differences are due to coincidence or not. P-values and confidence limits are common concepts within clinical trials.

For similar reasons as with respect to safety and efficacy one may state that economic data should also be gathered in an RCT. This could permit the use of a statistical paradigm similar to that used for safety and efficacy. However, while there are a number of well developed statistical procedures, for example concerning survival, such procedures are not available for the principal outcome of an economic evaluation: the incremental cost-effectiveness ratio (ICER).

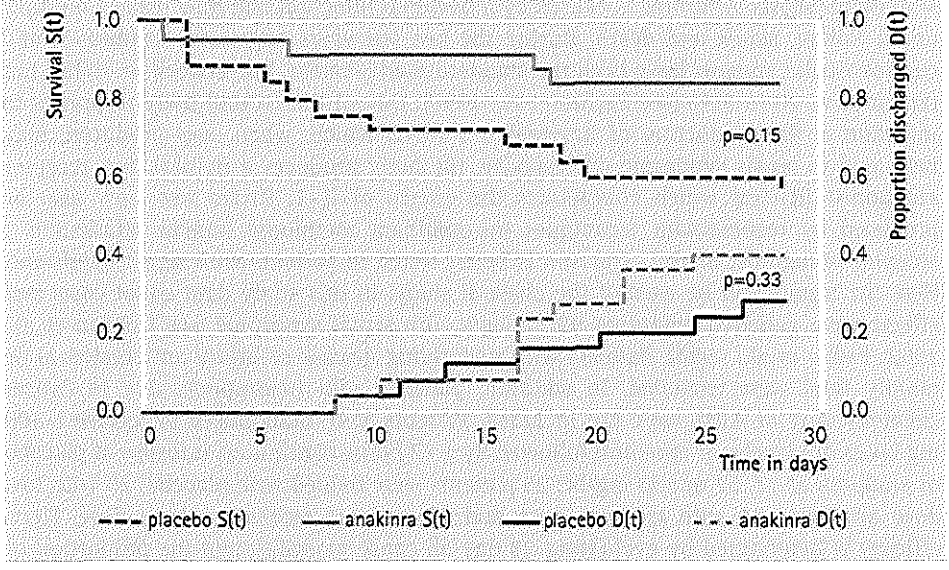
This chapter sets out to solve some of the problems in the statistical assessment of a ICER. The problems and methodology are illustrated using, as an example, data from a phase II trial in which the costs and effects of anakinra, a potential therapeutic agent for the treatment of sepsis syndrome, were investigated. We look at the possibilities for statistical inference regarding the costs and effects calculating a 95% probability ellipse and we introduce a new concept based on some of the ideas of O'Brien et al. (1994): the so-called 'C/E-acceptability curve'. This curve measures the probability that the ICER -resulting from a trial- is acceptable in comparison to a predefined ICER R for all positive values of R. The basic idea behind this concept is generally applicable to a large number of situations and its use is illustrated using the data from the anakinra-trial.

Anakinra versus placebo

Anakinra is a therapeutic agent developed for the treatment of patients with sepsis syndrome. Sepsis syndrome is caused by a bacterial infection and may lead to multiple organ failure, shock and death. During the last ten years aggressive antibiotics and cardiovascular support have been the standard management, but 4-week mortality has remained at 20 to 60% (see Parrilo et al., 1990). One of the natural body responses to infection is the production of interleukin 1 (IL-1) which, when over-produced, may help lead to sepsis syndrome. In response, the body produces a natural IL-1 receptor antagonist (IL-1ra) which counteracts the effects of IL-1. IL-1ra can now also be produced by recombinant techniques (recombinant methionyl human interleukin 1 receptor antagonist or rhIL-1ra) and is being tested in patients with sepsis syndrome. Here the results will be used of the phase II study. In this study data about costs and effects were gathered regarding 99 patients with sepsis syndrome (see Gordon et al., 1992; Fisher et al., 1994). Patients with underlying disease or injury anticipated to be rapidly fatal were excluded. In addition, patients with neutropenia due to cancer chemotherapy or AIDS were not included. Patients were randomised to receive a 100 mg bolus of placebo or anakinra followed by a 72 hours continuous infusion of placebo or anakinra at 17, 67 or 133 mg/hr: 25 patients were treated with placebo, 25 with low dose anakinra, 24 with medium dose and 25 patients received a high dose. Here, efficacy is determined by evaluating 28 day all cause mortality and the proportion of patients discharged alive. Additionally, costs are analysed by using length of stay data during the 28 day study period distinguishing between normal and intensive care.

Sepsis is an acute illness which usually results in either a cure or death within 5-10 days. Hence, in order to capture the effects on survival clinical trials such as those for HA-1A, another experimental agent (see Ziegler et al., 1991), and anakinra have limited their follow up period to 28 days. It is clear that 28 days is not enough for an estimate of costs per life year gained. For that purpose one needs long term data. When this is lacking, one tends to make assumptions, gather information from other sources and use epidemiological knowledge (e.g. Schulman et al., 1991), but this may produce unreliable estimates. Here, we limit ourselves to some short term parameters for effectiveness and we will concentrate on costs per additional survivor and costs per additional patient discharged alive. Additionally, as we are only using the data for illustrative purposes, we limit ourselves to the comparison between the high dose anakinra group and the placebo group. Figure 2.1 presents product-limit estimates of the duration distributions until death and discharge.

fig 2.1 Survival and proportion discharged after 28 days placebo versus high dose anakinra



The figure clearly indicates the process leading to the results after 28 days of using high dose anakinra: a trend towards an increase in survival, an increase in the discharge-rate and an increasing proportion of patients hospitalised. The P-values indicated in figure 2.1 result from tests on differences in durations (Wilcoxon and log-rank) under the obvious assumption that patients who have died cannot be discharged alive.

To obtain an estimate of costs, we distinguish between intensive care and non-intensive care. Using an estimate of NLG 2,500 for an intensive care day and NLG 1,000 for a non-intensive care day (see Hilgeman et al., 1992), total costs per patient are estimated at NLG 33,720 for the placebo group and at NLG 35,100 per patient in the high-dose anakinra group. The incremental cost-effectiveness or the additional costs per additional life saved (hereafter called 'costs per life saved') can be estimated at NLG 4,928 and the additional costs per additional patient discharged alive (hereafter called 'costs per patient discharged') at NLG 11,500. Both figures do not seem to be particularly high in comparison to the average costs per patient but it is emphasised that the costs of anakinra are not included since the price has not yet been fixed. Having only 25 observations the question remains how reliable these results are. In the next sections we will present a methodology to assess this in a formal way.

Costs, effects and c/e-ratios

The methodology presented in this section has primarily been developed in relation to the results for the anakinra trial in which effects are measured in terms of costs per life saved and costs per patient discharged. The results however are applicable to all types of ICERs such as 'costs per life-year gained', 'costs per percentage change in ejection fraction', 'costs per percentage change in cholesterol level' etc. Therefore, we will use a general notation and present the basic formula for the ICER as:

$$ICER = \frac{C_T - C_P}{E_T - E_P} = \frac{\Delta C}{\Delta E} \qquad \textbf{Eq.1}$$

in which:

C_T: average costs per patient in the treatment group
C_P: average costs per patient in the placebo group
E_T: average effects per patient in the treatment group
E_P: average effects per patient in the placebo group

Following the ideas of O'Brien et al. (1994), costs and effects per patient can be considered as the outcomes from a random distribution. Although costs and effects may not be distributed normally, with increasing numbers, the averages do approach a normal distribution. So,

$$C_T \sim N(\mu_{C_T}, \sigma^2_{C_T})$$
$$C_P \sim N(\mu_{C_P}, \sigma^2_{C_P})$$
$$E_T \sim N(\mu_{E_T}, \sigma^2_{E_T})$$
$$E_P \sim N(\mu_{E_P}, \sigma^2_{E_P})$$

In all equations μ represents the mean and σ^2 the variance of a normal distribution. Additionally E refers to average effects, C to average costs, T to treatment and P to placebo.

As a consequence, the denominator and numerator in equation 1, measuring differences between normally distributed random variables, approach normal distributions too. So,

$$\Delta C \sim N(\mu_{\Delta C}, \sigma^2_{\Delta C})$$
$$\Delta E \sim N(\mu_{\Delta E}, \sigma^2_{\Delta E})$$

in which:

$$\mu_{\Delta C} = \mu_{C_T} - \mu_{C_P}$$
$$\mu_{\Delta E} = \mu_{E_T} - \mu_{E_P}$$
$$\sigma^2_{\Delta C} = \sigma^2_{C_T} + \sigma^2_{C_P}$$
$$\sigma^2_{\Delta E} = \sigma^2_{E_T} + \sigma^2_{E_P}$$

Costs may be correlated with effects and consequently, both variables are not independent and the combination of both can be seen as the result from a bivariate normal distribution $f(\Delta E,\Delta C)$ including five parameters. The joint probability density function can be expressed as:

$$f(\Delta E,\Delta C) = \frac{1}{2\pi\sigma_{\Delta C}\sigma_{\Delta E}\sqrt{(1-\rho^2)}} e^{-Q/2}$$

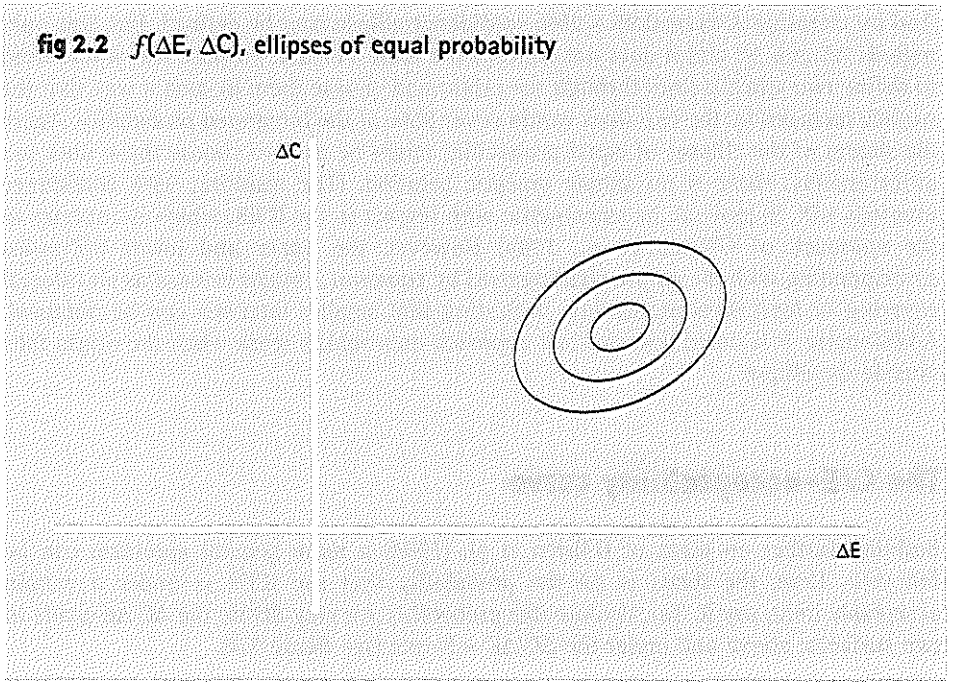
where:

$$Q = \frac{1}{1-\rho^2}\left[\left(\frac{\Delta C - \mu_{\Delta C}}{\sigma_{\Delta C}}\right)^2 - 2\rho\left(\frac{\Delta C - \mu_{\Delta C}}{\sigma_{\Delta C}}\right)\left(\frac{\Delta E - \mu_{\Delta E}}{\sigma_{\Delta E}}\right) + \left(\frac{\Delta E - \mu_{\Delta E}}{\sigma_{\Delta E}}\right)^2\right]$$

Here, ρ is the correlation coefficient of ΔC and ΔE taking values between -1 and 1. In the appendix we show how estimates of all parameters can be obtained when observations are available.

The joint probability distribution of ΔC and ΔE can be graphically represented in a so called $\Delta C/\Delta E$-plane (figure 2.2). The lines in figure 2.2 represent for three selected values of $f(\Delta E,\Delta C)$ the corresponding values of ΔC and ΔE. These are 'ellipses of equal probability', indicating the shape of the joint probability density function of ΔC and ΔE.

fig 2.2 $f(\Delta E, \Delta C)$, ellipses of equal probability



However, this is not what we are interested in, we are interested in the distribution of the ICER. Formally, the ratio of two normal distributed variables has neither a finite mean nor a finite variance. This is a well known problem to statisticians and publications about this go back to 1928 (e.g. Merril, 1928; Geary, 1930; Fieller, 1932). One of the consequences is that using a Taylor approximation to calculate 95% confidence limits (as proposed by O'Brien et al., 1994) is formally incorrect. Therefore, another approach has to be followed. One may propose the calculation of a surface that would cover 95% of the integrated probability. A possible choice for the shape of such a surface, leading to a relatively simple analytic solution, is the ellipse of equal probability. Such an ellipse, containing an integrated probability equal to λ ($0 < \lambda < 1$), is characterised by (see e.g. Gnedenko, 1968) :

$$\frac{1}{1-\rho^2}\left[\left(\frac{\Delta C-\mu_{\Delta C}}{\sigma_{\Delta C}}\right)^2-2\rho\left(\frac{\Delta C-\mu_{\Delta C}}{\sigma_{\Delta C}}\right)\left(\frac{\Delta E-\mu_{\Delta E}}{\sigma_{\Delta E}}\right)+\left(\frac{\Delta E-\mu_{\Delta E}}{\sigma_{\Delta E}}\right)^2\right]=-2\ln(1-\lambda)$$

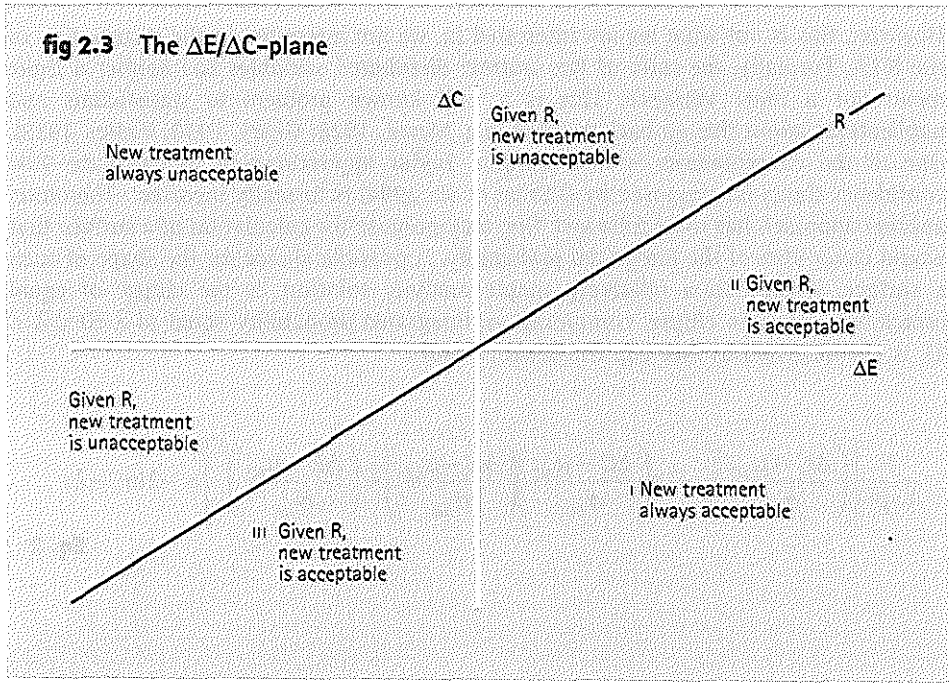
Eq. 2

And by choosing λ to be 0.95 a 95% confidence ellipse may be defined. But, this only regards ΔC and ΔE and it does not concern the ICER. For that purpose one would need to define two straight lines through the origin and choose them in such a way that the surface included by the two lines would cover 95% of the integrated probability. Again the shapes of the ellipses of equal probability might be used to choose those lines but unfortunately, there is no simple analytic solution. (The resulting 95% confidence intervals will, in general, be asymmetric with respect to the ICER, which is contrary to the approach of O'Brien et al., 1994) However, the upper and lower C/E-ratios corresponding with the ellipse characterised by equation 2 might be used as an interval covering *at least* 95% probability, indicating that smaller intervals can be obtained. Additionally, it is noted that this last approach only makes sense when the origin falls outside the ellipse.

The C/E-acceptability curve

Building further on ideas of O'Brien et al. (1994) a more general approach can be followed. Here, the focus is on the probability that the ICER is under a certain acceptable limit, say R. To calculate this probability we may divide the $\Delta C/\Delta E$-plane in two surfaces: above and under the $\Delta C/\Delta E$ = R-line (see figure 2.3).

fig 2.3 The $\Delta E/\Delta C$-plane



The origin of the ΔC/ΔE-plane corresponds with the situation where the new therapy is equivalent to the old therapy. At all other points the new therapy differs either in terms of costs or in terms of effects. Above the ΔC/ΔE=R-line the preference is to reject the new treatment (see also Black, 1990). Under this line the preference is to accept the new treatment. Both surfaces can be broken down into three components. Here we will concentrate on the last surface, indicated as the 'acceptability-surface'. The three components are:

I	$\Delta C < 0, \Delta E > 0$
II	$\Delta C > 0, \Delta E > 0, \Delta C/\Delta E < R$
III	$\Delta C < 0, \Delta E < 0, \Delta C/\Delta E > R$

The first component of the acceptability surface corresponds with the situation in which the new therapy is more effective and is saving costs. The second component corresponds with the situation that the new therapy is more effective but also more expensive. The last component corresponds with the situation that the new therapy is less effective but is saving costs. (This last component can be seen as the mirror of the second by imagining that the standard therapy is newly introduced and compared with the new therapy. The standard therapy is more effective but also more expensive and again, the ΔC/ΔE = R-line marks the acceptability from a cost-effectiveness point of view.)

Increasing the value of R corresponds with turning the ΔC/ΔE = R-line anti-clockwise. Consequently, while increasing the value of R (at positive values of R) does not affect the size of the first component of the acceptability surface, it increases the second component and it decreases the third component.

Having divided the ΔC/ΔE-plane into two surfaces, above and under the ΔC/ΔE =R-line, we may now calculate the probability that the ICER resulting from a trial is under R. This probability equals the integrated probability density of f (ΔE, ΔC) under the ΔC/ΔE = R-line. Additionally, when R is defined as the limit for which the ICER is acceptable, we may define this probability as the probability that the ICER that is found in the study is acceptable. Consequently, this probability changes when R changes and the exact relation between the two can be put forward by defining a so called 'C/E-acceptability curve'. In algebraic terms:

$$F(R)=\int_{-\infty}^{\infty}\int_{-\infty}^{R\Delta E} f_{\Delta E,\Delta C}(\Delta E,\Delta C)\, d\Delta C\, d\Delta E$$

This curve can be calculated for all non-negative values of R. When R = 0, it measures the probability that the new therapy will save costs (the integrated density under the ΔE-axis). When R is infinite, it measures the probability that the new therapy will be effective (the integrated density at the right side of the ΔC-axis). How such a curve can be used is illustrated in the next section where it is estimated on the basis of the data from the anakinra trial.

The C/E-acceptability curve of high dose anakinra versus placebo

To be able to illustrate the approach introduced in the former section we need patient specific data about costs and effects from a random sample. In the case of anakinra we define average costs per patient per day which means that we disregard the fact that the costs per inpatient day may differ per day and per patient. Additionally, effects are defined as the change in the proportion survived. It is emphasised that this limits the range of potential effects to the interval [-1, 1] but it can be shown that the distribution of the average effects will still approximate a normal distribution.

table 2.1 Parameter-estimates of the bivariate normal distribution $f(\Delta E,\Delta C)$

	ΔE=Δ (proportion survived)	ΔE=Δ (proportion discharged)
$\mu_{\Delta C}$	1380	1380
$\mu_{\Delta E}$	0.28	0.12
$\sigma^2_{\Delta C}$	3.2 x 107	3.2 x 107
$\sigma^2_{\Delta E}$	0.0152	0.0177
ρ	0.34	-0.31

Table 2.1 shows the estimates of the bivariate normal distributions regarding the proportion survived and the proportion discharged. Figure 2.4 and 2.5 present the corresponding probability contours. Here, the outer contours are the 95% confidence ellipses.

fig 2.4 $f(\Delta E, \Delta C)$, anakina versus placebo, $\Delta E = \Delta$(proportion survived)

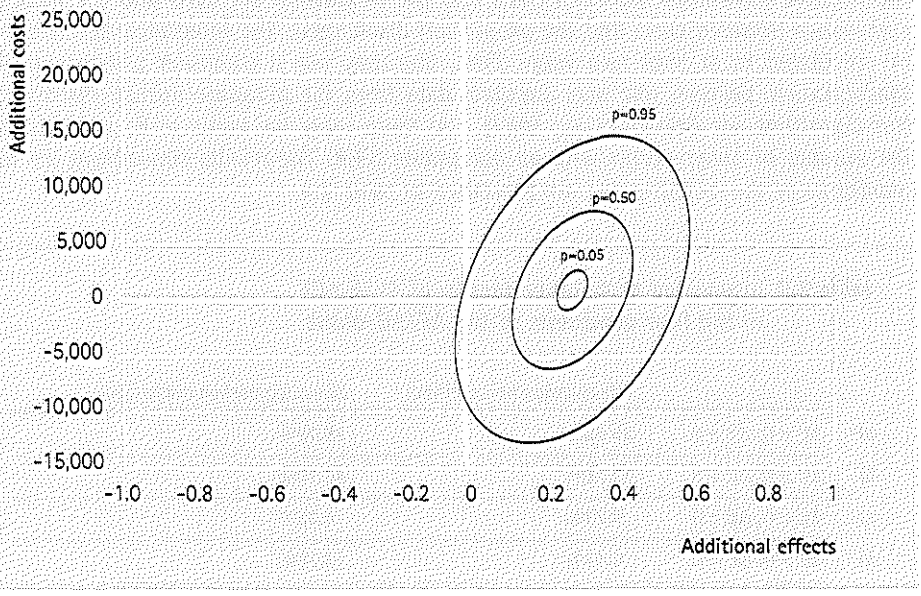
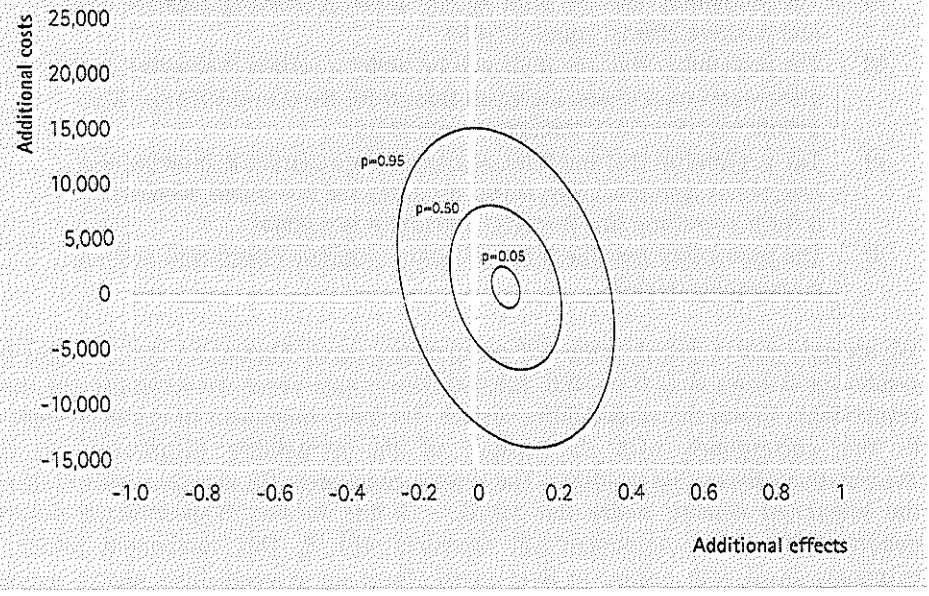


fig 2.5 $f(\Delta E, \Delta C)$, anakina versus placebo, $\Delta E = \Delta$(proportion discharged)



If the covariance would have been zero, the figures would have formed horizontal ellipsoids. Here, we recognise a positive and a negative shape in the ellipsoids reflecting the positive covariance between costs and the proportion survived and the negative covariance between costs and the proportion discharged. The integrated densities can be calculated in at least two ways. First, one may use a computer programme like Mathematica. Second, one may simulate data from the bivariate normal distribution (for algorithm, see Bratley et al., 1987), sort them, vary the value of R, and calculate the number of observations with an acceptable ICER. The results concerning the four quadrants of the $\Delta C/\Delta E$-plane are presented in table 2.2.

table 2.2 Integrated probability density of $f(\Delta E,\Delta C)$ in the four quadrants of the $\Delta C/\Delta E$-plane

		$\Delta E=\Delta$ (proportion survived)	$\Delta E=\Delta$ (proportion discharged)
a)	$\Delta C < 0, \Delta E > 0$	39.69%	35.79%
b)	$\Delta C > 0, \Delta E > 0$	59.16%	46.13%
c)	$\Delta C < 0, \Delta E < 0$	0.73%	4.33%
d)	$\Delta C > 0, \Delta E < 0$	0.42%	13.76%

Concerning the costs per life saved we may conclude that:

- **a)** The probability that the ICER will be under 0, showing negative costs and positive effects, equals 0.3969.
- **b)** The probability that the ICER will be under a certain R showing more effects and more costs, increases from 0 for R= 0 to 0.5916 for R is infinite (for R = NLG 100,000, it is 0.5865).
- **c)** The probability that the ICER will be under a certain R showing less effects but also less costs, starts at 0.0073 for R = 0 and decreases to 0 for R is infinite (for R = NLG 100,000 it is 0.0044).
- **d)** The probability that the ICER will show negative effects and positive costs equals 0.0042.

We are now able to calculate the C/E-acceptability curve. Figures 2.6 and 2.7 show the results for R from NLG 0 to NLG 100,000. Under this curve we recognise the three components within the acceptability surface. The first component (I) is independent of R; this corresponds with the surface of the probability distribution in the ΔC-negative/ΔE-positive quadrant. The second component (II) increases with R; this is the integrated density under the $\Delta C/\Delta E$ = R-line in the ΔC-positive/ΔE-positive quadrant. The third component (III) decreases with R. This component (which is extremely small with respect to lives saved where the relevant lines almost overlap

each other; figure 2.6) corresponds with the integrated probability under the R-line in the ΔC-negative/ΔE-negative quadrant.

Having noticed the problems in calculating confidence intervals surrounding a ICER, other approaches have to be followed. One of the possibilities will be the calculation of a C/E-acceptability curve as presented here for the case of high dose anakinra in comparison to placebo. In relation to the small number of observations it may be assumed that the shape of the curve is relatively gradual in this case. It is expected to be steeper when there are more observations and the exact interval between which it is to be estimated will have smaller limits.

In the case that the C/E-acceptability curve exceeds 0.95 it will be possible to give a 95% maximum level. Here, in the case of costs per life saved, the C/E-acceptability curve takes the value 0.95 at about NLG 42,000. One may conclude that the ICER - for which the central estimate was NLG 4,928 - is acceptable with 95% certainty when a limit to cost-effectiveness is used of NLG 42,000. A similar calculation cannot be made for the case of costs per patient discharged. There is a 13.4% probability that the ICER will indicate higher costs and less effects. Therefore, the C/E-acceptability curve will never exceed 95% and an upper 95%-limit cannot be estimated.

fig 2.6 C/E-acceptability curve for costs per life saved; placebo versus high dose anakinra

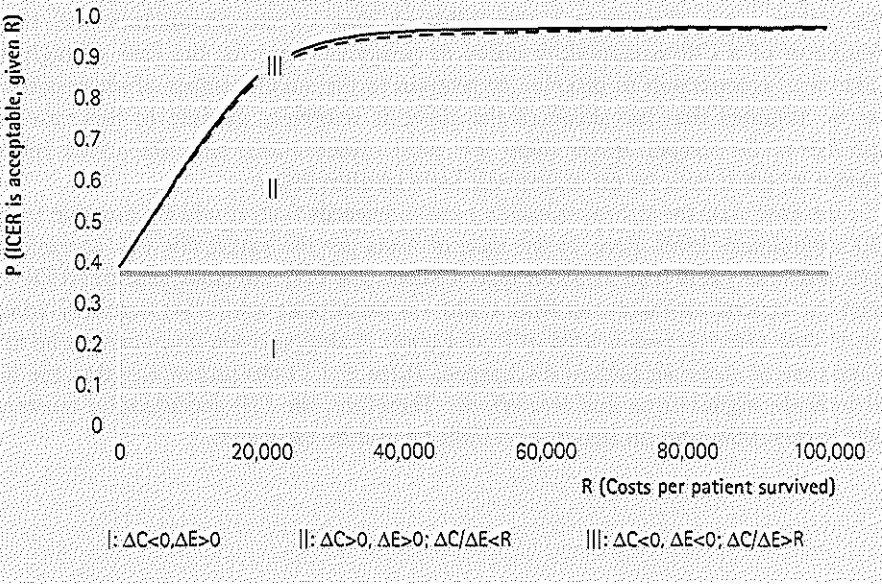
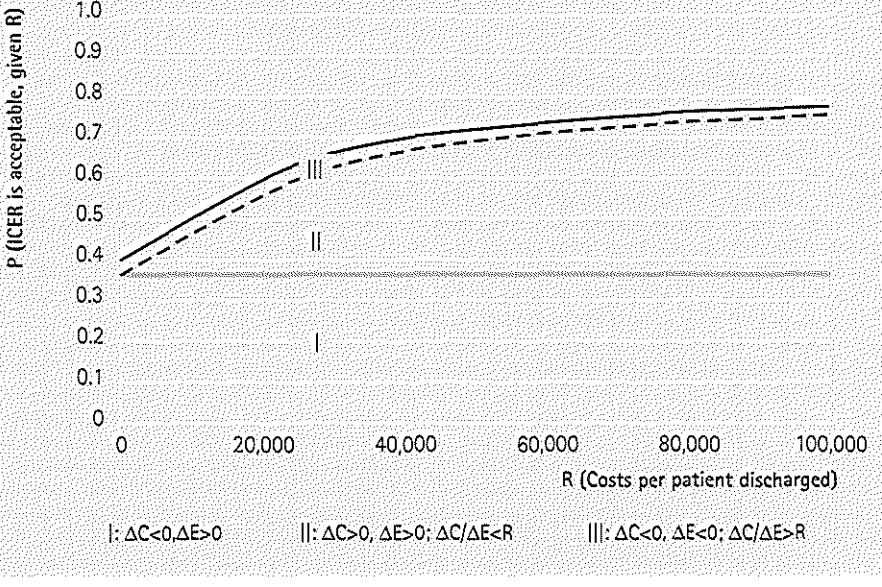


fig 2.7 C/E-acceptability curve for costs per patient discharged; placebo versus high dose anakinra



Discussion

In this chapter we have presented a methodology that is meant to give an idea about the uncertainty surrounding the estimate of the ICER. Additionally, the methodology may be useful in following the lines of thought as presented in earlier drafts of the guidelines for economic appraisal of pharmaceuticals in Ontario (see Laupacis et al., 1992). Here it was proposed to cluster various technologies according to their ICER and the C/E-acceptability curve can be used to indicate whether the ICER is within a certain cluster. As such the approach can be seen as an addition to common statistical analyses and therefore it is to be preferred that the relevant clusters and their boundaries are defined in advance together with the clinical hypotheses. This also prevents the use of the C/E-acceptability curve as a post-hoc instrument to safely indicate which limit has not been exceeded, an approach that might be seen as of questionable value.

The methodology has been illustrated using the results from a randomised clinical trial concerning anakinra. This is a typical example of a trial in which the number of observations is small, where the ICER looks relatively small in comparison to the average costs, but where there is considerable uncertainty about the reliability of the estimates. It is noted that the price of anakinra has not been included in the calculations and therefore it is emphasised that our example should be interpreted solely as an illustration.

With respect to the underlying assumptions, it is emphasised that the approach does not assume normally distributed costs and effects but it does assume that *average* costs and *average* effects are normally distributed. This assumption is based on the Central Limit Theorem and as such it is an approximation. Of course, the more observations, the better the approximation. Ideally, testing for normality would be possible by splitting the observations in a number of groups each with its own group mean, and testing whether these means are normally distributed. Such an approach would ask for many observations. When these are not available, like in the case of anakinra, one should be aware of the underlying assumption of normality, an assumption that is not formally tested.

A final remark concerns the usage of the methodology presented here for the outcome which is more commonly used: costs per life-year gained. For an unbiased estimate of the covariance between life-years gained and costs, complete patient specific observations are needed and the use of censored data may bias the estimates. Such complete observations are rarely gathered. Moreover, one often needs modelling techniques to estimate life-years gained and costs in later years. For these analyses sensitivity analysis will remain very useful and it may be expected that it will only be replaced by a formal procedure like the one presented here in some ideal cases.

Acknowledgements

We are grateful to Mark Wildhagen for his help with Mathematica and we thank Synergen for providing the Phase II data. We also thank Lesley Noe, John Pribble and two anonymous referees for their constructive comments.

Appendix

If observations per patient are available, estimates of means and variances are obtained by:

$$\hat{\mu}_{C_T} = 1/n_T \sum_{i=1}^{n_T} C_{iT} \qquad \hat{\sigma}^2_{C_T} = \frac{1}{n_T - 1} \sum_{i=1}^{n_T} \frac{(C_{iT} - \hat{\mu}_{C_T})^2}{n_T}$$

$$\hat{\mu}_{C_P} = 1/n_P \sum_{i=1}^{n_P} C_{iP} \qquad \hat{\sigma}^2_{C_P} = \frac{1}{n_P - 1} \sum_{i=1}^{n_P} \frac{(C_{iP} - \hat{\mu}_{C_P})^2}{n_P}$$

$$\hat{\mu}_{E_T} = 1/n_T \sum_{i=1}^{n_T} E_{iT} \qquad \hat{\sigma}^2_{E_T} = \frac{1}{n_T - 1} \sum_{i=1}^{n_T} \frac{(E_{iT} - \hat{\mu}_{E_T})^2}{n_T}$$

$$\hat{\mu}_{E_P} = 1/n_P \sum_{i=1}^{n_P} E_{iP} \qquad \hat{\sigma}^2_{E_P} = \frac{1}{n_P - 1} \sum_{i=1}^{n_P} \frac{(E_{iP} - \hat{\mu}_{E_P})^2}{n_P}$$

Here, the C_{iP} and E_{iP} refer to the costs and effects of patient i in the placebo group and the C_{iT} and E_{iT} refer to the patient specific data in the treatment group. The variables n_T and n_P refer to the number of observations in both groups.
Now:

$$\hat{\mu}_{\Delta C} = \hat{\mu}_{C_T} - \hat{\mu}_{C_P}$$
$$\hat{\mu}_{\Delta E} = \hat{\mu}_{E_T} - \hat{\mu}_{E_P}$$
$$\hat{\sigma}^2_{\Delta C} = \hat{\sigma}^2_{C_T} + \hat{\sigma}^2_{C_P}$$
$$\hat{\sigma}^2_{\Delta E} = \hat{\sigma}^2_{E_T} + \hat{\sigma}^2_{E_P}$$

The correlation coefficient can be estimated by:

$$\hat{\rho} = \frac{\hat{\rho}_T \hat{\sigma}_{C_T} \hat{\sigma}_{E_T} + \hat{\rho}_P \hat{\sigma}_{C_P} \hat{\sigma}_{E_P}}{\hat{\sigma}_{\Delta C} \hat{\sigma}_{\Delta E}}$$

where

$$\hat{\rho}_T = \frac{\sum_{i=1}^{n_T} (C_{iT} - \hat{\mu}_{C_T})(E_{iT} - \hat{\mu}_{E_T})}{\sqrt{\sum_{i=1}^{n_T} (C_{iT} - \hat{\mu}_{C_T})^2 \sum_{i=1}^{n_T} (E_{iT} - \hat{\mu}_{E_T})^2}} \qquad \hat{\rho}_P = \frac{\sum_{i=1}^{n_P} (C_{iP} - \hat{\mu}_{C_P})(E_{iP} - \hat{\mu}_{E_P})}{\sqrt{\sum_{i=1}^{n_P} (C_{iP} - \hat{\mu}_{C_P})^2 \sum_{i=1}^{n_P} (E_{iP} - \hat{\mu}_{E_P})^2}}$$

CHAPTER 3

Sample size calculation in economic evaluations

Summary

A simulation method is presented for sample size calculation in economic evaluations. As input the method requires: the expected difference and variance of costs and effects, their correlation, the significance level and the power of the testing method and the maximum acceptable incremental cost-effectiveness ratio. The method is illustrated with data from two trials. The first compares primary coronary angioplasty with streptokinase in the treatment of acute myocardial infarction, in the second trial, lansoprazole is compared with omeprazole in the treatment of reflux oesophagitis. These case studies show how the various parameters influence the sample size.

Given the large number of parameters that have to be specified in advance, the lack of knowledge about costs and their standard deviation, and the difficulty of specifying the maximum acceptable cost-effectiveness ratio, the conclusion of the study is that from a technical point of view it is possible to perform a sample size calculation for an economic evaluation, but one should wonder how useful it is.

Introduction

In general, randomised double-blind controlled clinical trials are seen as the optimal method to evaluate the efficacy and safety of a new medical technology. These trials may be either designed to prove superior efficacy or equivalence of the new treatment. An important part in the design of a clinical trial is the calculation of the sample size required to demonstrate, with high statistical significance, that the new treatment is equivalent or superior if indeed it is.

Nowadays, more and more decision makers are not only interested in safety and efficacy, but also in the balance between costs en effectiveness (e.g. life years or quality-adjusted life-years), and a goal of a study may be to prove that the new treatment is cost-effective, i.e. that the incremental cost-effectiveness ratio (ICER) is below a certain limit. As a result, in the design of a study increasingly the question is asked what the sample size should be to be able to prove cost-effectiveness. One may expect the required sample size to be larger compared to efficacy trials, since variation in costs is usually larger than the variation in treatment efficacy.

For efficacy trials we know that the required sample size is depended on the assumptions made about efficacy in both treatment groups. If efficacy is measured as a proportion, assumptions have to be made about only two parameters. Even with this small number, the influence of small changes in the assumptions on the required sample size can be large. This problem will only be exacerbated by assessing cost-effectiveness, since assumptions have to be made about many more variables.

The purpose of this study is twofold. First, we will show how, using a formula for the confidence interval of the ICER previously published (see Sacristan et al., 1995; Laska et al., 1997), the required sample size may be estimated. Second, we will assess how sensitive this estimate is for changes in assumptions about the expected costs, effectiveness and the correlation between them.

Data from two trials will be used to illustrate this: one comparing primary coronary angioplasty and streptokinase in the treatment of acute myocardial infarction and the other comparing the efficacy and safety of lansoprazole with omeprazole in the treatment of reflux oesophagitis.

Theory

Before sample size calculations can be performed, the hypothesis to be tested, the method of testing, and the acceptable errors should be known. In testing efficacy the null-hypothesis usually is $E_T - E_C = 0$, with E_T as the expected effects in the treatment group, and E_C representing the expected effects in the comparator group. If the efficacy parameter is a continuous variable a t-test will be performed to test the hypothesis, if it is a dichotomous variable, for instance cure, a chi square test might be used. In testing, the probability that an observed (or more extreme) difference could have been obtained if the null hypothesis were true is calculated. If this probability is $< \alpha$, the null hypothesis is rejected. Often 5% is chosen for α. Thus, the probability of falsely rejecting the null hypothesis (type I error), is less than 5%. However, there is also a chance that one might accept the null hypothesis, when the alternative is true (type II error). The probability that this happens is usually denoted by β. The power of a study to detect an effect of a specified size then is $1 - \beta$. If α, the sample size and the effect one wants to detect are specified, the power may be calculated. On the other hand, it is also possible to fix β in advance and to calculate the required sample size so that the study will have a high probability of finding a true effect of a given magnitude. Much literature has already been devoted to sample size calculations for efficacy trials (e.g. Lwanga & Lemeshow, 1991; Cohen, 1997).

But imagine that instead of efficacy cost-effectiveness has to be proven. What should be the null hypothesis, how should it be tested, and how can the required sample size be calculated? In economic evaluations the variable of interest is the ICER:

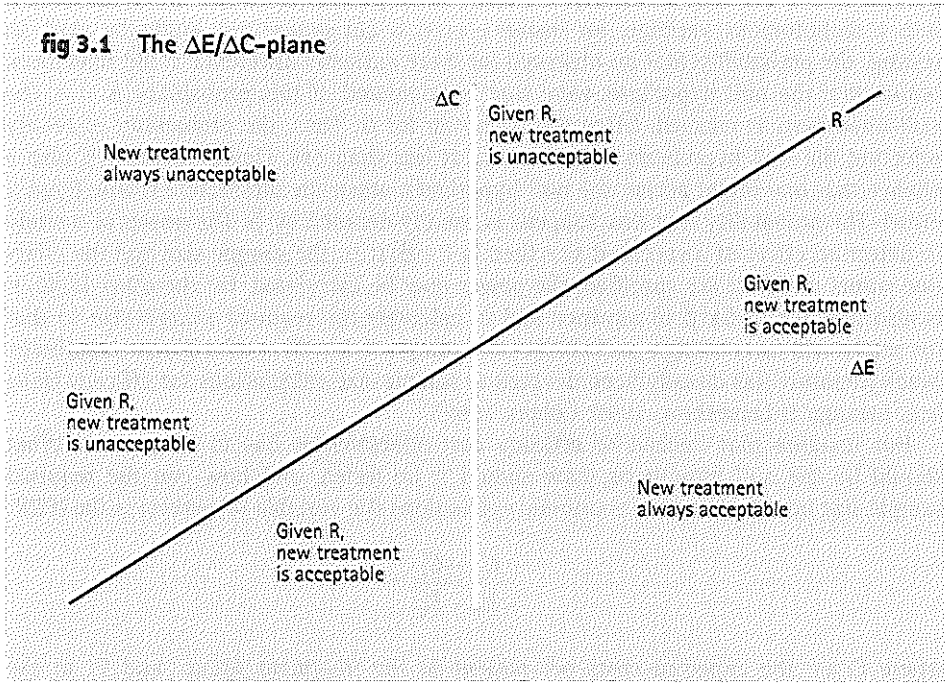
$$\frac{C_T - C_C}{E_T - E_C},$$

where C_T are the expected costs per patient in the treatment group, and C_C in the comparator group, respectively, and E_T and E_C here represent effectiveness rather than efficacy. This ratio may be estimated by

$$\frac{\overline{C_T} - \overline{C_C}}{\overline{E_T} - \overline{E_C}}.$$

One may assume that society will put a limit R on the additional costs that will be accepted to obtain additional effects. Now, the null hypothesis should be that $\Delta C/\Delta E$ is unacceptable, i.e. that it is above the R-line (above the R-line means $\Delta C/\Delta E < R$ if $\Delta E<0$ and $\Delta C/\Delta E > R$ if $\Delta E>0$, see figure 3.1). To test this hypothesis, a confidence interval for the ICER may be calculated and then one may check whether the interval is below the R-line.

fig 3.1 The ΔE/ΔC-plane



The formula for the confidence interval of a ICER is presented by several authors (see e.g. Sacristan et al., 1995; Laska et al., 1997) and is based on Fieller's theorem (Fieller, 1954). The limits L_1 and L_2 of the confidence interval of the true ICER may be written as:

$$\frac{(\Delta C\Delta E - t^2_{1-\frac{1}{2}\alpha}\,\mathrm{cov}_{\Delta C,\Delta E}) \pm [(\Delta C\Delta E - t^2_{1-\frac{1}{2}\alpha}\,\mathrm{cov}_{\Delta C,\Delta E})^2 - (\Delta C^2 - t^2_{1-\frac{1}{2}\alpha}\sigma^2_{\Delta C})(\Delta E^2 - t^2_{1-\frac{1}{2}\alpha}\sigma^2_{\Delta E})]^{1/2}}{(\Delta E^2 - t^2_{1-\frac{1}{2}\alpha}\sigma^2_{\Delta E})}$$

where $t_{1-1/2\alpha}$ is the upper $1-1/2\alpha$ percentage point of a student distribution with $(n_T-1)+(n_C-1)$ degrees of freedom, where n_T and n_C indicate the sample sizes in both groups (in the appendix it is shown how estimates of all parameters can be obtained when observations are available).

If the denominator of the above equation is <0, the limits L_1 and L_2 represent a confidence interval of the form $(-\infty, L_1] \cup [L_2,\infty)$ (this occurs when ΔE does not significantly differ from 0), if the denominator is >0, we find a confidence interval of

the form $[L_1,L_2]$. If both ΔC and ΔE do not significantly differ from 0 at the chosen significance level, no limits will be found. This means that at that significance level, the data are consistent with all possible hypotheses concerning the value of $\Delta C/\Delta E$, and thus the null-hypothesis can not be rejected.

Fieller's theorem, on which the confidence interval is based, assumes that the underlying distributions are normal. When working with costs, this will almost certainly not be the case. However, we are interested in the distribution of the average costs and the average effects, and the Central Limit Theorem tells that the distribution of those will approximate a normal distribution if the sample size is large enough. If our proposed sample size estimation method shows that the required sample size is small, one might suggest the use of a bootstrap method (see e.g. Chaudhary & Stearns, 1996; Briggs et al., 1997; Polsky et al., 1997) for the calculation of the confidence interval. However, performing a bootstrap simulation within a simulation will increase computation time considerably.

Since the hypothesis to be tested is one-sided, we seek a 95% one-sided confidence interval, which is equivalent to calculating a 90% two-sided confidence interval (of which one limit is disregarded). If the confidence interval is now below the R-line, we can reject the null hypothesis that the ICER is unacceptable, i.e. that it is above the R-line, at the 5% significance level.

Having answered the questions what the null hypothesis should be, and how it should be tested, the next section will present a simulation method to calculate the sample size for a given power and α.

Sample size calculation

If a trial that aims at demonstrating that the ICER is below a certain limit (say, arbitrarily, NLG 40,000 per additional survivor) with $\alpha = 5\%$ and a power of 80% is set up, the required sample size can be calculated by simulation. For this simulation, estimations of the average costs (μ_{CT}, μ_{CC}) and effects (μ_{ET}, μ_{EC}) in both groups, the standard deviation of the costs (σ_{CT}, σ_{CC}) and effects (σ_{ET}, σ_{EC}) for both groups and the correlation between costs and effects (ρ) are required. Furthermore, an assumption should be made about the distribution of the costs and effects in the patient groups. For the costs, often a log normal distribution will be a realistic assumption, if the effect measure is dichotomous a binomial distribution is appropriate. It is assumed that the sample size should be equal in both groups, so $n_T = n_C = n$.

Now the estimations are assumed to represent the true distribution of costs and effects, and thus we can also calculate the expected ICER, which could be for example NLG 20,000. Since this ICER is indeed below R = NLG 40,000, we expect that the trial outcome will lead to rejection of the null hypothesis that the ICER is above the R-line.

The next step is to simulate trial outcomes by drawing random samples from the specified cost and effect distributions. First this is done for a trial with the arbitrary sample size n=25. For this sample, it can be tested by means of the method explained

in the Theory section, whether the ICER is above the R-line. If the confidence interval includes R, the ICER is unacceptable, and thus the null hypothesis will be accepted. Remember that the distributions of costs and effects are specified in such a way that we know that the true outcome (NLG 20,000) lies below the line R = NLG 40,000. So, by accepting the null hypothesis we will make an error.

If 1,000 samples of size 25 are drawn, sometimes the null hypothesis is rejected, in the other samples it will be accepted. By counting how often the null hypothesis is rightly rejected, the power may be estimated. In this example, if for 400 of the 1000 trials with a sample size of 25 the right decision is made, the power will be 40%. Since this power is too low, the procedure can be repeated, now with a sample size of n=50. By subsequently increasing the sample size, the required sample size for a power of 80% can be found. It should be emphasised that the limit R should be specified in advance, since this limit has a large influence on the power of the study.

To summarise, a backward procedure is used to calculate the required sample size i.e. for a given n, the power is calculated. As long as the power does not equal the value predefined, the sample size is changed and the new power is calculated. This is repeated until the power has reached the required value. This simulation method is easily programmed, for this study Pascal was used.

Below, the proposed method will be illustrated with data from two clinical trials. In the first, both additional costs and additional effects occur, in the second, effects are equal for both treatment groups, and the additional costs are mainly dependent on the price difference between the two drugs considered.

Illustration 1: angioplasty vs. streptokinase

In the Zwolle trial patients with acute myocardial infarction either underwent primary coronary angioplasty or received intravenous streptokinase (see Boer et al., 1994). Alongside this trial, an economic evaluation was done (see Boer et al., 1995). Effects were measured in terms of event-free survival after 1 year, here defined as survival without recurrent myocardial infarction, stroke or additional revascularization procedures. In the angioplasty group 103 patients out of 152 (68%) survived without events, compared to 67 patients out of 149 (45%) in the streptokinase group.

Costs were defined as all direct medical costs during the first year, including all hospital admissions, additional procedures, and other medical events. The average costs per patient in the angioplasty group were NLG 27,354 (standard deviation NLG 16,245); in the streptokinase group the average costs per patient were NLG 26,264 (standard deviation NLG 14,570). The ICER was NLG 4,739 per event-free survivor, and the 95% one-sided confidence interval was (-∞, 24,137].

Consider the hypothetical situation where a new trial is to be set up, aiming at proving superior efficacy of angioplasty in terms of event-free survival after 1 year. Suppose that, from literature study, we postulate that the probability of success in the angioplasty group is $\mu_{ET} = 0.68$ and in the streptokinase group $\mu_{EC} = 0.45$. If the significance level α is set at 5%, in order to achieve a power of 80% we require a sample size of 65 patients per group.

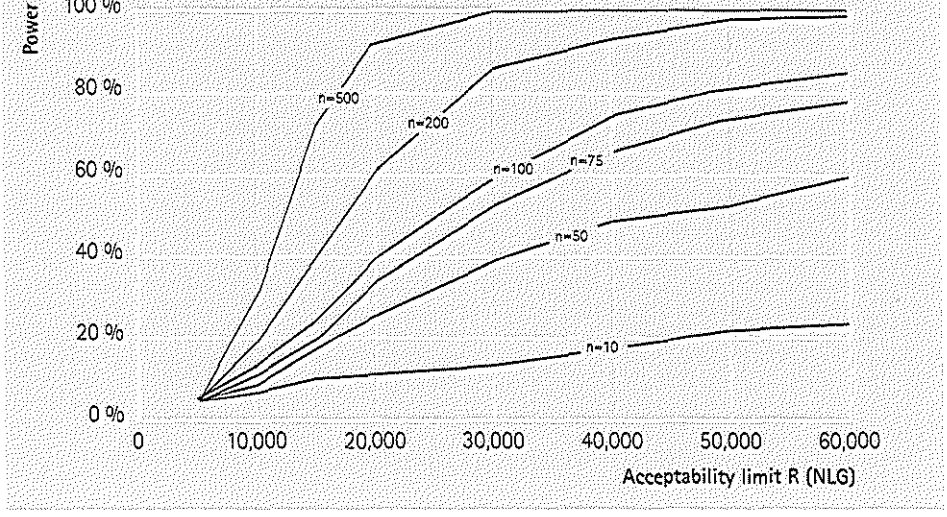
Now it is decided that proving superior efficacy is not sufficient, it should be proven

that the ICER is below the (arbitrarily chosen) limit of NLG 40,000 per additional event-free survivor. We use the same μ_{ET} and μ_{EC}, and for the costs in each group we postulate μ_{CT} = NLG 27,354, σ_{CT} = NLG 16,245, μ_{CC} = NLG 26,364 and σ_{CC} = NLG 14,570. For the moment, we assume ρ=0. The same α and power are assumed.

Using the simulation method described in the previous section, it is estimated that each group should consist of 125 patients. Figure 3.2 presents the relation between power and sample size for various limits. Clearly, as the limit increases, the sample size required to prove that the ICER lies below that limit decreases. For instance, if the limit is increased from NLG 17,500 to NLG 25,000, the sample size decreases from 500 to 250.

In this example, no correlation was assumed between costs and effects (ρ = 0). The simulations were repeated with ρ = -0.5 and ρ = 0.5, respectively. If the limit of R = NLG 40,000 is used, postulating a correlation of -0.5 results in a required sample size of 175 per group, whereas a correlation of 0.5 yields a required sample size of 70 per group. These results indicate that if a negative correlation is postulated, the required sample size increases, and if a positive correlation is postulated, the required sample size decreases. If other estimates of costs and effects are used, the above still holds. An intuitive explanation for this is that if costs and effects are positively correlated, the simultaneous distribution of the costs and effects is more or less 'parallel' to the R-line. If on the other hand the correlation is negative, this distribution is perpendicular to the R-line. Thus, a much larger sample size is required for the simultaneous distribution to lie below the R-line.

fig 3.2 **Relation between power, sample size and maximum acceptable incremental C/E ratio using data from the Zwolle trial**



We also used the simulation method to study the effect of increasing or decreasing the standard deviation of the costs, by either multiplying or dividing the standard deviation by 2. Table 3.1 shows that the closer R is to the ICER (NLG 4,739), the larger the effect of changes in the standard deviation is.

table 3.1 Effect of changes in standard deviation of the costs on sample size

R	n	SD *2 n	SD /2 n
20,000	350	1050	150
30,000	170	400	100
40,000	120	275	85
50,000	100	175	75

Finally, to see how small changes in all parameters simultaneously may effect the resulting sample size, we entered rounded estimates of costs and effects into the simulation method. So, our parameter estimates are now: μ_{CT} = NLG 27,000, σ_{CT} = NLG 15,000, μ_{CC} = NLG 26,000, σ_{CC} = NLG 15,000, μ_{ET} = 0.7, μ_{EC}=0.45, ρ=-0.5 and R=NLG 40,000. This results in a required sample size of 230 instead of 175, an increase of 31%.

Illustration 2: lansoprazole vs. omeprazole

Lansoprazole is a new proton pump inhibitor, developed for the treatment of acid-related gastrointestinal diseases like peptic ulcers and reflux oesophagitis. Proton pump inhibitors such as lansoprazole and omeprazole establish a more rapid and frequent healing of erosive and ulcerative lesions than histamine-receptor antagonists like ranitidine and famotidine (see Bardhan et al., 1995; Umeda et al., 1995). A trial was set up to compare lansoprazole 30 mg and omeprazole 40 mg in the treatment of acute reflux oesophagitis (see Mulder et al., 1996).

Patients were included if they had endoscopically proven ulcerative, non-bleeding, reflux oesophagitis grade II, III or IVa. The duration of treatment was 4 weeks, but if lesions had not been healed after 4 weeks, the treatment was extended to 8 weeks. The lansoprazole group consisted of 106 patients, the omeprazole group of 105. Cure was defined as grade 0. After 4 weeks 86% of patients in the lansoprazole group and 80% of patients in the omeprazole were cured (confidence interval for difference: [-5%,+18%]). After 8 weeks the healing rates were 96% for the lansoprazole group and 93% for the omeprazole group (confidence interval for difference: [-4%, +10%]).

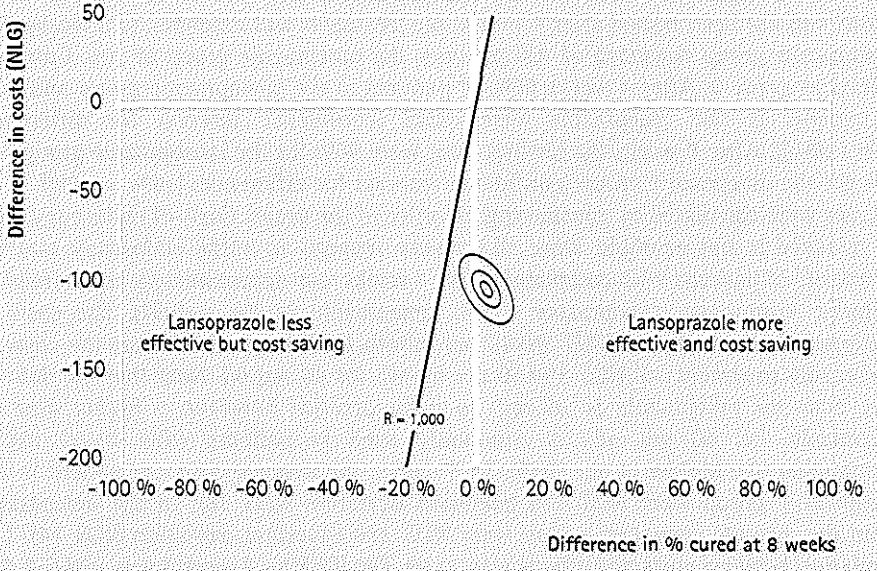
Costs were calculated by considering the price of the drugs. Costs of endoscopy were not included in this example, since all patients undergo an initial endoscopy, and the percentage of patients with complaints after 4 weeks was approximately equal (mild to moderate symptoms: lansoprazole 25.8%, omeprazole 27.2%, p=0.91). The costs of lansoprazole treatment are NLG 110.45 per month for 30 mg od whereas the costs of omeprazole treatment are NLG 193.50 per month for 40 mg od. Note that the defined daily dose for omeprazole is 20 mg. In daily practice however, a higher dose of omeprazole, i.e. 40 mg od., is commonly used in patients with severe reflux oesophagitis (see Mulder et al., 1996). At 8 weeks, the average costs per patient in the lansoprazole group were NLG 125 (standard deviation NLG 38); in the omeprazole group the average costs per patient were NLG 228 (standard deviation NLG 74). The correlation between costs and effects was -0.48. The ICER (at 8 weeks) was - NLG 3,430 per additional patient cured, and the 95% one-sided confidence interval was $(-\infty, 0] \cup [4{,}287, \infty)$. Figure 3.3 shows the 95% confidence ellipse, i.e. a simultaneous confidence region for both costs and effects. Clearly, even though efficacy of lansoprazole did not significantly differ from that of omeprazole, it is cost saving.

Again, consider the hypothetical situation where a new trial is set up, with the aim to prove that lansoprazole and omeprazole are equally effective. The assumption is that the probability of cure in the omeprazole group will be 90%. It now has to be decided which cure rate for lansoprazole will be considered equally effective and this decision should be based on which difference is considered clinically relevant. In this example we assume that the efficacy of lansoprazole will be regarded as equivalent if its cure rate is within 5 percent points of 90%, Furthermore, $\alpha = 5\%$ and $\beta = 20\%$ are chosen. By using formulas given by Makuch & Johnson (1989), it may be calculated that such equivalence study requires 756 patients in each group.

Starting a trial with 756 patients in each group will be very expensive, so it may be worthwhile to try to prove cost-effectiveness instead of equivalence.

We assume that effectiveness in both groups is equal, so we postulate $\mu_{ET} = \mu_{EC} = 0.90$; for the costs the assumptions are μ_{CT} = NLG 125, σ_{CT} = NLG 38, μ_{CC} = NLG 228 and σ_{CC} = NLG 74, and we expect the correlation to be ρ=-0.5.

fig 3.3 Bivariate normal distribution of costs and effects comparing lansoprazole and omeprazole



The outer ellipse is the 95%, the middle ellipse the 50%, and the inner ellipse the 5% confidence ellipse.

In deciding on the limit R, it is a good idea to first have a look at figure 3.3. It shows that lansoprazole will almost certainly be cost-saving, so the limit R now indicates how much costs we should save, before we are willing to accept 1 patient less cured by lansoprazole. If a limit of R = NLG 500 is chosen (curing 1 patient less should result in cost savings of at least NLG 500), the simulations show that, with α = 5%, already with 40 patients in each group a power of 80% is reached. If the limit is R = NLG 1,000 the required sample size per group becomes 125, and for R = NLG 1,500 the sample size should be at least 270 patients per group.

Just to see what the effect is of the difference in treatment costs, we ran the simulation again, now assuming the costs of lansoprazole per month are NLG 155, leading to μ_{CT} = NLG 176, σ_{CT} = NLG 53 (all other parameters remain unchanged).

If R = NLG 500, 175 patients per group are required to arrive at a power of 80%. Using a stricter limit of R = NLG 1,000, 520 patients per group are necessary, and for R = NLG 1500 the required sample size becomes 1100 per group.

To summarise, in this example of equivalent efficacy, the number of patients required to prove that the ICER is acceptable is much smaller than the number required to prove equivalency. Furthermore, decreasing the difference in treatment costs provided that all other parameters remain constant, leads to an increase of the required sample size.

Discussion

Analogous to the calculation of sample size in efficacy or equivalence trials, one would like to calculate sample size for economic evaluations. This chapter presents a simulation method that enables the calculation of sample sizes for economic evaluations, by using the confidence interval for ICERs that has been published by several authors (e.g. Sacristan et al., 1995; Laska et al., 1997).

It is important to realise that for sample size calculation in economic evaluations, more prior information is required compared to efficacy or equivalence trials. In the illustrations presented in this chapter only dichotomous effect measures have been used. If instead life-years gained or quality-adjusted life-years (QALYs) gained are used, not only the expected outcome should be estimated, but the standard deviation as well. Furthermore, the required sample size is influenced by the expected costs and their standard deviations and the correlation between costs and effects. In practice it may be difficult to find a reasonable estimate for the expected costs, and it will be even more difficult to find an estimate for the standard deviation of the costs. Moreover, the variation in costs will depend on how detailed resource utilisation is measured. For instance, if it is expected that some patients in a study may need a PTCA, one might decide to assign a fixed price to every PTCA performed, but one may also record length of hospital stay, all procedures, medication etc. per patient, to find the costs of the PTCA. The second method will lead to a larger standard deviation of the costs, and thus to a larger required sample size.

Besides the above mentioned variables the significance level and the power of the testing method should be specified in advance, but, moreover, the hypothesis to be tested should also be clearly defined, that is, it should be decided in advance which ICER is the maximum acceptable ratio. Defining this maximum acceptable ratio will not be easy in practice. If QALYs are used as a measure of effectiveness, one may look at the ICER of other technologies that have already been approved. Here, a distinction may have to be made between prophylactic treatment policies (e.g. breast cancer screening), common therapies (e.g. treatment of heart failure with ACE-inhibitors) and 'high tech' procedures (e.g. heart transplant). Unfortunately, many clinical trials do not use QALYs or life years as a measure of effectiveness, but intermediate (efficacy) measures, like disease-free period, or percentage cured. Such effectiveness measures make it difficult to compare technologies, and to decide what is acceptable. In our case studies, this problem clearly arises in the comparison of lansoprazole versus omeprazole. How does one determine how much society is willing to pay to cure one extra patient with reflux oesophagitis?

From the above it will be clear that it will not be difficult to manipulate the outcome of a sample size calculation. If one prefers a sample size of 100 for practical reasons, it is possible to specify parameters in such a way that the calculations result in a sample size of 100. So, one may conclude that from a technical point of view it is possible to perform a sample size calculation for an economic evaluation, but one should wonder how useful it is.

Regarding the technicalities of sample size calculation a remark should be made.

Stinnett (1996) has shown that the estimator for the ICER, as presented in the theory section, is biased but consistent. Consistent estimators are asymptotically unbiased, and therefore the bias may be negligible in studies with large sample sizes but is potentially important when sample sizes are small. This bias does not only influence the point estimate of the ICER, but also its confidence interval. Both Chaudhary & Stearns (1996) and Briggs et al. (1997) suggest that the so-called bias-corrected and accelerated bootstrap interval, which adjusts for bias of the estimator, may be used. For large sample sizes, the Fieller confidence interval and this bootstrap interval seem comparable. For smaller sample sizes, it might be better to use the bootstrap confidence interval, both because of the bias and because the parametric assumptions necessary to use Fieller may not be valid for small sample sizes. One should realise, however, that performing a bootstrap simulation within the sample size simulation will lead to a large increase of computation time.

Finally, one ought to be aware of the fact that the uncertainties in health care are not only of a statistical nature. An example might be the price difference between lansoprazole and omeprazole. Given the price difference one may state quite firmly that lansoprazole is very cost-effective. However, if the cost difference between both agents decreases, for example due to a price reduction of omeprazole, treatment with lansoprazole will be less or not cost-effective. One should take into account that just as with clinical trials research in health care is a snapshot in time and extrapolation to the future or to other countries may be difficult. That of course makes it even more important that statistical procedures are available to address those uncertainties in economic evaluations that *are* of a statistical nature.

Acknowledgements

The study was supported by a grant from Hoechst Marion Roussel, The Netherlands and Janssen-Cilag, The Netherlands. The authors are grateful for the comments made by Joseph C. Cappelleri.

Appendix

C_T represents the average costs per patient in the treatment group. For each patient i ($i=1...n_T$) an observation C_T is available. Now the average costs in this group may be estimated by:

$$\hat{C}_T = \frac{1}{n_T}\sum_{i=1}^{n_T} C_{iT} .$$

In the same way, C_C, E_T and E_C are estimated.

For the variances of the average costs and effects, the following estimates are applicable:

$$\hat{\sigma}^2_{\Delta C} = SE^2_{C_T} + SE^2_{C_C} \qquad \text{and} \qquad \hat{\sigma}^2_{\Delta E} = SE^2_{E_T} + SE^2_{E_C} \quad \text{where}$$

$$SE^2_{C_T} = \frac{1}{n_T(n_T-1)}\sum_{i=1}^{n_T}(C_{iT} - \hat{C}_T)^2 .$$

$SE^2_{C_C}$, $SE^2_{E_T}$ and $SE^2_{E_C}$ are calculated by the same formula.

Finally, the covariance between the average costs and effects may be estimated by:

$$\hat{\text{cov}}_{\Delta C,\Delta E} = \hat{\text{cov}}_{C_T,E_T} + \hat{\text{cov}}_{C_C,E_C} \qquad ;$$

$$\hat{\text{cov}}_{C_T,E_T} = \frac{1}{n_T(n_T-1)}\sum_{i=1}^{n_T}(C_{iT} - \hat{C}_T)(E_{iT} - \hat{E}_T) \qquad ;$$

$$\hat{\text{cov}}_{C_C,E_C} = \frac{1}{n_C(n_C-1)}\sum_{i=1}^{n_C}(C_{iC} - \hat{C}_C)(E_{iC} - \hat{E}_C) \qquad .$$

CHAPTER 4

A Bayesian approach to economic analyses of clinical trials:

the case of stenting versus balloon angioplasty

Summary

New results about the costs and effects of a new therapy may be weighted with prior information. As such, classical confidence intervals surrounding the costs and effects of a therapy may not reflect the real uncertainties. Bayesian techniques may improve this by formalising the way that prior information is taken into account in assessing the new evidence. Costs and effects can be analysed separately, but also, when considering the balance between costs and effects, they can be analysed simultaneously. Here, an example is given using data from two trials that compared costs and effectiveness of stent implantation versus balloon angioplasty. The Bayesian results make it clear that different prior distributions may lead to different decisions and it is concluded that even Bayesian analysis may not always reflect the process of capturing the remaining uncertainties.

Introduction

Economic evaluations assessing the balance between the costs and effects of a therapy may not only summarise the results in terms of single point estimates for the incremental costs, effects and cost-effectiveness ratio (ICER), but they may also indicate what the uncertainties are surrounding these estimates. Even when the uncertainty around the difference in efficacy is reasonable or small, the uncertainty around the difference in costs and the ICER can be quite large. However, the real uncertainties that remain after a study has been completed may vary from what is suggested by the data alone. There may be prior information or prior ideas and the results will often be valued with this prior knowledge in mind. Bayesian analysis offers a method to take explicit account of this, potentially reducing the uncertainty surrounding the cost-effectiveness ratio.

An additional advantage of using a Bayesian approach is that the interpretation of the results is more appealing. In the classical frequentist way of analysing data, results are often dichotomised: a difference is either significant or not. And although some journals now propagate the use of confidence intervals instead of P-values, the interpretation of the confidence interval within a frequentist framework is hardly appealing. A frequentist 95% confidence interval tells us that if the trial was repeated often enough and if a confidence interval was calculated each time, 95% of those intervals would contain the true difference. Most people, however, are inclined to interpret the interval as having a 95% chance that the true difference lies within the interval. Within the frequentist framework this interpretation is not allowed, whereas it is within the Bayesian framework.

Many articles have been published propagating the use of Bayesian analysis in biostatistics (e.g. Breslow, 1989; Spiegelhalter et al., 1993; Kadane, 1995; Brophy & Joseph, 1995; Lilford & Braunholtz, 1996) and pointing out the potential advantages of using Bayesian analysis in economic evaluations (e.g. Eddy et al., 1990; Manning et al., 1996; Jones, 1996). Moreover, Manning et al. (1996) state that a major goal of methodologists in cost-effectiveness analysis should be to further explore methods like Bayesian analysis. And while some first steps were taken by Eddy et al. (1990) about 10 years ago, concentrating on univariate analyses estimating single parameters, it was only recently that Heitjan et al. (1999) addressed Bayesian estimation of a cost-effectiveness ratio, using non-informative priors.

In this chapter, we summarise our first Bayesian experiences of an economic analysis based on data from a randomised clinical trial also using informative priors. First, a univariate approach for the analysis of the differences in costs and effects is presented. Second, some initial ideas about a multivariate approach for the analysis of the ICER are presented. Subsequently, the methodology is applied using data from the Benestent I and II trials (see Serruys et al., 1994; Serruys et al., 1996; Serruys et al., 1998). For both trials, the balance was assessed between the costs and effects of stent implantation in comparison with balloon angioplasty. Finally, the results are reviewed by asking how the Bayesian results compare with the classical ones in reflecting our uncertainties.

Background

Currently, there are a number of ways to assess the uncertainties surrounding estimates of costs, effects and the ICER. Within the recent literature, the assessment of the ratio has drawn most attention, and currently there are at least two methods that are commonly used to find a confidence interval surrounding this ratio: Fieller's method (Fieller, 1954) and bootstrapping (see e.g. Chaudhary & Stearns, 1996). Both methods do not use any prior information and inference is solely based on the data from the trial. Usually no assumption is made about the underlying distribution of costs and effects. In practice it is only assumed that the average costs and effects follow a normal distribution which is reasonable for sample sizes large enough. Which sample size is large enough depends on the underlying distributions (see Briggs & Gray, 1998a), for binomial distributions, for example, a rule of thumb is $np>5$ and $n(1-p)>5$. When the number of observations is large, both methods lead to similar results concerning the distribution surrounding the ICER.

In order to carry out a Bayesian analysis, it is necessary to assume parametric probability distributions for both costs and effects and their simultaneous distribution. When such a probability distribution is seen as a function of its parameters, for fixed observations, it is called the likelihood function. Additionally, prior distributions need to be specified for the parameters of both distributions. Subsequently, posterior distributions are obtained after combining the data from the experiment with the knowledge reflected by the prior distributions. Thus, the posterior distributions reflect all uncertainties after having carried out the study.

A variety of likelihood functions can be chosen for costs and effects and an additional variety of distributions can be chosen for the priors. Generally, both the likelihood function and the prior distribution may be extremely flexible, such as a mixture of normals. Most combinations of prior distribution and likelihood will result in posteriors that do not follow any standard parametric form. Subsequently, evaluation of such posteriors may require extensive simulation methods, like Markov Chain Monte Carlo (MCMC) methods (see e.g. Gelman et al., 1995; Carlin & Louis, 1996). If one does not want to use these simulation methods, the potential choices of distributions are rather limited. Here, we will avoid MCMC simulation by choosing convenient priors for the parameters of the distributions and we will concentrate on the situation where costs may be characterised by a log normal distribution and where effectiveness is described by a dichotomous variable.

If one has no prior ideas about the distribution of the parameters being studied, one may choose a non-informative prior. For the mean of a normal distribution, a non-

informative prior would be a distribution giving equal weight to all values from $-\infty$ to ∞ (e.g. Carlin & Louis, 1996). However, a prior distribution may also be more informative, for example based on the ideas of one or more experts. In this case, it is advisable to explore a number of different priors and see how the conclusions after combining data with these priors differ.

Another possibility is to use data from previously completed studies. This approach is followed by Eddy et al. (1990): updating the distribution of the parameter as data comes along. Here, if a technology is studied for the first time, a non-informative prior is used. Subsequently, when a second study is conducted, the posterior from the first analysis is used as a prior for the analysis of the second study, and so on. In practice, such an approach will lead to similar results as a 'classical' fixed-effect meta-analysis.

In the example discussed here, the last approach will be followed. Additionally, when addressing the multivariate model, we will also follow the idea of Brophy & Joseph (1995) and we will weigh the data from the previous trial when specifying the prior distribution. This is achieved by using the observed mean from the previous trial(s) as the mean of the prior, and by multiplying the observed variance from the previous trials with a factor greater than one to define the variance of the prior. This factor is chosen greater than one to reflect additional uncertainty due to unknown factors.

Methods

Costs - univariate analysis

A Bayesian analysis of costs starts with assuming a parametric distribution. This may be any kind but here we will assume - as indicated before - a log normal distribution (i.e. the log of the costs follow a normal distribution) characterised by two parameters, μ and σ^2. If we have n observations of c, the likelihood of all the observations is the product of the likelihood (the probability density function (pdf) $f(c|\mu,\sigma^2)$ regarded as a function of μ and σ^2, for a fixed observed value c) of each individual observation, i.e.

$$f(c \mid \mu,\sigma^2) = \prod_{i=1}^{n} f(c_i \mid \mu,\sigma^2) .$$

Additionally, it is assumed that the parameters μ and σ^2 themselves have a pdf $\pi(\mu,\sigma^2)$, the so-called prior distribution. Here we assume that $\pi(\mu,\sigma^2) = \pi(\mu)\pi(\sigma^2)$ and that $\pi(\mu) = N(\mu_0,\tau^2_0)$. Furthermore, following the literature we choose an inverse gamma distribution Inv-$\chi^2(\nu_0,\sigma^2_0)$ for $\pi(\sigma^2)$ (e.g. Gelman et al., 1995; Box & Tiao, 1973).

The posterior distribution of μ and σ^2, given our observations, can be calculated as:

$$f(\mu,\sigma^2 \mid c) = \frac{\pi(\mu,\sigma^2) f(c \mid \mu,\sigma^2)}{\int \pi(\mu,\sigma^2) f(c \mid \mu,\sigma^2) d\mu d\sigma^2} ,$$

$$\propto \pi(\mu,\sigma^2) f(c \mid \mu,\sigma^2)$$

or, in words, as the product of the prior distribution and the likelihood function, up to a normalising constant.

Given that the combination of the various distributions results in a posterior that does not follow a standard parametric form, we use the fact that the conditional marginal posterior distribution of μ given σ^2 is normal:

$$\mu \mid \sigma^2, c \sim N(\mu_n, \tau_n^2) \quad , \qquad \textbf{Eq. 1}$$

where

$$\mu_n = \frac{\sigma^2/n}{\sigma^2/n + \tau_0^2}\mu_0 + \frac{\tau_0^2}{\sigma^2/n + \tau_0^2}\overline{\log c_i} \qquad \text{and} \qquad \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

($\overline{\log c_i}$ denotes the average of the log costs).

For the marginal posterior distribution of σ^2 it can be shown that (e.g. Gelman et al., 1995):

$$f(\sigma^2 \mid c) \propto \tau_n \, N(\mu_n \mid \mu_0, \tau_0^2) \, \text{Inv}\text{-}\chi^2(\sigma^2 \mid \nu_0, \sigma_0^2) \prod_{i=1}^{n} N(\log c_i \mid \mu_n, \sigma^2) \; . \qquad \textbf{Eq. 2}$$

This posterior function can now be evaluated on a discrete grid of values of σ^2 by choosing values of σ^2 at regular intervals, calculating $f(\sigma^2|c)$ for each of these values, and by normalising the results so that they add up to 1. Subsequently, values can be drawn (for instance 1000) from this grid (i.e. from the posterior distribution), and these values can be used to estimate the posterior mean and a 95% posterior interval (the Bayesian equivalent of a confidence interval) for σ^2. Furthermore, every draw of σ^2 can be used to draw a posterior value of μ using the conditional posterior distribution of μ, thus allowing the estimation of the posterior mean of μ, and the estimation of a 95% posterior interval.

By saving all of the simulated values of σ^2 and μ, the expected costs can then be calculated by using

$$E(\text{costs}) = e^{\mu + 0.5 * \sigma^2} \; . \qquad \textbf{Eq. 3}$$

When starting the analysis, one may or may not use prior information or ideas about the parameters of the cost distribution. Without such information, so-called 'non-informative' distributions are used for μ and σ^2. The prior distribution for μ, $N(\mu_0,\tau^2_0)$ can be made non-informative by choosing the variance large enough to ensure that the choice of μ_0 has little influence on the posterior. Analogously, the inverse gamma distribution Inv-$\chi^2(\nu_0, \sigma^2_0)$ for σ^2 becomes non-informative if ν_0 is chosen sufficiently small to ensure that the choice of σ^2_0 has little influence on the posterior.

Effectiveness – univariate analysis

As with costs, a Bayesian analysis of effects starts with assuming a parametric distribution. When considering for example event-free survival at a certain point of time, one may use a binomial distribution with parameters n and p. In that case, the beta-distribution may be a convenient prior, characterised by two parameters (a,b) and defined on $p \in [0,1]$. It is convenient in the sense that it is conjugate to the binomial distribution, i.e. it yields a beta posterior distribution. If y is the number of event-free survivors and n is the sample size, the posterior pdf is Beta(a+y,b+n-y) (see Gelman et al., 1995), with

$$E(p \mid y) = \frac{a+y}{a+b+n} \quad \text{and} \quad Var(p \mid y) = \frac{E(p \mid y)[1 - E(p \mid y)]}{a+b+n+1} \; . \qquad \textbf{Eq. 4}$$

A non-informative distribution is found by choosing a=b=1, equivalent to a uniform prior.

Incremental cost-effectiveness ratio – multivariate analysis

For the assessment of the uncertainty surrounding the ICER, costs and effects have to be analysed simultaneously in order to take account of the correlation between costs and effects. Suppose that we have observations $(c_1,e_1),...,(c_n,e_n)$ on costs and effectiveness for n patients. These costs and effects have a pdf with one or more parameter $f(c,e|\theta_1,...,\theta_k)$, and it is assumed that the parameters $\theta_1,...,\theta_k$ themselves have a prior pdf $\pi(\theta_1,...,\theta_k)$. According to Bayes' rule the posterior distribution is

$$f(\theta_1,...,\theta_k \mid c,e) \propto \pi(\theta_1,...,\theta_k) f(c,e \mid \theta_1,...,\theta_k) .$$

Often, it will not be easy to define the likelihood function $f(c,e|\theta_1,...,\theta_k)$ directly. When applying the examples for our univariate case, a simultaneous distribution for costs and effects should be defined so that the marginal distributions are a log normal and a binomial distribution. If a suitable likelihood function is found, it will often lead to a complicated posterior distribution that requires advanced techniques like MCMC methods to be evaluated. However, when confronted with large sample sizes one might also use an approximation based on the Central Limit Theorem. In that case, the averages of costs and effects approximately follow a bivariate normal distribution, or

$$\begin{pmatrix} \bar{c} \\ \bar{e} \end{pmatrix} \mid \begin{pmatrix} \mu_c \\ \mu_e \end{pmatrix}, \Sigma \sim N\left(\begin{pmatrix} \mu_c \\ \mu_e \end{pmatrix}, \Sigma \right) ,$$

where Σ is the 2x2 variance matrix (here the variance is the square of the standard error).

If we write $\bar{y} = \begin{pmatrix} \bar{c} \\ \bar{e} \end{pmatrix}$ and $\mu = \begin{pmatrix} \mu_c \\ \mu_e \end{pmatrix}$, the likelihood function is

$$f(\bar{y} \mid \mu, \Sigma) \propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\bar{y} - \mu)^T \Sigma^{-1} (\bar{y} - \mu) \right) .$$

Our parameter of interest is μ. However, Σ is also an unknown parameter and the posterior distribution of μ when Σ is known differs from the distribution of μ when Σ is unknown (see Gelman et al., 1995). Without a formal solution for small sample sizes, we use the knowledge that for a large sample size the sample variance matrix converges to the true variance matrix. Thus, we assume that Σ is known and equal to the sample variance matrix (Heitjan et al., 1999 make the same assumption in their non-informative approach).

Now, given our former choices, it is only the prior distribution of μ - the parameter of interest - that remains to be specified. Choosing a normal prior may now be most convenient, considering that the likelihood is also normal and that the combination of both will lead to a normal posterior. Additionally, unless prior knowledge is heavily skewed, a bivariate normal prior will often be sufficient to describe the prior regarding μ. So, the prior can be defined as follows:

$$\mu \sim N(\mu_0, \Lambda_0),$$

and non-informative priors can be obtained by choosing large variances.
It can now be shown that the posterior can be written as (see Gelman et al., 1995)

$$f(\mu \mid \bar{y}, \Sigma) = N(\mu_1, \Lambda_1),$$

where

$$\mu_1 = \left(\Lambda_0^{-1} + \Sigma^{-1}\right)^{-1} (\Lambda_0^{-1}\mu_0 + \Sigma^{-1}\bar{y}),$$
$$\Lambda_1^{-1} = \Lambda_0^{-1} + \Sigma^{-1}.$$

After calculating the posterior distributions for two treatment groups, the posterior distribution of incremental costs and incremental effects can be found by using the fact that the difference of two independent normally distributed variables is itself normally distributed, with the mean equal to the difference between the two means, and the (co)variance equal to the sum of the two (co)variances. And, with this multivariate posterior distribution, the ICER can be calculated with its 95% posterior interval (see Fieller, 1954).

Another way of describing the uncertainty surrounding the estimate of the ICER is to use the acceptability curve presented in chapter 2. Wakker & Klaassen (1995) pointed out that, in the frequentist framework, the acceptability curve can not be seen as a confidence interval in a formal statistical sense. Within the Bayesian framework however, the interpretation of the acceptability curve becomes straightforward: this curve gives for every limit (R) that society might place on the ICER the probability that the ICER is indeed below this limit. As Briggs & Fenn (1998) argue, the information contained in the acceptability curve will often be more in line with the information required by decision makers. It should be noted that an ICER is considered acceptable either if the incremental effects are larger than zero and the ICER is smaller than R or if the incremental costs and effects are both smaller than zero and the ICER is larger than R. This means that, assuming R=50,000, an ICER of 1,000/0.1=10,000 is acceptable, whereas an ICER of -1,000/ -0.1=10,000 is not acceptable.

An example

Here we will apply the ideas presented above to the analysis of the Benestent I and Benestent II studies. In both trials, the balance between costs and effectiveness of stent implantation in comparison with balloon angioplasty was assessed. The first study, carried out between 1990 and 1993 (see Serruys et al., 1994), was a landmark study which showed that stents were superior in terms of the prevention of revascularisations but that treatment was associated with increased risk of bleedings and increased hospital stay. In the second study, carried out between 1995 and 1997, a heparin-coated stent was used in combination with a better anticoagulation regimen, and as tested in a pilot study, this decreased the risk of bleeding and lead to no excess hospital stay (see Serruys et al., 1996; Serruys et al., 1998). Effectiveness was defined as 'event-free survival' after 1 year, where events were: death, myocardial infarction, cerebrovascular accident and revascularisations. Costs included the costs of the initial procedure, follow-up (including the costs of events), medication and diagnostic procedures.

For the analysis of the stent group in the Benestent I trial, it seems very reasonable to assume a non-informative prior, since little was known about the costs and effectiveness of stent procedures at that time. For the balloon angioplasty group one might have a less vague idea about costs and effectiveness prior to the analysis of the Benestent I, but here we will assume a non-informative prior for this group too.

Before the results of the Benestent II were available, it seemed reasonable to expect the results of the balloon angioplasty group to be similar to the results of the Benestent I. For the stent group however, an improvement in effectiveness might be anticipated. With this in mind, a non-informative prior for the stent group was chosen for the analysis of the Benestent II. For the balloon angioplasty group, three prior distributions were specified: a non-informative prior (thus disregarding all information from the Benestent I), a prior equal to the posterior of the Benestent I analysis, and a prior using only 50% of the information from the Benestent I trial (this latter prior is only used in the multivariate analysis).

All calculations and simulations were done using a spreadsheet program (Quattro Pro).

The classical approach

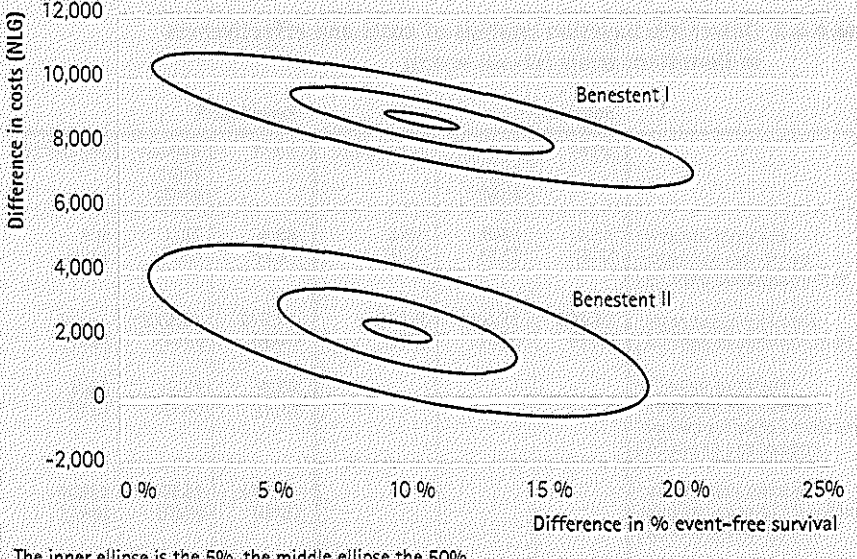
Table 4.1 shows the main summary statistics of the trials and figure 4.1 shows the estimated confidence ellipses for both studies. We immediately recognise that the outcomes will pose a problem in the interpretation of the results of the suggested Bayesian analysis. That is, contrary to our expectations we see substantial differences in the effectiveness of the balloon angioplasty: 67% in the Benestent I trial and 79% in the Benestent II trial. This might be a coincidence, but given the differences in timing it may not be. However, it was highly unexpected which is contrarily to the difference in the costs and the effectiveness in both the stent arms, which were expected.

table 4.1 Summary statistics for Benestent I & II (all costs in NLG)

	Benestent I			Benestent II		
	Stent	Balloon angio plasty	Difference	Stent	Balloon angio plasty	Difference
N	259	257		205	200	
% event-free survivors (SE)	77.6 (2.6)	66.9 (2.9)	10.7 (3.9)	88.7 (2.2)	79.0 (2.9)	9.7 (3.6)
average costs per patient (SE)	24,592 (553)	15,867 (633)	8,725 (840)	18,026 (832)	15,877 (746)	2,149 (1,118)
ICER			81,688			21,975

SE= Standard error
ICER: Incremental cost-effectiveness ratio

fig 4.1 Bivariate normal distribution of costs and effects comparing stent and balloon angioplasty in the Benestent I and II



The inner ellipse is the 5%, the middle ellipse the 50% and the outer ellipse the 95% confidence ellipse

Bayesian analysis of costs and effects

Given the lack of prior knowledge, non-informative priors were used for the Bayesian analysis of the Benestent I data. For costs we defined the priors for μ and σ^2 as N(10,100) and Inv-χ^2(0.01, 0.26), respectively. Both distributions have a large variance, ensuring that our arbitrary choices for μ_0 (10) en σ^2_0 (0.26) have little influence on the posteriors. The non-informative priors for effectiveness were defined as beta(1,1), which is equivalent to a uniform prior on [0,1].

In the method section, we described how the posterior distribution of σ^2 (equation 2) can be evaluated through simulation. Using this method, we find (with 2000 draws from the grid) a posterior mean of 0.269 and a posterior variance of 0.00055 for σ^2, when considering the balloon angioplasty group in the Benestent I data. Every draw of σ^2 is now entered into equation 1, and from this conditional distribution a posterior value of μ is drawn. From these draws, we find that the posterior mean of μ is 9.52 and the posterior variance is 0.001. Every pair of μ and σ^2 is then entered into equation 3 resulting in an estimate for the posterior expected costs in the balloon angioplasty group of 15,599, with a standard error of 546. The same can be done for the stent group. Then, by taking the difference between the costs in the two groups for every draw, we also find the distribution of the difference of the expected costs.

To calculate the posterior effectiveness, equation 4 can be used. This leads, for example in the balloon angioplasty group in the Benestent I trial where 172 patients out of 257 survived without an event, to a beta(173, 86) posterior distribution with a posterior mean of 0.668 and variance 0.00085. The distribution of the difference in

table 4.2 Univariate Bayesian analysis of costs and effectiveness in Benestent trials (all costs in NLG)

	Benestent I			Benestent II		
	Stent	Balloon angio plasty	Difference	Stent	Balloon angio plasty	Difference
Posterior mean % event-free survivors; non-informative prior (SE)	77.4 (2.6)	66.8 (2.9)	10.6 (3.9)	88.4 (2.2)	78.7 (2.9)	9.7 (3.7)
Posterior mean costs; non-informative prior (SE)	24,468 (475)	15,599 (546)	8,858 (711)	17,572 (546)	15,572 (614)	1,997 (831)
Posterior mean % event-free survivors; Benestent I prior for balloon angioplasty group (SE)				88.4 (2.2)	72.1 (2.1)	16.3 (3.0)
Posterior mean costs; Benestent I prior for balloon angioplasty group (SE)				17,572 (546)	15,639 (430)	1,934 (696)

SE = Standard error

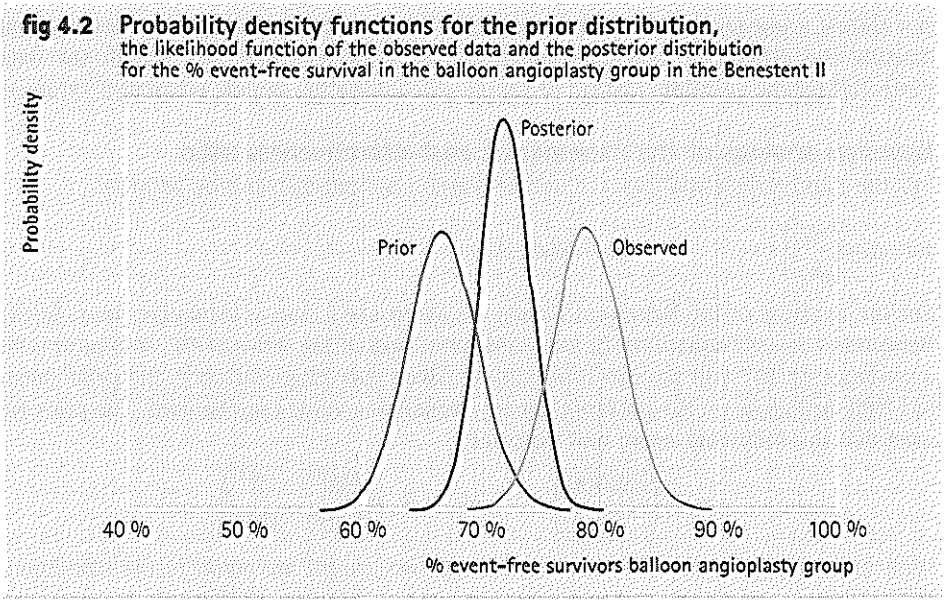


fig 4.2 **Probability density functions for the prior distribution,** the likelihood function of the observed data and the posterior distribution for the % event-free survival in the balloon angioplasty group in the Benestent II

effectiveness can be evaluated by drawing from both posterior beta distributions and calculating the difference for each draw. Table 4.2 presents the results for both arms and their difference.

For the analysis of the Benestent II, we first used the same non-informative priors as for the Benestent I (N(10,100); Inv-χ2(0.01, 0.26); beta(1,1)) and followed the same procedure as above to estimate the posterior costs and effectiveness. The results of this analysis are given in table 4.2. Additionally, we used the posterior results for the balloon angioplasty group from the Benestent I as prior for the same group in the Benestent II study while using a non-informative prior for the stent group. So, for the balloon angioplasty group the prior of μ is now N(9.52, 0.001). Since the posterior distribution of σ2 was very close to a normal distribution, the prior of σ2 is defined as N(0.269, 0.00055). Notice that in equation 2, the inverse gamma distribution is replaced by a normal distribution. The prior for effectiveness in the angioplasty group is the beta(173, 86) distribution, equal to the posterior of the Benestent I analysis. In figure 4.2, this prior distribution is shown together with the distribution of the observed data and the resulting posterior.

In the same way as before, we may simulate from the posteriors of both groups and calculate the differences in costs and effectiveness. Again, table 4.2 shows the results of the analysis with an informative prior for the angioplasty group.

If we compare the results of a Bayesian analysis with a non-informative prior (table 4.2) to the average costs per patient as we calculate them in the classical framework (table 4.1), we see that the difference is very small. This is what we expected to find, given the non-informative priors. The reason for this difference is that the data does

not follow a log normal distribution exactly. So it is not the Bayesian approach itself causing the difference, but the explicit use of a model (that does not fit the data perfectly) for the cost data instead of simply calculating the sample mean and variance.

Bayesian analysis of the incremental cost-effectiveness ratio

For the multivariate analysis, we use a bivariate normal distribution as prior

$$\text{for } \mu = \begin{pmatrix} \mu_C \\ \mu_E \end{pmatrix} : \qquad \mu \sim N(\mu_0, \Lambda_0) ,$$

For the analysis of the Benestent I data, we again use a non-informative prior. In the Benestent trial, the square of the standard errors was of the order of 10^6, so a prior with a variance of 10^{10} will be non-informative. Given this large variance, the prior mean has little influence on the posterior, so any reasonable value may be used. The prior we used had the following mean and variance:

$$\mu_0 = \begin{pmatrix} 20{,}000 \\ 0.9 \end{pmatrix} \quad \text{and} \quad \Lambda_0 = \begin{pmatrix} 10^{10} & 10^9 \\ 10^9 & 10^{10} \end{pmatrix} .$$

With these, the posterior distributions for the two groups are equal to the likelihood. Table 4.3 presents the posterior for both groups and for the differences in costs and effects.

For the analysis of the Benestent II, we again used a non-informative prior for the stent group. For the angioplasty group three different priors were defined: a non-informative prior, a prior equal to the posterior of the Benestent I and a prior giving a weight of 50% to the Benestent I data (this is done by multiplying the variance matrix of the Benestent I by 2).

table 4.3 Results of multivariate Bayesian analysis of costs and effectiveness in Benestent I

(presented are the five parameters of the bivariate normal distribution; all costs in NLG; effectiveness in proportion event-free survival)

	μ_C	μ_E	σ^2_C	σ^2_E	ρ
Prior stent	20,000	0.9	10^{10}	10^{10}	0.1
Likelihood/ posterior stent	24,593	0.776	305,759	0.00067	-0.80
Prior angioplasty	20,000	0.9	10^{10}	10^{10}	0.1
Likelihood/ posterior angioplasty	15,869	0.669	401,220	0.00086	-0.80
Posterior difference	8,724	0.107	706,979	0.00153	-0.80

Note: for the likelihood, sample means, their squared standard errors and the sample correlation are reported.

Table 4.4 presents the results of the multivariate Bayesian analysis of the Benestent II data. One might notice from this table that the posterior mean costs for priors 2 and 3 (NLG 15,462 and NLG 15,460) are not between the prior mean costs (NLG 15,869) and the mean observed costs (NLG 15,877) as one might expect intuitively. Furthermore, the posterior mean costs for prior 3 (NLG 15,460) does not lie between those associated with prior 1 (NLG 15,877) and 2 (NLG 15,462) which one might have expected. It should be noted that in general, when working with multivariate models in a Bayesian framework, intuitions developed in basic models may not always correspond with the outcomes in more complicated settings. In our situation, the posterior for the costs is influenced by the effects and vice versa. Carlin & Louis (1996) discuss several examples of non-intuitive posteriors.

table 4.4 Results of multivariate Bayesian analysis of costs and effectiveness in Benestent II

(presented are the five parameters of the bivariate normal distribution; all costs in NLG; effectiveness in proportion event-free survival)

	μ_C	μ_E	σ^2_C	σ^2_E	ρ
Prior stent	20,000	0.9	10^{10}	10^{10}	0.1
Likelihood/posterior stent	18,026	0.888	692,771	0.00049	-0.61
Likelihood angioplasty	15,877	0.79	557,199	0.00083	-0.65
Prior angioplasty1	20,000	0.9	10^{10}	10^{10}	0.1
Prior angioplasty 2	15,869	0.669	401,220	0.00086	-0.80
Prior angioplasty 3	15,869	0.669	802,440	0.00172	-0.80
Posterior angioplasty 1	15,877	0.79	557,199	0.00083	-0.65
Posterior angioplasty 2	15,462	0.73	223,918	0.00042	-0.75
Posterior angioplasty 3	15,460	0.75	315,547	0.00056	-0.72
Posterior difference 1	2,149	0.098	1,249,971	0.0013	-0.62
Posterior difference 2	2,564	0.158	916,641	0.00091	-0.64
Posterior difference 3	2,567	0.137	1,008,270	0.00105	-0.64

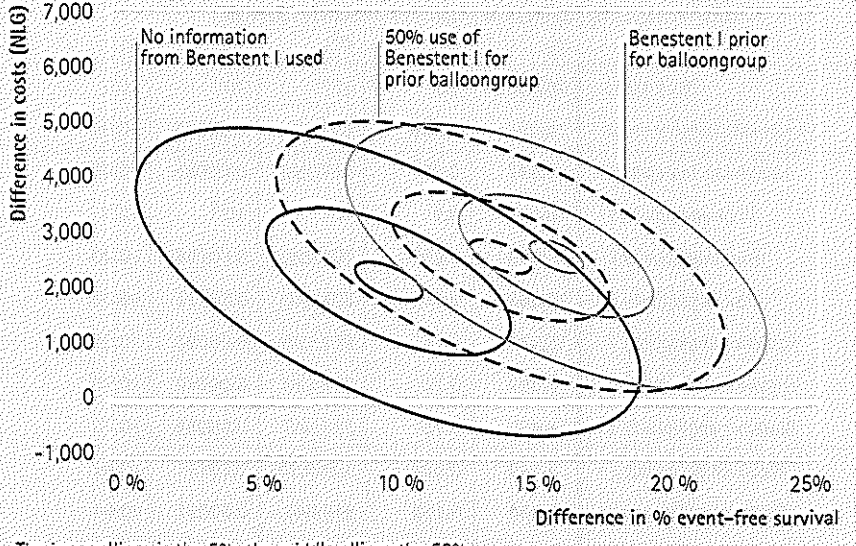
Prior angioplasty 1: non-informative prior; *Prior angioplasty 2*: posterior Benestent I is used as prior for Benestent II; *Prior angioplasty 3:* only 50% of the information from the Benestent I trial is used.

Note: for the likelihood, sample means, their squared standard errors and the sample correlation are reported.

Figure 4.3 shows the confidence ellipses for the three posterior distributions for the angioplasty group. From left to right the ellipses become smaller, indicating the decrease in uncertainty that is observed when some knowledge from the previous trial is included in the analysis.

For each of the posterior distributions of additional costs and effects, the ICER and its 95% posterior interval can be calculated (see Fieller, 1954). When no information from the Benestent I trial is used, the ICER is NLG 21,984 with a 95% posterior interval [2,265 , 90,612]. If the prior for the balloon angioplasty group is based on the posterior from the Benestent I, we find an ICER of NLG 16,269 with a 95% posterior interval of [5,157 , 35,413] and if only 50% of the information in the Benestent I on the balloon angioplasty group is used as prior, these numbers become NLG 18,642 and [5,247 , 45,576]. So, the point estimates vary substantially, and while NLG 16,269 per additional event-free survivor might be considered acceptable, NLG 21,984 might not be. Here, it may be more appropriate to consider the three acceptability curves

fig 4.3 **Bivariate normal density functions for the posterior distributions of the difference in costs and effects in the Benestent II,**
using three different prior distributions in the balloon angioplasty group



The inner ellipse is the 5%, the middle ellipse the 50% and the outer ellipse the 95% confidence ellipse.

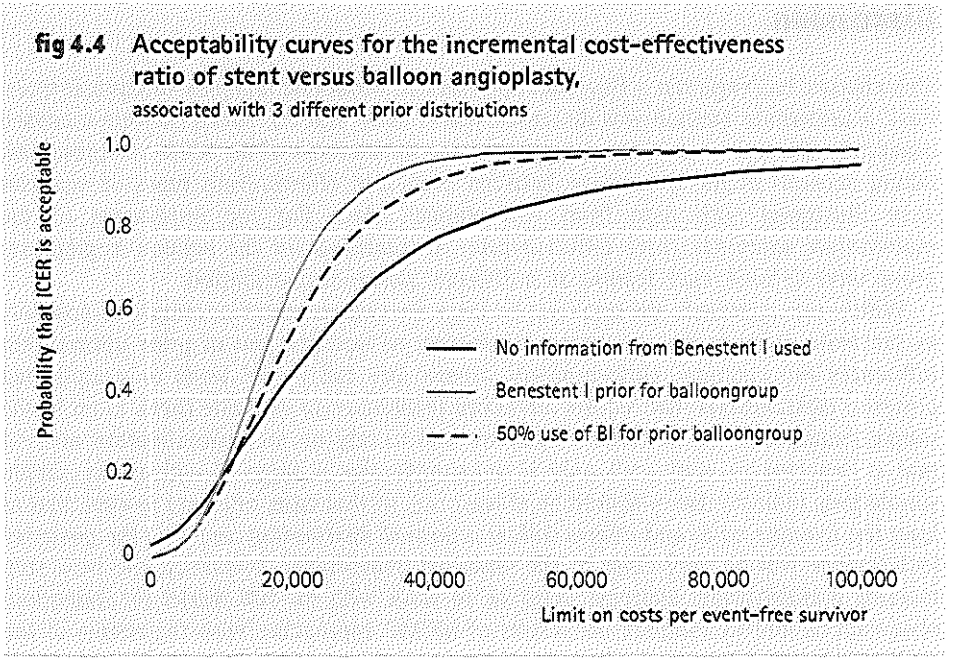


fig 4.4 Acceptability curves for the incremental cost-effectiveness ratio of stent versus balloon angioplasty, associated with 3 different prior distributions

associated with the three different prior distributions used as presented in figure 4.4. (Note also that all values have positive incremental effects.) The question whether the three different posteriors lead to different decisions depends on the limit that society puts on the ICER. If the limit is about NLG 10,000 per additional event-free survivor, the probabilities that the ICERs are acceptable are almost identical (20%). However, if the limit is about NLG 30,000, the probabilities differ substantially (65%, 80%, 90%), presumably leading to quite different decisions.

Discussion

This chapter has explored a Bayesian approach to economic analyses of clinical trials. First, a univariate method was discussed where costs and effectiveness were analysed separately. Second, a multivariate approach for the analysis of the ICER was presented, taking account of the covariance between costs and effects. Both types of analyses were illustrated with data from two clinical trials. For the illustration of the multivariate approach, three different prior distributions for the balloon angioplasty group were defined, each leading to a different posterior distribution.

Within our approach, the goal was to keep the computations as simple as possible, as the main purpose was to get an impression of how the use of prior information influences the outcome. Unfortunately, almost any full model leads to complicated posterior distributions that require advanced techniques to evaluate them. Even in the simple example of the Benestent study we had to rely on an approximation method that is expected to work well only if the sample size is large enough (some simple simulations indicate that n should be at least 50). In that case, one may assume that at least the average costs and effects follow a normal distribution and that the posterior (co)variance will be very close to the observed (co)variance. Naturally, the next step towards a multivariate approach must be to work with the full model for individual observations and to compare the results of MCMC approaches with the results of the current approximation method.

In general, if non-informative priors are used, the results of the Bayesian and frequentist analysis will be similar, but as was mentioned in the introduction, an important advantage of using Bayes is the straightforward interpretation of the confidence interval and the acceptability curve. The 95% confidence interval (or posterior interval) now tells us that the probability that the ICER is in this interval is 0.95, and the acceptability curve actually tells us for various limits on the ICER what the probability is that the ICER is acceptable.

As such, using non-informative priors has its advantages. However, in order to take full advantage of everything Bayesian analysis has to offer, one may want to use informative priors. Here, an example was used where data from an earlier trial were used as the prior for a later trial, an approach that leads to similar results as obtained by weighted pooling of the data in a frequentist meta-analysis. The example led to some striking results. At the start of the Benestent II trial, it was expected that the costs and effects in the balloon angioplasty arm would be similar to the results in the Benestent I trial. However, while costs were very similar, the effectiveness differed by more than 10%, suggesting that the two balloon angioplasty groups may not have been as comparable as thought beforehand.

This is why it is recommended, when using a Bayesian approach, to calculate a posterior distribution for a number of prior distributions. These priors should represent the various ideas that experts and decision makers may have on the subject. If the posterior distribution is more or less the same for each of these priors, the choice of prior is apparently not very important. If on the other hand the posterior distributions clearly differ, it might be an indication that more research is required.

One aspect of Bayesian analysis that has not been discussed in this chapter is model checking. As in frequentist analysis, it is important to check the adequacy of the fit of the model to the data. Gelman et al. (1995) devote a chapter of their book on methods for model checking. In frequentist analysis, assumptions about the underlying distribution of costs and effects are seldom made explicit, and consequently model checking is hardly ever seen as part of a study. Within the Bayesian framework, explicit assumptions are made about the distribution of costs and effects and about the prior distribution, and thus an explicit, critical evaluation of the model may always be needed. An advantage of a Bayesian approach is that testing hypotheses about different models (model refers here both to the likelihood function and the prior distribution) is fairly straightforward, with no requirement about hypotheses being nested and no limit on the number of hypotheses considered simultaneously.

In the analysis of the Benestent II, we found unexpected results in the balloon angioplasty group, i.e. an improved event-free survival compared to the Benestent I. This was probably the result of the availability of bailout stenting, permitting a more aggressive approach to balloon angioplasty. With this in mind, it might have been better to have used a completely different model for the data. The method used in this chapter, using the posterior of the Benestent I as a prior for the Benestent II, is comparable to a fixed-effect meta-analysis. Given the outcome of the Benestent II, a so-called hierarchical model might have been preferable. Such a model is similar to a random-effects model in frequentist meta-analysis. However, exploring such a model would require computations that are far more complex.

Finally, our exploration of Bayesian analysis in economic evaluations does not answer the question whether all health economists should become Bayesians. We expect that disadvantages of a technical nature, for example having to calculate the posterior distribution, may be overcome. However, we also expect the problems about formulating the prior distributions will never be resolved. What is known and what is not, and how to reflect this, may remain an art and may never become a science. On the one hand, one would like to define one's prior knowledge beforehand. However, one may then get the peculiar results as presented in our example. On the other hand, one might be tempted to redefine the prior after assessment of the posterior, so that the posterior will lead to a certain decision. This seems extremely opportunistic, and this may also not be what we are searching for.

In conclusion, we feel that the potentials of the Bayesian approach need to be investigated further. In Bayesian terms, we may conclude that adding our experience to our prior expectations has left us with a posterior distribution with a large variance.

Acknowledgements

The work of Ms. Al was funded through the Pfizer Health Economics Fellowship Program. The authors are grateful for the comments made on earlier versions of this chapter by Joseph C. Cappelleri, Ken Redekop, and Christine A. Getter.

CHAPTER 5

The cost-effectiveness of diclofenac plus misoprostol compared with diclofenac monotherapy in patients with rheumatoid arthritis

Summary

The objective of this study was to estimate the cost-effectiveness of treatment with a fixed dose combination of diclofenac and misoprostol compared with diclofenac monotherapy in the prevention of nonsteroidal anti-inflammatory (NSAID)-induced ulcers in rheumatoid arthritis (RA) patients.

A model was used to incorporate estimates of costs, incidence of ulcers and their complications, death rates and the efficacy of misoprostol. The costs per ulcer-free period gained and costs per additional survivor were calculated. Cost-effectiveness was calculated for the treatment of all RA patients, and of risk groups only.

The analysis showed that if 100 RA patients receive 3 months of treatment with diclofenac plus misoprostol, instead of diclofenac alone, this will lead to overall additional costs of NLG 773, while 0.82 symptomatic ulcers and 0.019 deaths will be prevented. If this treatment is given only to patients at high risk for NSAID-induced ulcer, cost-savings will occur instead of additional costs. Univariate sensitivity analysis showed that the outcomes are sensitive to changes in: the percentage of ulcers treated in the ambulatory setting, the price difference between diclofenac and the fixed dose diclofenac/misoprostol combination, the percentage of ulcers with complications, and the efficacy of misoprostol.

In conclusion it can be stated that treatment with diclofenac/misoprostol is cost-saving in RA patients at high risk for NSAID-induced ulcers. For RA patients in general, the cost-effectiveness of this intervention compares favourably with that of other preventive policies.

Introduction

Rheumatoid arthritis (RA) is a chronic, multisystem disease that occurs in about 1 to 2% of the Dutch population (see Valkenburg, 1980). Women are affected about 2.5 times as frequently as men, and the total direct medical costs related to rheumatic diseases amount to 160 million Dutch Guilders (NLG) per year (see e.g. Valkenburg, 1980; Koopmanschap et al., 1991). Although pain, swelling and erosions of multiple joints are the most common characteristics of established RA, extra-articular features may also occur, for example in the lungs, heart, kidney and skin.

Nonsteroidal anti-inflammatory drugs (NSAID) play an important role in the treatment of RA, because of their anti-inflammatory effect and ability to relieve pain. Unfortunately, NSAID treatment is associated with serious adverse effects such as gastritis, peptic ulceration (complicated by bleeding or perforation) and platelet dysfunction, which may even be fatal. Much research has been done to find drugs that may prevent these gastrointestinal (GI) adverse effects. In 1988, Graham et al. found that NSAID-treated osteoarthritis patients who were receiving misoprostol developed significantly fewer gastric ulcers compared with controls.

Several cost-effectiveness analyses of prophylactic treatment with misoprostol were based on this study (e.g. Hillman & Bloom, 1989; Edelson et al., 1990; Knill-Jones et al., 1990; Drummond et al., 1992; Jönsson & Haglund, 1992; Gabriel et al., 1993). Some of these analyses showed that cost-savings were associated with misoprostol prophylaxis, whereas others did not. A critical appraisal of these studies was published by Stucki et al. (1994), addressing the causes for those differences. However, since then, the results of trials investigating the use of misoprostol in RA patients, trials including patients with both gastric and duodenal ulcer as endpoints, and trials in which misoprostol was administered in a fixed dose combination with the NSAID diclofenac have been published (see Bolten et al., 1992; Verdickt et al., 1992; Melo Gomes et al., 1993; Graham et al., 1993; Barradell et al., 1993). Also, the effects of misoprostol in serious GI complications, such as bleeding ulcers, has been investigated by Silverstein et al. (1995). With this new information, the balance between costs and effects of misoprostol prophylaxis may be re-evaluated (see Pouvourville, 1995).

The objective of this study was to compare the costs and effects of a fixed dose combination (fdc) of diclofenac and misoprostol with those of diclofenac alone in the treatment of RA patients.

Methods

The current analysis evaluated the cost-effectiveness of an fdc of misoprostol and diclofenac in comparison with diclofenac monotherapy in patients with RA, for the Dutch situation. An attempt was made to calculate medical costs from a societal viewpoint, i.e. where possible, real costs were used. Nonmedical and indirect costs were not taken into account. The analysis is mainly based on two recently published studies

of the clinical effectiveness of misoprostol in patients with RA.

The first study compared an fdc of diclofenac/misoprostol with diclofenac alone in the prevention of endoscopically diagnosed ulcers, in treatment groups of about 170 patients (see Verdickt et al., 1992). The trial lasted for 12 weeks. At the end of this period, all patients had an endoscopy to investigate the gastric and duodenal mucosae for lesions and ulcers. In the fdc group, 6 out of 137 (4.4%) patients had developed an ulcer, compared with 17 out of 153 (11.1%) patients in the diclofenac group. 11% of patients in the fdc group withdrew because of adverse events, compared with 8.5% of those in the diclofenac group.

In the second study, patients receiving continuous treatment for RA symptoms with NSAIDs were either assigned to misoprostol or to placebo to study the effectiveness of misoprostol in the prevention of serious GI complications (perforated or bleeding ulcer, endoscopically proven, see Silverstein et al., 1995). Both treatment groups consisted of about 4400 patients and the duration of the trial was 6 months. In the NSAID plus misoprostol group, 16 out of 4404 (0.36%) patients had a perforated or bleeding ulcer (with a proven lesion), while in the NSAID plus placebo group this proportion was 33 out of 4438 (0.74%) patients. Premature withdrawal because of adverse effects occurred in 27.5% of the patients in the misoprostol group and 20% of those in the placebo group.

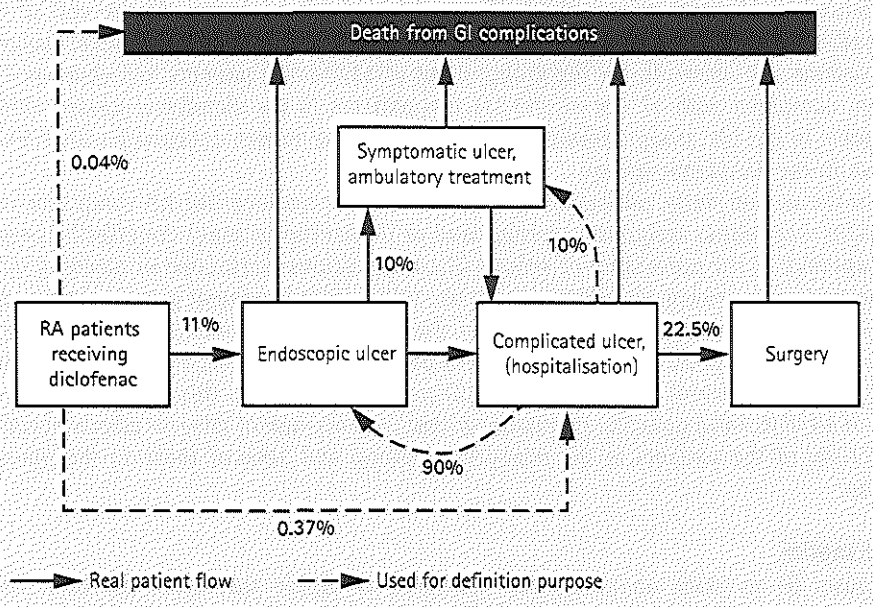
The model

Because: (I) no data on costs were gathered alongside the trials; (II) both trials addressed different aspects of NSAID-induced GI complications; and (III) both trials did not completely reflect common medical practice in The Netherlands, we constructed a mathematical model to convene data from various sources. Using the model, the costs and effects of several scenarios were compared.

We defined two baseline scenarios. In the first scenario, 100 RA patients receive diclofenac [50 mg, twice daily (bid) or 3 times daily (tid)] for 3 months. In the second scenario, they receive fdc diclofenac/misoprostol (50 mg + 200 μg, bid or tid). In several alternative scenarios, only RA patients with an elevated risk for ulcers are treated with fdc diclofenac/misoprostol. We used the results of a study by Fries et al. (1991) to identify the most important risk factors, namely prednisone use, history of GI events, age > 75 years, or a combination of these.

The patient flow for RA patients receiving diclofenac treatment is represented in figure 5.1, from left to right. The first step in the model is the development of an endoscopic ulcer (i.e. an ulcer verifiable with gastroscopy). Only some of these ulcers cause symptoms, so a separate stage was defined for patients who require ambulant treatment for a symptomatic ulcer. Both asymptomatic and symptomatic ulcers may lead to serious complications (bleeding or perforation) requiring admission to a hospital, and sometimes even surgical treatment. In addition, a small number of patients may die as a result of GI-complications.

fig 5.1 Flowchart for adverse gastrointestinal (GI) events in patients with rheumatoid arthritis (RA) taking diclofenac



For those patients receiving fdc diclofenac/misoprostol, the model is basically identical except for the addition of a stage in which patients may stop treatment as a result of misoprostol-related adverse effects (mainly diarrhoea). Since RA patients need NSAID treatment for pain relief, it is assumed they either fully comply to their fdc treatment or switch to diclofenac monotherapy after discontinuation of their fdc treatment. Because it has been shown that the median time to the onset of misoprostol adverse effects such as diarrhoea and abdominal pain is 2 to 3 days, it was assumed that patients who discontinue fdc treatment do so in the first 2 weeks (see Geis, 1992).

The cost-effectiveness of prophylactic treatment with fdc diclofenac/misoprostol was defined both as additional costs per symptomatic ulcer-free period gained, and as additional costs per additional survivor. Symptomatic ulcers were defined as ambulatory treated ulcers plus asymptomatic ulcers leading to complications.

Epidemiology

Table 5.1 presents the estimates of the probabilities included in the model. Probabilities for the occurrence of endoscopic ulcers and complicated ulcers were directly obtained from Verdickt et al. (1992) and Silverstein et al. (1995), respectively. Note that both probabilities are defined as a percentage of all 100 RA patients (see figure 5.1).

table 5.1 Probability estimates used in the model (3-month probabilities)

Variable (treatment)	Baseline estimate (%)	Range (%)	Distribution
Endoscopic ulcer (diclofenac)	11.1	6.1 – 16.1	Normal
Endoscopic ulcer (fdc)	4.4	0.95 – 7.8	Normal
Probability endoscopic ulcer requires ambulatory treatment	10	1.2 – 44.0	Log normal
Complications (diclofenac)	0.37	0.25 – 0.50	Normal
Complications (fdc)	0.18	0.09 – 0.27	Normal
Probability complication is preceded by ambulatory treatment	10	5 – 15	Uniform
Death from GI complications (diclofenac)	0.04	0.01 – 0.06	Normal
Death from GI complications (fdc)	0.02		
Surgery	22.5	15 – 30	Normal
Withdrawal from treatment because of adverse effects (fdc)	4.2	1.4 – 7.0	Normal
Proportion GU : DU	1.2 : 1	1.0:1 – 1.4:1	Normal

fdc = fixed dose combination; GI = gastrointestinal; GU = gastric ulcer; DU = duodenal ulcer

The percentage of patients requiring ambulatory treatment for a symptomatic ulcer was also derived from Silverstein et al. (1955). In that study, 86 ulcers were reported in the group that did not receive prophylaxis, resulting in an 1% incidence of symptomatic ulcers requiring treatment within 3 months of NSAID therapy. Combining this 1% incidence of symptomatic ulcers with the 11% incidence of endoscopically proven ulcers found by Verdickt et al. (1992) we estimated that 10% of endoscopic ulcers cause symptoms requiring ambulatory treatment.

The 1% incidence of symptomatic ulcers from Silverstein et al. (1995) is confirmed by two trials that studied the GI adverse effects of NSAID treatment, in which incidences of 0.5% and 1% were found for ulcers and bleedings (see Piercksky et al., 1989; Eversmeyer et al., 1993). These studies were not endoscopy controlled, so these percentages represent spontaneous reporting of symptoms by patients. In other studies, much higher percentages (44% and 30%) of the ulcers were symptomatic (see Larkai et al., 1987; Farah et al., 1988). In these studies, a random selection of RA patients underwent gastroscopy and were asked whether they had experienced GI symptoms. However, it is not to be expected that all patients who reported GI symptoms visited a physician to seek treatment for these symptoms. Since we were only interested in ulcers that require ambulatory treatment, the higher percentages were not used in the baseline scenario, but were included in a separate sensitivity analysis.

The proportion of complicated ulcers that are preceded by ambulatory treatment was derived from two studies in which patients admitted to hospital for complications were asked about previous symptoms and about medication for ulcer treatment (see Armstrong & Blower, 1987; Jorde et al., 1988). With these findings, it was estimated that approximately 10% of the patients who were admitted to hospital for complications were started on antiulcer medication for an endoscopically proven ulcer before their admission.

Using Dutch nationwide hospital registration data (see SIG Zorginformatie, 1993), we determined the number of admissions for complicated ulcers in The Netherlands in 1991. This was related to the number of surgical procedures for complicated ulcers in 1989, after adjusting this number to 1991 values according to the linear time trend between 1985 and 1989 (see De Nederlandse Lever Darm Stichting, 1992). Thus, we calculated that 22.5% of patients with complicated ulcers require surgery.

Since treatment of ulcers varies with type, an assumption about the relative proportions of gastric and duodenal ulcers had to be made. In the baseline scenarios, it was assumed that diclofenac treatment causes 20% more gastric than duodenal ulcers (i.e. 6 gastric ulcers for every 5 duodenal ulcers; see e.g. Mellem et al., 1985; Farah et al., 1988; McCarthy, 1989).

Several studies have reported the percentage of patients taking fdc diclofenac/misoprostol that discontinued the study medication because of adverse effects. The overall percentage of patients withdrawing was estimated to be 4.2% (see Geis, 1992).

To estimate the proportion of patients who die as a result of GI complications we used the results of a large prospective study by Fries (1991) of 2,700 RA patients (average follow-up 3.5 years) that reported an excess GI death rate of 0.15% per year.

The proportional decrease in complicated ulcers among patients treated with fdc compared with patients taking diclofenac was used as a proxy for the proportional decrease in the death rate of patients receiving fdc.

Based on the same study, Fries et al. (1991) developed a simple scoring system (by means of multivariate analysis of risk factors) enabling the direct calculation of the risk of ulcer complications for RA patients. With this model, we calculated the probability of complications for various risk groups (table 5.2).

table 5.2 Estimates of the risk of developing a complicated ulcer in the various risk groups (3-month probabilities)

Risk groups	Probability (%)
Prednisone	0.625
History GI events	0.7
Age 75	0.65
Age 80	0.725
Age 85	0.775
Age 75 & prednisone	0.825
Age 75 & prednisone & history GI events	1.075

GI = gastrointestinal

Costs

Combining information from the literature (see Van Berge Henegouwen & Bijlsma, 1991; McGuigan, 1994) and information about treatment policy in two Dutch general hospitals, we developed treatment profiles for an 'average' patient. For ambulatory treatment, it was assumed that patients are referred for gastroscopy by either their general practitioner (GP) or rheumatologist. During the gastroscopy, biopsies are done (gastric ulcer is checked for both helicobacter pylori and malignancy, duodenal ulcer for helicobacter pylori only). If a gastric ulcer is found, the patient has to return for a control gastroscopy after 6 or 8 weeks, to check whether the ulcer has healed. Of the patients with an ulcer, 43% receive treatment with ranitidine (for 8 weeks), 38.6% with omeprazol (gastric ulcer 6 weeks, duodenal 3 weeks) and 18.4% with cimetidine (for 8 weeks) [data on file, Searle]. After the initial gastroscopy, patients with a duodenal ulcer visit their GP or rheumatologist once or twice, and patients with a gastric ulcer visit a gastro-enterologist once on an outpatient basis. Patients admitted to hospital for a bleeding or perforated ulcer have one gastroscopy on admission and one before discharge. Furthermore, they receive the same medication as ambulatory patients.

For gastroscopy and in-hospital days real costs were calculated; for consultations and surgery, tariffs were used (see COTG (Central Organisation on Tariffs in Health

Care), 1995). Medication costs were based on 1995 prices. (Exchange rate at time of study: NLG 1 = 0.60 $US.) Only direct medical costs were taken into account. Table 5.3 shows the various cost estimates. Since only 3 months of treatment were evaluated, no discount factor was applied to the costs.

table 5.3 Cost estimates used in the model (baseline estimates and the range used in the multivariate sensitivity analysis)
Costs are in 1995 Dutch Guilders (NLG; NLG1=$US0.60, 1995) and are for 3 months' treatment. The drug costs are based on twice daily dosage

	Baseline cost estimate (NLG)	Range
Diclofenac	38	10%
Fdc diclofenac/misoprostol	47	10%
Gastroscopy	246	10%
Biopsy	90	10%
Consultations GU	130	20%
Consultations DU	90	20%
Medical treatment GU	224	20%
Medical treatment DU	183	20%
Cost of hospitalisation GU	9,280	20%
Cost of hospitalisation DU	8,120	20%
Surgery GU	1,500	10%
Surgery DU	2,150	10%

fdc = fixed dose combination; GU = gastric ulcer; DU = duodenal ulcer

Sensitivity analysis

The sensitivity of the results to variation in the input variables was studied both univariately and multivariately. The univariate analysis shows which variables influence the outcome significantly. This analysis was carried out by subsequently increasing and decreasing all variables with 20%. If changing a variable with 20% altered the outcomes by 10% or less, the outcome was considered not sensitive to that variable. Changes in the outcomes of more than 10% were reported. The multivariate analysis determines the most probable outcome, and also the 'worst' and the 'best' possible outcomes. For this purpose, a probability distribution was defined for each variable. Where possible, distributions were based on trial data. For cost estimates, a range of either ±10% or ±20% (depending on the uncertainty regarding the estimate) was used. For the percentage of ulcers that were treated ambulatory, a log normal distribution was defined with a mean of 10% and an upper boundary of 44%.

Results

The model showed that, if 100 RA patients receive 3 months of treatment with diclofenac monotherapy, 1.45 symptomatic ulcers will occur and 0.0375 patients will die as a direct result of GI complications. The total costs for this treatment policy are NLG 19,825 per 100 patients. If patients are treated for 3 months with fdc diclofenac/misoprostol, it was estimated that the number of symptomatic ulcers will decrease to 0.63 and the number of deaths to 0.0189, whereas costs will increase to NLG 20,598. Comparing the two scenarios, we found that treatment with fdc diclofenac/misoprostol costs NLG 773 extra, but prevents 0.82 symptomatic ulcers, 0.18 hospital admissions for complicated ulcers, 0.04 operations on bleeding or perforated ulcers and 0.019 deaths (all per 100 patients). Thus, the additional costs per symptomatic ulcer-free period gained amount to NLG 949 and the additional costs per additional survivor to NLG 41,790.

Cost-effectiveness ratios of other prophylactic treatment policies are mainly published as costs per life-year gained. Therefore, we made a rough calculation of the number of life years that may be gained by fdc therapy compared with diclofenac monotherapy. It was assumed that RA patients with GI events have an average age of 65 (see Fries, 1991). In The Netherlands, the life expectancy at the age of 65 is 18 years for women and 15 years for men, and the median life expectancy of RA patients is approximately 3 years less in women and 7 years less in men (see Netherlands Central Bureau of Statistics, 1993; Vandenbrouke et al., 1984). Because 71% of RA patients is female (see Valkenburg, 1980), this leads to an expected number of life-years gained per death prevented of 13 without discounting, and 10 after discounting by 5%. When

table 5.4 **Costs and effects of treatment with a fixed dose combination of diclofenac and misoprostol compared with diclofenac treatment for high risk groups**
Costs are in 1995 Dutch Guilders (NLG; NLG1 = $US0.60, 1995) and are for 3 months' treatment of 100 patients

Risk groups	Additional costs (NLG)	Symptomatic ulcer-free periods gained	Additional survivors
Prednisone	- 801	1.37	0.031
History GI events	- 1,267	1.53	0.035
Age 75	- 957	1.42	0.032
Age 80	- 1,423	1.59	0.036
Age 85	- 1,733	1.70	0.038
Age 75 & prednisone	- 2,044	1.81	0.041
Age 75 & prednisone & history GI events	- 3,598	2.36	0.054

combined with the costs per additional survivor, an estimate of NLG 4,179 per life-year gained results.

If treatment with fdc diclofenac/misoprostol is limited to patients who have a high risk for NSAID-induced ulcers, savings will occur instead of additional costs. Table 5.4 gives the results for the various risk groups. Older patients who take prednisone and have a history of GI events showed the highest savings and the greatest effects. If those patients received diclofenac, it was estimated that 4.18% would experience a symptomatic ulcer and 0.11% would die as a result. If 3 months of treatment with fdc diclofenac/misoprostol is given instead, the percentage of symptomatic ulcers decreases to 1.83 and only 0.055% of the patients will die from a complicated ulcer.

Univariate sensitivity analysis

Table 5.5 gives an overview of those variables for which the results of the cost-effectiveness analysis are most sensitive. Clearly, changes in the percentage of complicated ulcers (i.e. an ulcer requiring hospitalisation) have more influence on the outcome than changes in the percentage of endoscopic ulcers. This is easily explained by the fact that only 10% of patients with a endoscopic ulcer seek (ambulatory) treatment, whereas all complicated ulcers require treatment, at much higher costs.

Regarding the cost estimates, the results were most sensitive to changes in the costs of diclofenac and fdc diclofenac/misoprostol. The impact of changing the treatment costs for complicated ulcers by ±20% is less remarkable, but still important: both costs per symptomatic ulcer-free period gained and costs per additional survivor change by 45%.

Two variables were studied more closely: the percentage of endoscopic ulcers treated ambulatory, and the price difference between diclofenac and fdc diclofenac/misoprostol. For both variables, a 'break-even point' was calculated, i.e. the value at which the costs associated with fdc treatment will be equal to those associated with diclofenac monotherapy. For the percentage of endoscopic ulcers treated ambulatory this value is 25%, and for the price difference between diclofenac and fdc diclofenac/misoprostol it is NLG 6.50 (price per month at the standard daily dosage).

table 5.5 Results of the univariate sensitivity analysis

	Baseline value	Additional costs (NLG)	Symptomatic ulcer-free periods gained	Additional survivors
Baseline scenario		773	0.82	0.0185
% Endoscopic ulcers diclofenac 13.3%	11.1%	604	1.03	0.0185
% Endoscopic ulcers diclofenac 8.9%		942	0.60	0.0185
Efficacy of fdc on endoscopic ulcers 70%	60%	694	0.92	0.0185
Efficacy of fdc on endoscopic ulcers 50%		862	0.70	0.0185
% Complicated ulcers diclofenac .44%	0.37%	70	0.88	0.0210
% Complicated ulcers diclofenac .3%		1,477	0.75	0.0141
Efficacy of fdc on complicated ulcers 60%	50%	461	0.84	0.0217
Efficacy of fdc on complicated ulcers 40%		1,165	0.78	0.0145
% Death from complicated ulcers +20%	0.04%	773	0.82	0.0222
% Death from complicated ulcers -20%		773	0.82	0.0147
Costs diclofenac +20%	NLG 38	-2,180	0.815	0.0185
Costs diclofenac -20%		3,727	0.815	0.0185
Costs fdc diclofenac/misoprostol +20%	NLG 47	4,343	0.815	0.0185
Costs fdc diclofenac/misoprostol -20%		-2,797	0.815	0.0185
Costs of ambulatory treatment +20%	NLG 788	671	0.815	0.0185
Costs of ambulatory treatment -20%		876	0.815	0.0185
Costs of complicated ulcers +20%	NLG 9,404	428	0.815	0.0185
Costs of complicated ulcers -20%		1,118	0.815	0.0185

fdc = fixed dose combination

Multivariate sensitivity analysis

A multivariate sensitivity analysis was done by running the model 5000 times, each time with values for the parameters that were randomly chosen from the probability distributions specified in table 5.1. Table 5.6 presents the mean, median and a 95% range for costs and effects. The frequency distributions of the additional costs and the additional survivors are approximately normal, whereas the distribution of the symptomatic ulcer-free periods gained is skewed to the right. Because of this skewness, the upper limit of the 95% range is relatively high. If we consider only the 50% range, we find that half of the outcomes are between 0.35 and 1.00.

With the outcomes of the simulation, a so-called cost-effectiveness acceptability curve was constructed (see chapter 2). Such a curve shows, for every limit that society

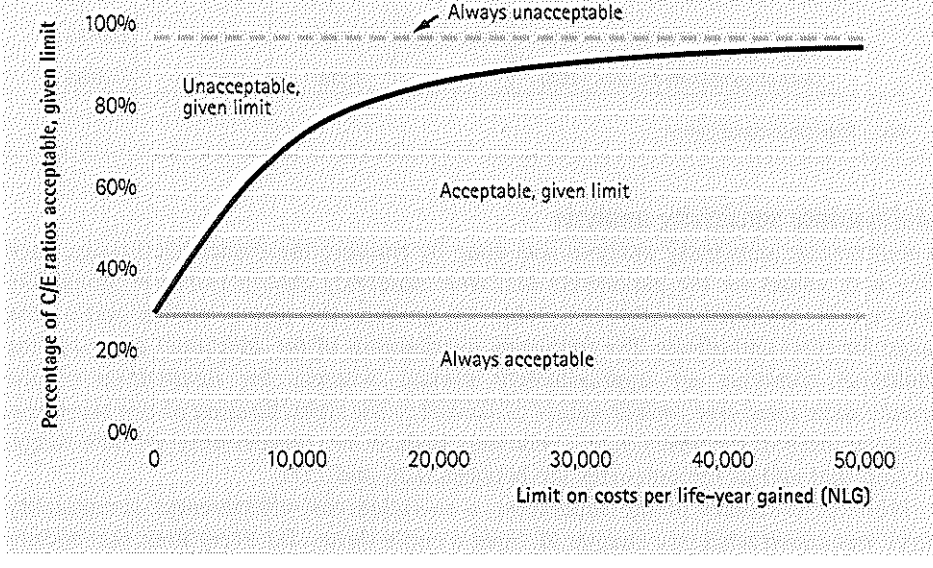
table 5.6 Results of the multivariate sensitivity analysis

	Mean	Median	95% range	
Additional costs (NLG)	778	801	-2,144	to 3,542
Symptomatic ulcer-free periods gained	0.84	0.57	0.15	to 3.10
Additional survivors	0.0180	0.0171	0.0018	to 0.0366
Costs per symp. ulcer-free period gained (NLG)	2,539	1,183	-2,665	to 16,833
Costs per additional survivor (NLG)	74,838	42,788	-171,689	to 568,506

may put on additional costs per additional survivor, the percentage of simulated cost-effectiveness ratios that are below that limit. In interpreting this curve, it should be realised that savings combined with lives gained are always deemed acceptable while additional costs combined with lives lost are always deemed unacceptable. With this method, in our model 1% of the simulation outcomes were unacceptable and 29% acceptable, regardless of the limit used. For the remaining 70% of the outcomes, acceptability depended on the limit chosen by society. Figure 5.2 presents the relationship between the limit for the costs per life-year gained and the percentage of our simulation outcomes that are deemed acceptable.

fig 5.2 Cost-effectiveness acceptability curve

This shows the relationship between the limit placed by society on costs per life-year gained and the percentage of simulations that yield a result that is deemed acceptable



Discussion

The current cost-effectiveness analysis shows that preventive treatment with fdc diclofenac/misoprostol in all RA patients will generate symptomatic ulcer-free periods and life years in return for extra costs. However, if this preventive treatment is given only to risk groups, such as patients treated with prednisone and patients with a history of GI events, it will also save costs.

Of course, this conclusion depends on the assumptions incorporated in this model. Therefore, an extensive sensitivity analysis was performed. The univariate analysis demonstrated that, in particular, the results are influenced by the costs of diclofenac and fdc diclofenac/misoprostol, the probability of endoscopically proven ulcers and complicated ulcers, and the percentage of ulcers that require ambulatory treatment. The multivariate analysis showed that the median costs per additional survivor amount to NLG 42,788, but that these costs may vary between -NLG 172,000 and NLG 569,000 (table 5.6).

Our estimates on the number of complicated ulcers and the efficacy of misoprostol in preventing these ulcers were obtained from a study by Silverstein et al. (1995). It should be noted that in this study the average daily dosage of misoprostol was 650 µg, whereas in our model an average dosage of 500 µg was used. The latter equals the average recommended daily dosage for fdc diclofenac/misoprostol of 50 mg/200 µg bid or tid in The Netherlands (see Dutch Health Insurance Funds Council, 1995). Since the effectiveness of misoprostol is known to be dosage-dependant, the percentage of complications in the fdc group may well be an underestimate (see Graham et al., 1988). On the other hand, in the trial by Silverstein et al., various NSAIDs were used, some of them known to cause more, others to cause fewer, ulcers than diclofenac. Clearly, every adjustment of the estimate would be arbitrary, so the results as reported by Silverstein et al. were used without modification.

To ensure comparability of our results with those of other cost-effectiveness studies, we also included costs per life-year gained as a measure of effectiveness. As already pointed out by Stucki et al. (1994), it would have been even more appropriate to use quality-adjusted life-years. Quality of life may be affected by misoprostol in two different ways. On the one hand, patients who have an ulcer may experience a temporary decrease in their quality of life, particularly patients with a bleeding ulcer. Misoprostol will prevent part of this decrease. On the other hand, some patients taking misoprostol will develop diarrhoea, which may also affect their quality of life. Since this diarrhoea is often self-limiting, it is most likely that the decrease in quality of life is transient in nature. However, without additional research it is hardly possible to quantify both effects, and therefore we decided not to estimate quality-adjusted life-years.

Several studies on prophylactic treatment with misoprostol have been published. However, most of these studies evaluate the cost-effectiveness of misoprostol prophylaxis in osteoarthritis patients with abdominal pain (see Hillman & Bloom, 1989; Knill-Jones et al., 1990; Jönsson & Haglund, 1992; Gabriel et al., 1993). In a study by Edelson et al. (1990) prophylaxis for RA patients was considered. Comparing our results with those of Edelson et al. is not straightforward, since they measured costs per episode of GI bleeding prevented as the main outcome. Furthermore, in our study, misoprostol was administered as an fdc with the NSAID, as opposed to separately, which leads to a better compliance. What can be said is that the cost estimates for the treatment of ulcers and for misoprostol are much higher in the study by Edelson et al. than in our study, which may be explained by the differences in the structure of the health care systems (see OECD, 1987).

If the costs per life-year gained (NLG 4,179) are compared with those of other preventive policies, treatment with fdc diclofenac/misoprostol seems favourable. For instance, the costs of screening for breast cancer were estimated at NLG 7,650 per life-year gained (see De Koning et al., 1991), screening for cervix carcinoma at NLG 24,000 per life-year gained (see Van Ballegooijen et al., 1992), and cholesterol-lowering therapy for men with cholesterol level above 8 mmol/l at about NLG 30,000 per life-year gained (see Martens et al., 1989).

At present, the Dutch Health Insurance Funds Council approves of the use of fdc for long term NSAID users at high risk for GI complications. They currently define high risk patients as the elderly and patients with a history of proven GI ulcers. From our analysis, it appears that fdc diclofenac/misoprostol is not only cost-saving for these patients, but also for patients using prednisone. The choice of whether or not to approve of fdc for other RA patients taking NSAIDs long term remains arbitrary. Payers will probably take into account not only the cost-effectiveness ratio - which seems favourable compared with other (approved) preventive therapies - but also the nationwide costs involved. Whereas the additional costs per patient per year are only minor (NLG 31), the relatively high prevalence of RA (1 to 2%) could lead to an annual investment of several million NLG.

Even though the results clearly indicate that treatment with fdc diclofenac/misoprostol in RA patients at high risk for GI adverse reactions leads to both health gains and economic benefits, physicians may be reluctant to adopt such a treatment policy. This reluctance may be attributable to a dislike of prescribing combination therapy, the feeling that fdc treatment does not leave any space for experimentation with combinations of dosages, and the feeling that it is not rational to adopt a prophylactic treatment against the adverse effects of NSAIDs when the treatment itself may cause unpleasant adverse effects. This may be outweighed by the much lower price of misoprostol in the fixed dose combination (NLG 9 per month in an fdc, compared with NLG 64 per month when prescribed separately) and the convenience of an fdc for patients who are obliged to take many different drugs each day.

Based on currently available data on the effectiveness of misoprostol, epidemiology of ulcers and costs, it may be concluded that treatment with fdc diclofenac/misoprostol is cost-saving for RA patients at high risk for NSAID-induced GI adverse effects. For RA patients in general, the cost-effectiveness of the treatment policy compares favourably with that of other preventive policies.

Acknowledgements

The study was supported by an unrestricted grant from Searle. We gratefully acknowledge the helpful comments of M. Janssen, rheumatologist and H.P.M. Festen, gastro-enterologist.

CHAPTER 6

Cost-effectiveness of lung transplantation in The Netherlands: a scenario analysis

Summary

The objective of the study was to calculate the cost-effectiveness of scenarios concerning lung transplantation in The Netherlands. A microsimulation model was developed, predicting survival, quality of life, and costs with and without transplantation program, based on data of the Dutch lung transplantation program of 1990-1995. Included were 425 patients referred for lung transplantation, of whom 57 underwent transplantation.

The analysis of the baseline scenario showed that the estimated costs per life-year gained are NLG 194,000, and the costs per quality-adjusted life-year (QALY) gained NLG 167,000. Restricting patient inflow ("policy scenario") lowers the costs per life-year gained: NLG 172,000 (costs per QALY gained: NLG 144,000). The supply of more donor lungs could reduce the costs per life-year gained to NLG 159,000 (NLG 135,000 per QALY gained, 1NLG = 0.6US$, based on the exchange rate at the time of the study).

In conclusion it may be stated that lung transplantation is an expensive, but effective intervention: survival and quality of life improve substantially after transplantation. The costs per life-year gained are relatively high, compared with other interventions and other types of transplantation. Restricting the patient inflow and/or raising donor supply improves cost-effectiveness to some degree. Limiting the extent of inpatient screening or lower future costs of immunosuppressives may slightly improve the cost-effectiveness of the program.

Introduction

Lung transplantation is a fast-growing and expensive medical intervention. Worldwide, about 6,000 lung transplants have been performed in more than 100 centres (see St. Louis International Lung Transplant Registry, January 1997). However, reliable information on cost-effectiveness of lung transplantation programs is lacking. To our knowledge, only one retrospective pilot study on this topic was published thus far, by Ramsey et al. (1995a). Its small sample size and limited cost-analysis prohibited firm conclusions. In this chapter, we will estimate the cost-effectiveness of various scenarios of future lung transplantation, based on detailed data from a large technology assessment of the Dutch lung transplantation program, as performed during 1990 to 1995 (see TenVergert et al., in press). This study was initiated by the Dutch National Health Insurance Board, to support public reimbursement decisions.

The scenario analysis describes future transplantation programs, operational for 15 years. This simulation period is necessary to reach a stable number of patients on the waiting list, which has not yet been established during the observation period. Societal costs, survival and quality of life are followed up to 40 years, comparing the situation with and without a lung transplantation program.

Several scenarios will be presented: a baseline scenario, a prolongation of the current program; a policy scenario, restricting the inflow of patients on the waiting list; and a donor scenario, assuming a larger supply of donor lungs.

Material and methods

From November 1990 until April 1995, data were gathered on all patients who entered the Dutch lung transplantation program. Patients are eligible for the program if they have irreversible, progressively disabling end-stage pulmonary or cardiopulmonary disease with a predicted life expectancy of less than 12 to 18 months (see Boer & Mannes, 1993). The first phase of the program is the application phase in which potential candidates are identified on the basis of written information of the referring physician. The other phases of the program are outpatient screening, inpatient screening, pretransplantation, waiting list, transplantation (perioperative and intensive care), inpatient follow-up, and outpatient follow-up. A total of 425 patients was referred to the program. Of these patients, 303 entered the outpatient screening phase and 179 were accepted for the inpatient screening. One hundred-and-twenty patients were placed on the waiting list. Finally, 57 patients received a transplantation. Two patients died during the transplantation phase and nine patients died during the follow-up.

During the screening phase, patients may be rejected, screening may be deferred, or patients do not contact the lung transplantation team for more than 12 months. With exception of the application phase, for all phases and for all patients, length of stay was registered, with the reason for leaving the phase. Furthermore, all costs (direct

medical, direct nonmedical and indirect nonmedical) and data on quality of life were collected. In addition, several other patient characteristics were registered, of which we used diagnosis, age, body length and blood type. Diagnosis and age were used as explanatory variables for length of stay in various phases and for survival. Length and blood type were used for matching donor and recipient.

For the same period, we also collected data on donor lungs: acquisition date of the lung and length and blood type of the donor.

Quality of life

The health-related quality of life of the patients was measured through a self-administered questionnaire. It contained several domain-specific instruments (Karnofsky Performance Index, Index of Well-Being, Self-rating Depression Scale, State-Trait Anxiety Inventory, activities of daily living) and two generic instruments: the Nottingham Health Profile and the EuroQol. Patients were asked to fill out the questionnaire at the entry of the outpatient screening phase, and from then on every 3 months. Following transplantation, quality of life was measured after 1, 4, and 7 months and from then on every 6 months.

In this analysis, effectiveness is measured both as life-years gained and quality-adjusted life-years (QALY) gained. For the latter, it is necessary to express quality of life as a number between 0 and 1 (utility), where 0 represents the worst possible health state, and 1 the best. The EuroQol group has developed an algorithm that allows the calculation of the utility that represents the health state reported by the patient (see Bjork, 1992). Note that this utility reflects the value the general population assigns to health states.

Table 6.1 presents the average quality of life (as measured by the EuroQol score) of patients who did not die while on the waiting list. It shows that the health-related quality of life, already poor during the screening, deteriorates further if patients remain on the waiting list for a year or longer. For patients who died within 1 year after being placed on the waiting list, the utility was lower: on average 0.4 (n=10). For those patients, the utility in the 3 months before death was 0.31.

table 6.1 Quality of life on waiting list (patients who did not die while on waiting list)

Average EuroQol score and standard deviations (in parentheses)

Phase	Screening	Waiting list 0-6 months	Waiting list 6-9 months	Waiting list 9-12 months	Waiting list 12-15 months	Waiting list > 15 months
Utility (SD)	0.52 (0.2)	0.55 (0.16)	0.50 (0.18)	0.45 (0.2)	0.40 (0.15)	0.40 (0.12)
N	169	30	30	27	18	11

Immediately after transplantation, while the patient is still in hospital, for survivors the average utility has increased to 0.69 (N=24) and improves further, reaching normal values. Table 6.2 presents the utilities associated with patients' quality of life during the outpatient follow-up phase. The other quality of life instruments also showed a substantial improvement of quality of life after transplantation (see Ten Vergert et al., 1998).

In all phases, except for the waiting list, fewer than five observations were available of quality of life during the last 3 months before death. Therefore, we assume that in every phase, patients' quality of life is 0.30 during the last 3 months before death. Furthermore, for the phases until outpatient follow-up, we have used the utilities as presented before. For the first 2 years of outpatient follow-up, we have assumed that quality of life has a value of 0.85; after 2 years, this increases to 0.90.

table 6.2 Quality of life after transplantation

Average EuroQol score and standard deviations (in parentheses)

Phase	1-3 months follow-up	4-6 months follow-up	7-12 months follow-up	13-19 months follow-up	20-25 months follow-up	>25 months follow-up
Utility (SD)	0.83 (0.16)	0.85 (0.14)	0.84 (0.15)	0.86 (0.12)	0.91 (0.1	0.90 (0.12)
N	30	24	17	15	12	11

Costs

Data on all direct medical, direct nonmedical and indirect nonmedical costs, i.e., value of production losses (paid or unpaid work), related to the lung disease, were gathered for all patients, from the moment they entered the outpatient screening phase until they left the program. Where possible, full resource costs were estimated (see Enckevort et al., 1997; base year 1992). Table 6.3 presents for each phase the average costs per patient per cost category. The highest costs occur during inpatient screening, on the waiting list, in the transplantation phase and during follow-up. In general, average costs per patient are higher than median costs and standard deviations are substantial, reflecting skewed distributions of costs (see table 6.3). This skew is due to a minority of patients causing very high costs (e.g. due to complications). This pattern is very normal in numerous studies of medical consumption (e.g. Duan et al., 1983).

We used the sum-limit method as described by Hout et al. (1993) to calculate cumulative costs by length of stay per phase, per patient, and per reason for leaving the phase. These cumulative costs were then used to extrapolate the cost data beyond the observation period (or beyond the date for which fewer than five observations were available). Almost all cumulative costs could be estimated by a linear function or by a combination of two linear functions (R^2 > 95%). For instance, for patients who died during the inpatient screening phase, direct nonmedical costs were NLG 104 per week whereas indirect nonmedical costs were NLG 174 per week in the first 15 weeks, and from then on NLG 110 per week.

table 6.3 Average costs per patient per phase during study period,
in NLG (1 NLG = 0.6 US$)
Median costs and standard deviations in parentheses (median/ standard deviation)

Phase	Direct medical costs within UHG*	Direct medical costs outside UHG*	Direct nonmedical costs[1]	Indirect nonmedical costs[2]
Outpatient screening	964 (260/1,941)	8,107 (2,583/16,289)	1,244 (590/1,859)	2,678 (1,010/4,326)
Inpatient screening	24,334 (22,587/14,003)	10,290 (8,092/8,441)	2,749 (2,302/2,095)	4,966 (3,752/5,711)
Pretransplantation	603 (37/1,459)	3,471 (1,453/6,155)	691 (465/1,344)	838 (570/1,078)
Waiting list	16,448 (9,299/23,043)	30,174 (20,736/28,380)	5,099 (4,255/4,403)	9,372 (7,178/8,917)
Transplantation	82,557 (67,956/44,228)	121 (0/329)	602 (451/608)	121 (15/234)
Inpatient follow-up	55,766 (53,796/21,005)	517 (0/948)	1,401 (1,076/1,248)	522 (0/950)
Outpatient follow-up (on average 510 days)	71,521 (70,129/47,422)	39,186 (28,473/30,964)	5,589 (5,130/3,989)	6,966 (3,429/9,368)

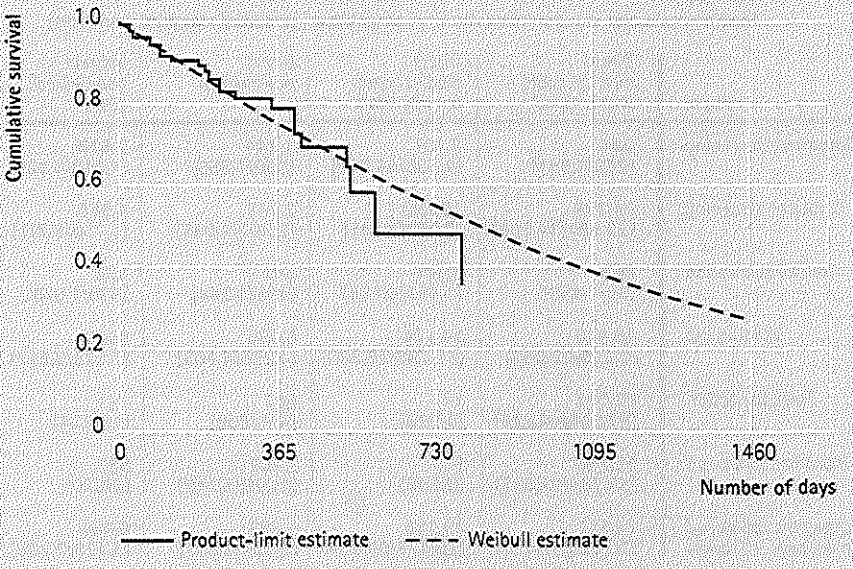
*) University Hospital Groningen
1) such as travel costs, diet costs, costs of medical supplies
2) costs of absence from work (paid work and unpaid work, e.g. household work)

Costs for the situation without transplantation were derived from the cost data as gathered for the situation with the transplantation program. It was assumed that until transplantation, the conventional treatment of patients was not influenced by the existence of the transplantation program. The following cost categories are only relevant in case of a transplantation program: all direct medical costs within the University Hospital Groningen (except for a few patients who receive their conventional treatment in Groningen), all costs in the transplantation phase and follow-up, indirect nonmedical costs during the inpatient screening phase, conditioning costs on waiting list (medication, special diets and physiotherapy) and travelling costs to Groningen.

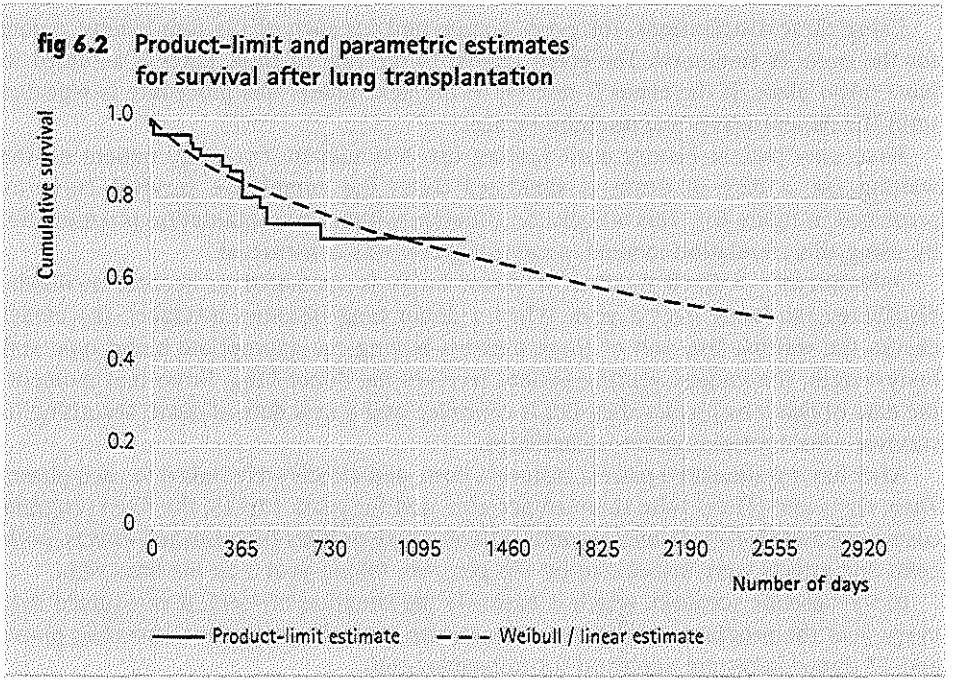
Survival

To estimate survival without transplantation, survival on the waiting list was used, by defining transplantation as censoring. A parametric model (Weibull) was used to estimate survival, thus allowing extrapolation beyond the observation period. Figure 6.1 shows both the product-limit and the parametric estimates for survival on the waiting list.

fig 6.1 Product-limit and parametric estimates for survival on the waiting list for lung transplantation



After transplantation, the 1- and 2- year survival rates were 86% and 75%, respectively. Not enough data were available to extrapolate survival beyond 3 years. We therefore combined our data with international data on survival after heart and lung transplantation (see Hosenpud et al., 1995; St. Louis International Lung Transplant Registry, January 1997). The data available from the St. Louis International Lung Transplant Registry clearly show that cumulative survival after lung transplantation decreases with 5% per year from year 3 till year 6. Furthermore, data on heart transplantation show that after 1 year, the cumulative survival decreases with 4% per year until year 11. Combining this information, we have estimated survival after transplantation for the first 3 years with a Weibull model and after 3 years, cumulative survival decreases with 5% per year (figure 6.2).



Method

A micro-simulation model was used to calculate the cost-effectiveness of the Dutch lung transplantation program in the next 15 years. A period of 15 years was chosen to make sure that a steady state was reached, i.e., a situation in which the number of patients per phase (before transplantation) is stable. After these 15 years, no transplantations are performed, but survival, costs and quality of life are calculated for a follow-up period of 25 years after the program has stopped; thus, the total evaluation period is 40 years.

The model simulates individual patient histories, containing the exact date a patient enters and leaves a phase. By linking length of stay in each phase with cost and quality of life estimates, the total costs and effects of the program are estimated. To simulate length of stay in, for instance, the outpatient screening phase, four distributions were estimated, based on the Groningen data: one regarding duration until death, one until referral to the next phase, one until rejection and one until the patient had not contacted the lung transplantation team for more than 12 months. These distributions were estimated by means of survival analysis, thus allowing for censoring of the data. If the duration distribution until death is estimated, censoring will occur if the patient is rejected, referred or 'has no contact', or if no event has taken place before the end of the study. With the same technique, duration distributions for the other three events are calculated.

Using these distributions, for every patient and for each of the four events, a date for the event is simulated. The event related to the first of these dates is assumed to have taken place. If the event is referral to the next phase, this procedure is repeated until the patient leaves the program or dies.

First, only patient history until waiting list is simulated. For every patient on the waiting list, a date of death is simulated using the cumulative survival curve as presented in figure 6.1. This reflects the situation without transplantation program. So, by using this simulation method, a control group was constructed.

Subsequently, a fixed number of donor lungs is simulated to become available, distributed randomly over a year. When a donor lung becomes available, the model checks the waiting list, and of those patients having the appropriate blood type and body length, the longest waiting patient receives the donor lung. From that moment, the date of death as predicted earlier is cancelled and a new date of death according to survival after transplantation is simulated.

After having simulated patient histories, costs and effects were linked to each phase. Then, costs and effects were summed per year, and for years 1 to 15, the fixed program costs were added. To take into account different time profiles for costs and effects, both costs and effects were discounted by 5% per year, taking year 1 as the base year.

In the baseline scenario, which is basically the situation of 1995, it is assumed that every year the program is effective, 100 patients enter the outpatient screening phase and 17 donor lungs are approved for transplantation.

The cost-effectiveness of two other scenarios will be assessed as well. First, it is anticipated that the baseline scenario will show a rapid increase of the number of patients on the waiting list, due to the small number of available donor lungs. Therefore a scenario (the policy scenario) will be assessed in which the number of patients entering the program is restricted.

Second, it has been estimated that with extensive effort, the supply of donor lungs in The Netherlands may be increased to 27 per year (see TenVergert et al., in press). In the donor scenario the impact of such an increase on the cost-effectiveness of the program is calculated.

Results

Base-line scenario

Patient flow: Each program year, 100 patients will enter the outpatient screening phase. After four years, the number of patients per year who enter a specific phase becomes stable: 65 patients enter the inpatient screening; 50 patients are placed on the waiting list; and 17 undergo transplantation.

The first years of the program, more patients enter the waiting list than leave (either because of transplantation or death). After 10 program years, the number of patients on the waiting list at the end of the year stabilises (n=105). In this situation, 50 patients enter the waiting list, 17 undergo transplantation, and 33 die. During the 15 program years, the number of patients in the follow-up phase at the end of each year increases. This reflects the fact that each year, more patients undergo transplantation than die. The number of deaths per year during the follow-up phase is 12 in year 10, and 15 in year 15.

Because survival, both on the waiting list and after transplantation, differs for each diagnosis group, we also studied the distribution across diagnosis groups before and after transplantation. In the outpatient phase, 42% of patients have emphysema/COPD versus 58% of the patients with transplants. The percentage of patients with pulmonary hypertension drops from 17% in the outpatient screening phase to 11% after transplantation; for lung fibrosis patients the percentage decreases from 19 to 10%. The share of patients having cystic fibrosis or other diagnoses remains the same before and after transplantation.

Costs: The total discounted costs with the program amount to NLG 246 million and without the program to NLG 130 million. Table 6.4 presents the total costs in four categories for the evaluation period of 40 years.

table 6.4 Baseline scenario Total costs for the full evaluation period, per type of costs, discounted by 5%, in million NLG (1 NLG = US$ 0.6). Costs for 1992

	Direct medical costs inside UHG*	Direct medical costs outside UHG*	Direct nonmedical costs	Indirect nonmedical costs	Total
Costs with program	88	112	16	30	246
Costs without program	0	88	13	29	130
Incremental costs	88	24	3	1	116

*) University Hospital Groningen

For the situation with the program, the costs outside the transplant centre are the highest, whereas during the study period 1991 to 1995, the costs inside the transplant centre dominated. This is explained by the rapid increase of the number of patients on the waiting list, predicted with the simulation model. Patients on the waiting list induce much higher costs outside the centre

Health effects: Without discounting, the total number of life-years during the evaluation period (40 years), in the situation with the program, amounts to 5,494. The number of life-years gained, compared to the situation without program, are 1,232. The total number of transplantations during the evaluation period are 242, yielding 5.1 life-years gained per patient with transplant. The number of QALYs gained is somewhat higher: 1,358, due to the large difference between quality of life on the waiting list and after transplantation. Table 6.5 presents both life-years and QALYs after discounting by 5% per year.

table 6.5 Baseline scenario Life-years and quality-adjusted life-years (QALYs) for the full evaluation period, discounted by 5%

	Life-years	QALYs
Effects with program	3,264	1,996
Effects without program	2,664	1,297
Life-years or QALYs gained	600	699

The total costs per life-year gained (after discounting) amount to NLG 194,000 (table 6.6), the costs per QALY gained are lower: NLG 167,000. If only direct medical costs are taken into account, the cost-effectiveness ratios are slightly lower. From table 6.6 it is clear that the cost-effectiveness ratios are notably influenced by discounting: the cost per life-year/QALY gained increased by 15 to 20% after discounting.

table 6.6 Baseline scenario Costs per life-year/ QALY gained of lung transplantation in NLG (1 NLG = US$ 0.6). Costs for 1992

	Costs per life-year gained		Costs per QALY gained	
	0% disc	5% disc	0% disc	5% disc
Direct medical costs	155,000	188,000	140,000	161,000
Total costs	162,000	194,000	147,000	167,000

Policy scenario

In the baseline scenario, two thirds of patients on the waiting list die, which is a highly undesirable situation, both for the patients and for the physicians involved in the program. Therefore, we calculated cost-effectiveness if the program inflow is restricted in such a way that the number of patients admitted to the outpatient screening is such that no more than 50% of the patients eventually die on the waiting list (assuming that the probability of entering the waiting list remains unchanged). This would mean that no more than 68 patients per year should enter the outpatient screening phase. Of these 68 patients, 35 would be placed on the waiting list, and 17 patients would then undergo transplantation per year. If the number of patients entering the program would decrease to <65, not enough patients would be on the waiting list to find a match for all 17 donor lungs. After 10 years, a steady state would be reached, where at the end of the year, 55 patients would be waiting for transplantation.

In this scenario, the additional costs with the program, after discounting, amount to NLG 95 million, NLG 22 million less then in the baseline scenario. The number of life-years gained and QALYs gained are 550 and 656, respectively (after discounting). Thus, the costs per life-year gained are more favourable, NLG 172,000, and the costs per QALY gained are NLG 144,000. If only direct medical costs are taken into account, these ratios are NLG 168,000 and NLG 141,000, respectively.

Donor scenario

In this scenario, the number of patients entering the screening phases and waiting list are the same as in the baseline scenario. From year 4 on, 27 patients undergo transplantation per year, whereas 23 patients die on the waiting list. In the steady state, which is reached after 10 years, the number of patients waiting for a transplant is 75, compared to 100 in the baseline scenario.

With 100 patients entering the outpatient screening each year, all 27 available donor lungs match with at least one patient on the waiting list. If the number of patients entering the program falls below 95, this may not always be the case.

The total costs in this scenario are 26% higher than in the baseline scenario, NLG 147 million. The number of life-years gained and QALYs gained are about 55% higher, 923 and 1,089, respectively. The cost-effectiveness ratios are NLG 159,000 per life-year gained and NLG 135,000 per QALY gained (after discounting costs and effects).

Very recently, the legislation in The Netherlands concerning organ donation has changed (see Eerste Kamer der Staten-Generaal, 1995-1996). The previous system assumed no permission for donation, unless explicit permission was given by the donor (e.g. by means of a 'donor codicil') or his/her relatives. In the new system, any Dutch citizen will be invited to fill in a response card giving (or not) permission to donate specific organs. A national registry will keep an up-to-date database of these responses, which can be consulted if necessary. It is expected that if this system is fully operational, the number of donor organs will increase.

Sensitivity analysis

The lifelong use of immunosuppressive medication during follow-up after transplantation is a major element in the costs of lung transplantation. For one of the most often used immunosuppressive drugs, cyclosporine, the future costs may fall as a result of completing the patent period. It is difficult to predict the extent of a possible price decrease. However, if we would assume a 50% cost reduction in follow-up medication, total incremental costs for the baseline scenario would fall by NLG 11.8 million (5% discounting). The costs per life-year gained (and per QALY gained) would be 10% lower. However, new, more expensive, immunosuppressive medication is already being used. Widespread application of these drugs will lead to cost increases, but this might be offset by a better survival and/or less drug toxicity, which may improve quality of life.

During the inpatient screening phase, patients are hospitalised for several weeks in the University Hospital Groningen to undergo an extensive number of tests. As lung transplantation is still relatively new, it may be expected that evaluation of the screening process will result in a more limited, but equally effective, screening in the future. If the costs of inpatient screening in the transplant centre could be halved, incremental costs and cost per life-year gained (and QALY gained) would decrease by 6% in the baseline scenario.

The long-term survival after lung transplantation is still very uncertain. If future long-term survival would deviate significantly from the survival as assumed above, it could clearly result in a substantial change in effectiveness. The influence on cost-effectiveness could be considerable, but depends also on the specific costs during the additional life-years gained.

Discussion

Lung transplantation is an expensive, but effective intervention; survival and quality of life improve substantially after transplantation. This analysis suggests that regardless which scenario would effectuate in the near future, lung transplantation remains expensive in terms of costs per life-year (or QALY) gained.

Crucial elements determining cost-effectiveness are the number of patients screened and placed on the waiting list as compared to the number of available donor organs as well as the substantial costs of follow-up after transplantation. As more patients are screened and more patients are waiting (longer) for transplantation, then more costs are incurred without any gains in health effects. Restricting the inflow in the screening phase (e.g. being even more restrictive concerning contraindications) can improve the balance between the costs of screening and the health effects of transplantation.

We did not try to establish a true "optimal scenario" in terms of cost-effectiveness, but combining the policy and donor scenario (restricted inflow and more donors) would result in NLG 151,000 per life-year gained (NLG 124,000 per QALY), which is slightly more favourable than the results of the donor scenario. This scenario has the disadvantage that not all donor organs will be used for transplantation.

The quality of life and utility scores for patients with transplants improved substantially. This is in accordance with the findings of Ramsey et al. (1995b) for lung transplantation patients. About the same improvements in utility scores were found in the Dutch heart transplantation study (see Hout et al., 1993).

Sensitivity analysis showed that a more limited inpatient screening process could result in some cost reduction. However, the feasibility of such a rationalisation should first be investigated.

Comparison with Dutch programs for heart and liver transplantation shows that cost-effectiveness for lung transplantation is relatively unfavourable. Costs per life-year gained (5% discounting) for heart and liver transplantation were NLG 66,000 and NLG 54,000 (see Michel et al., 1992; Hout et al., 1993; costs adjusted to 1992). This difference can not be explained fully by different methods of analysis or inclusion of different cost categories. For heart and liver transplantation, the average number of life-years gained per patient with transplant is higher: 10.5 and 7.6 years, respectively. Furthermore, the costs during a year on the waiting list or a year of follow-up after transplantation are substantially higher for lung transplantation as compared to heart and liver transplantation.

In the meantime, the Dutch National Health Insurance Board advised the minister of Health Affairs for the moment not to include lung transplantation in the benefit package. The transplantation program will proceed (subsidised by a development grant) but further research should indicate if costs can be reduced, especially during the screening phase (by reducing the number of patients screened and/or lowering the costs per patient screened) and the follow-up phase.

Financial support

This study was supported by Grant OG 91-053 from the Fund for Investigative Medicine of the Dutch Health Insurance Board.

CHAPTER 7

Optimal allocation of resources over health care programs: dealing with decreasing marginal utility and uncertainty

Summary

This chapter addresses the problem how to value health care programs with different ratios of costs to effects, specifically when taking into account that these costs and effects are uncertain. First, the traditional framework of maximising health effects with a given health care budget is extended to a flexible budget using a value function over money and health effects. Second, uncertainty surrounding costs and effects is included in the model using expected utility. It is shown that - depending on risk attitude - taking account of uncertainty may result in a different ordering of health care programs. Other approaches to uncertainty that do not specify a utility function are discussed and it is argued that these too include implicit notions about risk attitude. Hence, to take account of uncertainties in the ordering of health care programs, attention needs to be paid to the marginal values of money and health effects and to risk attitude.

Introduction

Cost-effectiveness analyses often conclude by stating that the obtained cost-effectiveness ratio is well within the acceptable limit and it is thus advisable to implement the therapy under investigation on a wide scale. An example can be found in the Dutch guidelines on cholesterol-lowering therapy. Supported by a cost-effectiveness analysis, the guidelines advise to treat all patients with a 22% or higher 10-year risk for a cardiovascular event. This type of advice is problematic. Although it has been estimated that for such a strategy the costs per life-year gained are under EUR 20,000, no indication is given where the budget for treatment may be found. Given the number of untreated individuals in The Netherlands, the additional costs are estimated at EUR 100 million, a considerable amount. Thus, the question arises whether budget should be allocated to the new therapy. A related problem concerns the uncertainties surrounding this cost-effectiveness ratio. A point estimate of EUR 20,000 may be deemed acceptable in a league table, but this may not hold true for the uncertainty margins. Fear of being confronted with a higher than expected cost-effectiveness ratio when implementing this strategy may induce the decision maker to value the cost-effectiveness lower than the point estimate. More generally, the question can be raised how budget should be allocated when programs differ not only in terms of their cost-effectiveness but also in terms of the uncertainty surrounding both costs and effects. These questions are addressed here using a formal mathematical framework.

First we will present the mathematical rules of cost-effectiveness as introduced by Weinstein & Zeckhauser (1973) (see also Torrance et al., 1972). As we will see, these rules (that do not account for uncertainty) imply that the required budget for a new cost-effective therapy has to be taken from existing less cost-effective treatments in the health care sector. However, the additional budget to implement new medical treatments could also be sought in other governmental departments (e.g. education), or might be obtained from an increase in e.g. taxes or insurance premiums. A pragmatic alternative to the standard model is the use of a fixed trade-off between costs and health effects. However, as argued by Gafni & Birch (1993), that approach would lead to an ever-increasing health care budget.

Next, a rule similar to that derived by Weinstein and Zeckhauser (still assuming certainty) is explored, that can be derived to support a more flexible trade-off between budgets. Formulating the model clarifies the need to establish the right moment to stop increasing the health care budget at the cost of either money available for other departments or increases in taxes and premiums. In other words, consideration should be given to the formulation of a value function characterised by the decreasing marginal value of both money and health effects.

Then an approach including uncertainty is presented which combines the results from the previous section with the theory of expected utility. This approach requires the explicit formulation of a utility function, expressing risk attitude towards money and health effects. A separate section discusses how to handle uncertainty without explicitly formulating such a utility function. In the final section, we discuss our results

and compare them with other recent work on uncertainty and budget allocation (e.g. Stinnett & Paltiel, 1997; O'Brien & Sculpher, 2000; Meltzer, 2001).

The classical approach

The classical approach to support thinking in terms of cost-effectiveness ratios was introduced in 1973 by Weinstein & Zeckhauser, who addressed the problem of maximising health effects within a given budget mathematically. They assumed (as we will do) independence of programs, constant returns to scale, and a finite, fixed size of each program. Conditions can be adjusted to take mutually exclusive programs into account (see Laska et al., 1997).

Denote, for n health care programs, the total costs of each program i by c_i and its health effects by e_i. Let d_i, $0 \leq d_i \leq 1$, represent the fraction[1] of the i^{th} program undertaken and B the total budget. The budget allocation problem can be formalised as:

$$\max_{d_i} \sum_{i=1}^{n} d_i e_i$$

$$\text{subject to} \sum_{i=1}^{n} d_i c_i \leq B.$$

Weinstein & Zeckhauser (1973) showed that the optimal solution is always characterised by[2]:

$$\begin{cases} \frac{c_i}{e_i} < \alpha \;\rightarrow\; d_i^* = 1 \\ \frac{c_i}{e_i} > \alpha \;\rightarrow\; d_i^* = 0 \\ \frac{c_i}{e_i} = \alpha \;\rightarrow\; d_i^* = \pi \end{cases} \qquad \textbf{Eq. 1}$$

Here α and π have to be chosen such that the budget constraint is met with equality. The parameter α can be interpreted as the maximum cost-effectiveness as accepted within the given budget. Programs with a lower ratio of costs to effects than α are undertaken, while programs with a higher ratio are not. Note that if the budget increases, all else remaining equal, the critical ratio α increases, which means that more programs become acceptable. Thus, the new, higher budget is again used

1. If programs are indivisible, d_i should be either 0 or 1, indicating whether or not the program is undertaken. See Birch & Donaldson (1987) and Stinnett & Paltiel (1996)
2. We reformulated Weinstein & Zeckhauser's solution in terms of cost-effectiveness ratios, so that their γ equals $1/\alpha$.

completely. Stated differently, the value of $1/\alpha$ indicates the shadow price of the budget constraint. If this value approaches zero, the constraint is not binding and all available programs can be financed. Throughout this chapter we will assume that in the optimum (for each approach) some existing programs are not (completely) implemented.

To consider the effect of the addition of new programs to an existing set of programs already implemented, define d_i^* as the optimal fractions for the given set of n programs, with optimal α^* and a budget B^*. Let a new program have costs c_{new} and effects e_{new}. If $c_{new}/e_{new} > \alpha^*$, the program is not worth introducing and the optimal set of programs and the values d_i^* do not change.

If $c_{new}/e_{new}=\alpha^*$, the new program is as cost-effective as the least cost-effective program(s) with fraction $d_{last}^*>0$. One is indifferent as to whether or not the new program is (partially) implemented. The fractions d_{last} and d_{new} make no difference to the optimal health effects that can be obtained.

If $c_{new}/e_{new} < \alpha^*$, the new program is better than program "last". A new optimal allocation of the budget must be found, whereby some of the programs currently implemented are now (partially) abandoned, because the budget is needed to finance the new program. For $c_{new}/e_{new} < \alpha^*$, the new critical ratio will be smaller than or equal to α^*. This means that, for a fixed budget, the greater the number of efficient programs existing, the lower the ratio of costs to effects of the programs included in the set of optimal programs, i.e. the higher the requirements will be for new programs to enter.

In practice, researchers often neglect the dependency between the critical ratio and the available budget and choose the critical ratio to be fixed at some value γ (e.g. Centraal begeleidingsorgaan voor de intercollegiale toetsing, 1998; Niessen et al., 2000). According to this approach the optimal set of programs is given by the following conditions:

$$\begin{cases} \frac{c_i}{e_i} < \gamma \;\rightarrow\; d_i^* = 1 \\ \frac{c_i}{e_i} > \gamma \;\rightarrow\; d_i^* = 0 \\ \frac{c_i}{e_i} = \gamma \;\rightarrow\; 0 \le d_i^* \le 1 \end{cases} \qquad \textbf{Eq. 2}$$

In that case, the introduction of a new intervention has no effect on the critical ratio of costs to effects. If $c_{new}/e_{new} > \gamma$, then the optimal set of programs will not change. If $c_{new}/e_{new} = \gamma$, the new program is just on the boundary of being cost-effective and d_{new} is a matter of choice. If $c_{new}/e_{new} < \gamma$, a new optimum with a larger amount of money spent on health care and more health effects is obtained. No current programs are reduced. Note that this would lead to unlimited growth of the health care budget if new, more efficient programs continue to become available. This approach may allow one to forget the problem of finding the additional money to fund the new "acceptable" therapies.

Examples

As a theoretical example, suppose there are three programs that might be implemented. Table 7.1 presents the costs and effects of these programs.

table 7.1 Costs and health effects for three fictitious health care programs

Program	Total costs	Total health effects	Ratio costs to effects
1	400,000	1,900	211
2	450,000	1,800	250
3	300,000	1,300	231

If the goal is to maximise health effects for a fixed budget of 750,000, then the optimum is achieved for $d_1^* = 1, \quad d_2^* = 1/9, \quad d_3^* = 1$. In this optimum, the total health effects are 3,400 units, and the critical ratio is 250.

If the question whether or not to implement a program is based on a fixed critical ratio of costs and effects, and a critical ratio of 250 is chosen, programs 1 and 3 will again be fully implemented, but the decision maker is indifferent to all fractions of program 2. As soon as a marginally larger critical ratio is chosen, all programs will be implemented, while a marginally smaller critical ratio will lead to the implementation of programs 1 and 3 only. Thus, total costs will be between 700,000 and 1,150,000 and health effects will be between 3,200 and 5,000.

Now suppose that a new program becomes available, with costs 225,000, effects 1,000 and a ratio of costs to effects of 225. For the situation with a fixed budget, the new optimal set becomes $d_1^* = 1, \quad d_2^* = 0, \quad d_3^* = 5/12, \quad d_{new}^* = 1$. In this optimum, the total health effects are 3,442 units, and the critical ratio is 231. For the situation with a fixed critical ratio of 250, programs 1, 3 and "new" are fully implemented and, again, the decision maker is indifferent to all fractions of program 2.

As a practical example, consider the introduction of cholesterol-lowering therapy on a large scale. Suppose that the health care budget in The Netherlands was sufficient to fund all programs up to a limit of EUR 40,000 per life-year gained. This means that programs (starting with those on the border) have to be cancelled to free the budget needed for cholesterol-lowering therapy until everything fits the total budget. Moreover, a new critical ratio results, for example EUR 39,000.

This example raises a number of questions: for instance, which health care programs are on the border, or what is the critical ratio (as derived from the total health care budget), or even what is the total budget? The last question is already hard to answer, let alone the second and the first. This may be one of the reasons behind using the approach where an acceptable critical ratio is defined without considering the interaction between this ratio and the total health care budget.

A wider perspective

The standard approach starts with a given health care budget without consideration of the decision process that defined the budget; this process may be similar to that of choosing between alternative health care programs. When considering the whole governmental budget, money must be allocated to sectors such as health care, education, and public safety. The determination of the health care budget may be seen as the result of negotiations between the department of health, the department of defence, the department of education etc. In these negotiations, implicitly a comparison is made between the effects of spending money in the different sectors: if the effects of spending EUR 1 million on education are valued higher than the effects of spending EUR 1 million on health care, this money will be allocated to the department of education.

One formulation of the health care budget allocation problem that allows for more flexibility than a fixed budget, is to use a value function defined both over health and money as the objective of the maximisation problem. That is, the decision maker chooses to implement those health care programs that maximise the value function. Value is derived from budget available for purposes competing with health care spendings and from health. This value function represents the decision maker's preference structure for these attributes (e.g. Keeney & Raiffa, 1993). Note that the value function is characterised by decreasing marginal value from both money and from health effects.

Once a function V is defined on budget and health effects, the goal is to find values d_i so as to maximise value, given n health care programs with costs c_i and effects e_i. This can be written as:

$$\max_{d_i} V(Y-c,e)$$
$$\text{subject to } 0 \le d_i \le 1.$$

Here c= $\sum_i d_i c_i$, e= $\sum_i d_i e_i$, and Y is the total budget of the decision maker.

The first order conditions for an optimum are:

$$\begin{cases} \frac{c_i}{e_i} < \frac{V_2(Y-c,e)}{V_1(Y-c,e)} & \rightarrow \quad d_i^* = 1 \\ \frac{c_i}{e_i} > \frac{V_2(Y-c,e)}{V_1(Y-c,e)} & \rightarrow \quad d_i^* = 0 \\ \frac{c_i}{e_i} = \frac{V_2(Y-c,e)}{V_1(Y-c,e)} & \rightarrow \quad 0 \le d_i^* \le 1 \end{cases} \qquad \textbf{Eq. 3}$$

Here $V_1 = \frac{\partial V}{\partial (Y-c)}$ and $V_2 = \frac{\partial V}{\partial e}$

Note that decreasing marginal value implies that $V_{11}<0$ and $V_{22}<0$,

with $V_{11} = \frac{\partial^2 V}{\partial (Y-c)^2}$ and $V_{22} = \frac{\partial^2 V}{\partial e^2}$

Clearly, there is a similarity to the conditions presented in the previous section (see equations 1 and 2). The marginal rate of substitution takes the place of the critical cost-effectiveness ratio α (γ). As such the critical cost-effectiveness ratio is defined by the value function, which is characterised by decreasing marginal value. From this follows a certain budget for health care.

Note that the approach with a fixed critical ratio can be represented by a linear value function with constant marginal derivatives, that is, with a fixed trade-off between costs and effects. This is written as $V = (Y-c) + \gamma * e$, with $V_1 = 1$ and $V_2 = \gamma$. In that case, a value function is assumed without decreasing marginal value in either money or health effects. It is unlikely that such a function can adequately describe a decision maker's preferences.

In general, the critical ratio decreases with less money and more health until it is equal to the cost-effectiveness of the last program to be (partially) implemented.[3] This determines the budget that is spent on health care.

Consider again the introduction of a new intervention, with d_i^* being the fractions of each program in the current optimum, V_2^*/V_1^* being the marginal rate of substitution in that optimum, and program "last" being the last program to be (partially)

implemented. In this situation, $\frac{V_2^*}{V_1^*} = \frac{c_{last}}{e_{last}}$ or, in words, the marginal rate of substitution is equal to the largest (i.e. least favourable) ratio of costs to effects in the current optimal set of programs.

Suppose that a new program is considered for implementation, with a ratio of costs to effects of c_{new}/e_{new}. In order to be at least partially implemented,

this ratio must satisfy the condition $\frac{c_{new}}{e_{new}} \leq \frac{c_{last}}{e_{last}}$.

---

3. Corner solutions with $d_i^*=1$ and binding for all programs implemented are also possible. Denote by "not" the best program among those not implemented.

In this situation $\frac{c_{not}}{e_{not}} > \frac{V_2^*}{V_1^*} > \frac{c_{last}}{e_{last}}$, i.e. the marginal rate of substitution in the optimum is larger than the ratio of costs to effects of the least favourable program implemented, but smaller than that of the best program among those not implemented. In order for a new program to be at least partially implemented,

the ratio must now satisfy $\frac{c_{new}}{e_{new}} \leq \frac{V_2^*}{V_1^*}$ and $\frac{V_2^{new}}{V_1^{new}} \leq \frac{V_2^*}{V_1^*}$.

Whether or not the new program is fully implemented will depend on the size of the program. Independent of the size of the new program,

$$\frac{V_2^{new}}{V_1^{new}} \leq \frac{V_2^*}{V_1^*},$$

i.e. the critical ratio never increases as a new program becomes available.

Since the budget is flexible, adjustment to the new situation takes place not only through changes in the di, but also through a change in the health care budget. The therapies at the border, that would have been cancelled in the standard approach, are now evaluated against programs on the borders of other budgets.

Examples (continued)

For the theoretical example, assume a decision maker wants to allocate budget over the three programs mentioned in the previous section (see table 7.1). However, this decision maker has a non-linear value function that can be written as $V(Y-c,e)=(Y-c)^{0.75}e^{0.6}$. Furthermore, assume that the total budget Y is 2,000,000. The optimum solution is $d_1^*=1$, $d_2^*=0.3$, $d_3^*=1$, and the critical ratio is 250. In this optimum, the total costs are 835,000, with total health effects of 3740.

If we compare this with the example in the previous section, we see that the critical ratio in this example is the same as for the fixed budget approach, but that the fraction of program 2 implemented is now somewhat larger. The preferences stated in the value function above indicate a willingness to increase the budget in exchange for more health effects.

As a practical example, imagine that the total health care budget is defined by the marginal rates of substitution such that the corresponding critical ratio is approximately EUR 40,000. Again, the introduction of cholesterol-lowering therapy might push some therapies that are already implemented out of the health care system. Within this framework, however, these programs might also compete with other non-health care expenditures. Thus, building hospitals would not only be compared with building outpatient clinics but also with building, e.g. police stations. This implies that the changes in the optimal fractions are smaller than they would have been in case of a fixed budget. Therefore, some interventions that would be removed completely in case of a fixed budget, may now still be undertaken. Moreover, while in the fixed budget approach the critical ratio might decrease to EUR 39,000 it might now only decrease to EUR 39,500.

Decisions under uncertainty using the wider perspective

Until now we have assumed that costs and effects for all programs are known with certainty, without having any example for which this is true. To include risk attitude into decision making, we will combine the results from the previous section with the well-known concept of expected utility (e.g. Keeney & Raiffa, 1993). The expected utility approach implies the definition of a utility function over the set of uncertain outcomes (Y-c, e) that satisfies the expected utility property, i.e. the utility of a stochastic event is represented by the expected value of the utilities of all possible realisations of the event (see e.g. Keeney & Raiffa, 1993; Varian, 1984).

Again, $c=\sum_i d_i c_i$, $e=\sum_i d_i e_i$, and Y is the total budget of the decision maker.

Let $x_i = (c_i, e_i)$ be stochastic and assume that the x_i are independent for all i.
The general goal is now to find d_i so as to maximise expected utility:

$$\max_{d_i} E[U(Y-c,e)] \text{ subject to } 0 \le d_i \le 1.$$

The first order conditions for an interior optimum can now be derived:

$$E[\frac{\partial U}{\partial d_i}] = 0 \text{ for } 0 < d_i < 1.$$

Using similar expressions for corner solutions, and defining

$U_1 = \frac{\partial U}{\partial (Y-c)}$ and $U_2 = \frac{\partial U}{\partial e}$ as the marginal utility of money (i.e., the budget available for purposes competing with health care spendings) and the marginal utility of health effects, respectively, we find the following characterisation of the optimum:

$$\begin{cases} E[e_i U_2 - c_i U_1] > 0 & \rightarrow \quad d_i^* = 1 \\ E[e_i U_2 - c_i U_1] < 0 & \rightarrow \quad d_i^* = 0 \\ E[e_i U_2 - c_i U_1] = 0 & \rightarrow \quad 0 < d_i^* < 1 \end{cases} \qquad \textbf{Eq. 4}$$

Comparison of this characterisation with the one under certainty (equation 3) shows that they are analogous, with the only difference being that now the expected value is used.

When assuming a linear utility function $U=(Y-c)+a^*e$ with $U_1 = 1$ and $U_2 = a$ the first order conditions can be rewritten to:

$$\begin{cases} \dfrac{E(c_i)}{E(e_i)} < a & \rightarrow \quad d_i^* = 1 \\ \dfrac{E(c_i)}{E(e_i)} > a & \rightarrow \quad d_i^* = 0 \\ \dfrac{E(c_i)}{E(e_i)} = a & \rightarrow \quad 0 < d_i^* < 1 \end{cases} \qquad \textbf{Eq. 5}$$

As expected, this means that for a linear utility function one can simply use the expected values of costs and effects to find the same characterisation of the optimum as with a linear value function under certainty. However, such a function would assume risk neutrality as well as no diminishing marginal utility of either money or health effects.

When using more realistic utility functions, assuming decreasing marginal utility or risk aversion, in general $E[e_1U_2 - c_1U_1]$ cannot be restated to a condition on c_1/e_1 and the decision whether or not a new program will be added to an existing set of programs cannot be made by a simple comparison of $E(c_{new})/E(e_{new})$ with a critical ratio.

For certain combinations of utility function and probability density function it is possible to simplify equation 5 further. As an example, suppose the decision maker is constantly risk averse on both money and health effects, and assume mutual utility independence. This is reflected by the following utility function:

$$U(Y-c,e) = k_1(1-\exp^{-\delta(Y-c)}) + k_2(1-\exp^{-\gamma e}) + k_3(1-\exp^{-\gamma e})(1-\exp^{-\delta(Y-c)}) \qquad \textbf{Eq. 6}$$

Furthermore, assume that for each intervention costs and effects follow a bivariate normal distribution. The expected utility $E[U(Y-c,e)]$ can be found using the fact that if $x \sim N(\mu,\sigma^2)$ then $\exp^x$ follows a lognormal distribution and $E(\exp^x) = \exp(\mu + \tfrac{1}{2}\sigma^2)$. Maximising this expected utility we find the following characterisation of the optimum:

$$\begin{cases} R_1^i + R_2^i \geq 0 & \rightarrow \quad d_i^* = 1 \\ R_1^i \leq 0 & \rightarrow \quad d_i^* = 0 \\ else & \rightarrow \quad d_i^* = -R_1^i / R_2^i \end{cases}$$

with

$$R_1^i = -(k_1 + k_3)\exp^{S_1}\delta\mu_{c_i} + (k_2 + k_3)\exp^{S_2}\gamma\mu_{c_i} + k_3\exp^{S_3}(\delta\mu_{c_i} - \gamma\mu_{c_i})$$

$$R_2^i = -(k_1 + k_3)\exp^{S_1}\delta^2\sigma_{c_i}^2 + (k_2 + k_3)\exp^{S_2}\gamma^2\sigma_{c_i}^2 + k_3\exp^{S_3}(\delta^2\sigma_{c_i}^2 + \gamma^2\sigma_{c_i}^2 - 2\gamma\delta\sigma_{ce_i})$$

$$S_1 = -\delta(Y - \sum d_i\mu_{c_i}) + \tfrac{1}{2}\delta^2\sum d_i^2\sigma_{c_i}^2$$

$$S_2 = -\gamma\sum d_i\mu_{c_i} + \tfrac{1}{2}\gamma^2\sum d_i^2\sigma_{e_i}^2$$

$$S_3 = S_1 + S_2 - \delta\gamma\sum d_i^2\sigma_{ce_i}$$

Note that compared to the situation under certainty, the variances and covariances of costs and effects of all programs (the σ's) now appear in these conditions. If the degree of uncertainty becomes very small, the variances converge towards zero, the means converge towards the certain value of the costs or effects, and the conditions become similar to those outlined in the previous section.

Example (continued)

Suppose again that the total budget Y is 2,000,000. Assume that the risk attitude of the decision maker can be described by equation 6 with the following parameters:

k_1 = 0.49*1.0186; δ = 0.000002;
k_2 = 0.51*1.125; γ = 0.00022;
k_3 = 0.

For clarity k_1 and k_2 are written as a product of two factors, a weighing factor, giving relative weight of costs and effects in the utility function and a scaling factor, chosen such that on the domain considered, the utility is between zero and one. It is furthermore assumed that costs and effects follow normal distributions and (to simplify calculations) we will assume that they are not correlated. Table 7.2 presents the parameters of the distributions.

table 7.2 Mean costs and health effects + variances for three fictitious health care programs

Program	μ_c	μ_e	σ_c	σ_e
1	400000	1900	120000	125
2	450000	1800	40000	90
3	300000	1300	70000	110

table 7.3 Changes in optimum budget allocation as uncertainty for program 1 changes

σ_c program 1	Optimum	Total expected costs	Total expected health effects
120,000	$d^*_1=1$, $d^*_2=0.9$, $d^*_3=1$	1,105,000	4820
150,000	$d^*_1=1$, $d^*_2=0.89$, $d^*_3=1$	1,100,500	4802
200,000	$d^*_1=0.97$, $d^*_2=0.89$, $d^*_3=1$	1,088,500	4745
250,000	$d^*_1=0.81$, $d^*_2=1$, $d^*_3=1$	1,074,000	4639
350,000	$d^*_1=0.68$, $d^*_2=1$, $d^*_3=1$	1,022,000	4392

An optimum is achieved for $d_1^* = 1$, $d_2^* = 0.90$, $d_3^* = 1$. In this optimum, the total expected costs are 1,105,000 and the total expected health effects are 4820.

To see how the optimum may change for changes in the degree of uncertainty, we vary the standard deviation of the costs of program 1. Clearly, as the uncertainty about the total costs of program 1 increases, the decision maker is less willing to spend money, which results in lower expected total costs and lower expected health effects. Also, after the standard deviation has passed a threshold, program 1 is implemented only partially, even though its ratio of costs to effects is the lowest among the three programs. It is interesting to see that, for a standard deviation of 200,000, both programs 1 and 2 are implemented partially. Under certainty, this will never be optimal for programs with different ratios of costs to effects.

We also changed the standard deviations of all programs to zero. For that situation, the optimum becomes $d_1^* = 1$, $d_2^* = 0.94$, $d_3^* = 1$, with total expected costs of 1,123,000 and total expected health effects of 4892.

Often decreasing risk aversion is more plausible than constant risk aversion. This risk attitude may be described by a utility function that is a linear combination of exponential functions. For such utility function, in combination with a bivariate normal distribution, an analogous derivation as for equation 6 will yield an algebraic characterisation of the optimum.

To conclude, for the general case of budget allocation under uncertainty using the expected utility approach, the optimal budget allocation not only depends on the ratio of costs to effects and the size of the programs, but also on the (co)variance of costs and effects. As shown in the example above, under uncertainty and risk aversion, it may be optimal that a program with a more favourable ratio than all other programs under consideration is implemented only partially. Under certainty that would never be the case. Under uncertainty, however, this may occur if the variance of costs and effects in this program is large relative to that of the other programs. Being risk averse implies a willingness to give up some effects or budget in return for more certainty on the total effects and budget left after health care spendings.

Since, in the general case, marginal utility is not independent of total costs or effects, the first order conditions cannot be restated to simple conditions in terms of

cost-effectiveness ratios. Hence it is not easily seen from its cost-effectiveness ratio alone whether a new program should be (partially) implemented. A new optimum must be derived, taking into account expected costs and effects of this program and all other programs as well as the (co)variances of costs and effects.

Decisions under uncertainty using the classical approach

The expected utility approach is general and includes most other approaches as special cases. However, this general approach requires the specification of a utility function by the decision maker who may be unwilling or unable to make such specification. Therefore a few alternative approaches are discussed below, since they are the most straightforward extensions towards uncertainty of the optimisation problems presented earlier.

The first approach is similar to the optimisation problem with a fixed budget. Instead of maximising deterministic health effects for a fixed budget, expected health effects are maximised under the constraint that the probability that the total costs will exceed the budget is smaller than a given small percentage, say, 5%. This may be written as:

$$\text{maximise } \sum_{i=1}^{n} d_i E(e_i) \text{ subject to } P\left(\sum_{i=1}^{n} d_i c_i \leq B\right) \geq 0.95$$

If the (e_i, c_i) have a bivariate normal distribution with mean costs μ_{ci}, mean effects μ_{ei}, variance in costs σ^2_{ci}, and variance in effects σ^2_{ei}, then this problem is equivalent to:

$$\text{maximise } \sum_{i=1}^{n} d_i \mu_{c_i} \text{ subject to } z_{0.95}\sigma^c + \mu^c \leq B \ ,$$

where $z_{0.95}$ denotes the 95th percentile of the standard normal distribution,

$$\sigma^c = \sqrt{\sum_{i=1}^{n} d_i^2 \sigma_{c_i}^2} \text{ and } \quad \mu^c = \sum_{i=1}^{n} d_i \mu_{c_i} \ .$$

The first order conditions for optimality are given by:

$$\begin{cases} \dfrac{\mu_{c_i}}{\mu_{e_i}} + \dfrac{2z_{0.95}\sigma_{c_i}}{\mu_{e_i}} \leq \lambda & \rightarrow \quad d_i^* = 1 \\ \dfrac{\mu_{c_i}}{\mu_{e_i}} \geq \lambda & \rightarrow \quad d_i^* = 0 \\ else & \rightarrow \quad 0 < d_i^* < 1 \end{cases}$$

This approach can also be written as the maximisation of a specific expected utility function[4]. Thus, it is a special case of the expected utility formalisation.

The critical ratio λ is found as the shadow price $1/\lambda$ belonging to the constraint that $z_{0,95}\sigma^c + \mu^c \leq B$. This constraint is stricter than a budget constraint on expected costs, as it includes a mark-up to reflect uncertainty. If the costs are more uncertain, the critical ratio is, for the same budget B, lower. It should be noted that this approach includes an implicit assumption about preferences over additional costs and effects under uncertainty.

Example (continued)

Using the information from table 7.2, this approach leads to the following optimum: $d_1^* = 0.26$, $d_2^* = 1$, $d_3^* = 0.34$. In this optimum, the expected total costs are 655,400 and the expected total effects are 2733.

If we compare this to the example using the expected utility approach, it becomes clear that the standard deviations of the costs of the various programs are given more weight than the ratio of costs to effects, indicating a very strong risk aversion.

Note that this approach accounts for the uncertainty in program costs, but only takes account of uncertain effects in so far that expected values are used. Variance of health effects and correlation of costs and effects are ignored. Thus, the decision maker is assumed to be risk neutral with respect to health effects. However, as Arrow & Lind (1970) have argued, it is more rational for the decision maker to be risk neutral towards costs and, since benefits only accrue to some individuals, risk averse towards health effects. This leads to a second approach. We formulate an optimisation problem that reflects risk aversion to effects and risk neutrality to costs. The decision maker minimises expected costs under the constraint that the probability that total effects exceed some aspiration level L is at least, say, 95%. This can be written as:

$$\text{minimise } \sum_{i=1}^{n} d_i E(c_i) \text{ subject to } P\left(\sum_{i=1}^{n} d_i e_i \geq L\right) \geq 0.95$$

The appendix presents the characterisation of the optimum allocation.

---

4. Using the utility function U(Y-c,e)= $\sum_{i=1}^{n} d_i e_i + G(Y-c)$ with

$$G(Y-c) = \begin{cases} 0 & \text{if } P[\sum_{i=1}^{n} d_i c_i - B > 0] - \alpha \leq 0 \\ -\infty & \text{if } P[\sum_{i=1}^{n} d_i c_i - B > 0] - \alpha > 0 \end{cases},$$

expected utility maximisation is equivalent to the problem defined by equation 6.

A third approach is implied by the calculation of confidence intervals (e.g. Chaudhary & Stearns, 1996) and acceptability curves (see chapter 2) for cost-effectiveness ratios. This can be seen as a straightforward extension towards uncertainty of the fixed trade-off formulation. The idea is that a decision maker wants to be reasonably sure that the cost-effectiveness ratio is below a certain limit. There is a clear link with the situation under certainty, where we know for certain whether or not a ratio is smaller than the critical ratio. Under uncertainty, the current ratio is accepted if it is below the critical ratio with a given high probability (instead of with absolute certainty). Like the fixed trade-off approach this is a pragmatic approach, that was developed without reference to an underlying optimisation problem. Analogous to the former approach, here too an implicit assumption is made about preferences over additional costs and effects under uncertainty. Note that, in this situation, the decision maker requires a certain level of certainty for each individual program, instead of for the portfolio of programs together.

Example (continued)

Based on table 7.2, using Fieller's theorem (see Fieller, 1954; Chaudhary & Stearns, 1996), the following 90% two-sided confidence intervals are found for the ratios of costs to effects: program 1 [106, 320]; program 2 [210,294]; program 3 [140,331]. Assuming the decision maker wants 95% certainty that the ratio of costs to effects is below a certain limit, the order in which to implement programs would be: 2, 1, 3. If, again, a threshold of 250 is used, no program would be implemented. This reflects the very risk averse attitude towards both costs and effects implied by requiring 95% certainty of sufficiently low cost-effectiveness ratios.

Discussion

In this chapter the question was raised how a decision can be made between different health care programs with different ratios of costs to effects, specifically when taking account of uncertainty surrounding the estimates of costs and effects. It was suggested that attitudes towards risk should be taken into account as well as the idea of decreasing marginal value of money and health effects. In order to formalise this problem, we first considered the reasoning behind the use of cost-effectiveness ratios as introduced by Weinstein and Zeckhauser (1973). We then extended the standard approach of maximising health effects for a given budget to maximising the value derived from health effects and the budget available for purposes competing with health care spendings.

When a value function is used, adding a new program means that some programs may have to be abandoned (but fewer than with a fixed budget approach) and at the same time that the budget will increase (but not unlimited as with a fixed critical ratio). Furthermore, the critical ratio will decrease. Moreover, it was shown that accepting a program without removing other programs may only be justified when assuming a very specific value function: one without decreasing marginal value for either money or health effects. It is unlikely that any decision maker has such value function, emphasising that budgets determine critical ratios instead of the other way around. Using a value function including health effects as well as money to decide between programs proves particularly fruitful when the ordering of programs of which the outcomes are surrounded with uncertainty is addressed.

We have shown that, if uncertainty is taken into account, the optimum budget allocation may differ from the allocation that would have been derived for the situation with no uncertainty surrounding costs and effects. However, for that purpose we assumed that a utility function for both money and health effects was defined. In practice, it may not be easy for a decision maker to specify preferences over uncertain outcomes so explicitly. Thus, we also addressed some other approaches to handling uncertainty without such an explicit utility function.

The intuitive generalisation of the fixed budget approach is to maximise the expected health effects while staying within budget with some specified probability. However, this has some less favourable implications, because it assumes that the decision maker is risk neutral towards health, but risk averse towards costs. Furthermore, the approach ignores interdependencies between a program's costs and effects. Finally, the approach can be rewritten as a special case of the expected utility approach.

The generalisation of the fixed trade-off formalisation is to derive confidence intervals around cost-effectiveness ratios and implement programs only if the upper limit of the interval is below the critical ratio. These intervals are very informative; however, using them to decide on the implementation of programs may imply rather strong risk aversion towards both costs and effects. One might say that any alternative approach, without assuming an explicit utility function, also has its own implicit utility function. Moreover, these functions are arbitrary as they suggest that all decision

makers have the same preferences over uncertain situations, which is unlikely to be the case. This stresses the need for further research into the risk attitude of decision makers.

It is important to realise that in this chapter we have discussed the characterisation of the optimum for each approach, which is not the same as a decision rule or algorithm to find the optimal allocation. When only independent health care programs are considered, for the fixed critical ratio approach the decision rule follows directly from the optimality conditions, and for the fixed budget approach it is also relatively easy to derive the decision rule from the characterisation of the optimum. However, once the collection of programs also contains some mutually exclusive programs, optimality conditions concern the ratio of costs to effects of a single program whereas the decision rule is concerned with the incremental cost-effectiveness ratio.

The value function approach under certainty and the expected utility approach under uncertainty look very similar. Therefore, it is tempting to compare these two approaches, for instance assuming risk neutrality and then comparing results. However, because both approaches are built on different sets of axioms, the results are not directly comparable. As was pointed out by Dyer & Sarin (1982), in the expected utility approach, the strength of preference and attitude towards risk taking are confounded. Thus, a linear utility function indicating risk neutrality automatically also indicates constant marginal preference. In contrast, the formalisation under certainty had decreasing marginal value.

Some other approaches to the problem of budget allocation under uncertainty have been published. For example, O'Brien & Sculpher (2000) propose formulating the problem of budget allocation under uncertainty as a portfolio selection problem. In portfolio analysis the ultimate goal is to maximise expected return on investment and minimise uncertainty. These two goals will often be conflicting, hence a trade-off needs to be made between expected return on investment and uncertainty. This might be done, for instance, by setting an aspiration level for expected return and minimising uncertainty for that level. A more general approach described by O'Brien and Sculpher is to use a utility function defined on expected return and variance of return. Note that, although a different approach is used here than that presented in this chapter, both methods require that a utility function be defined describing preferences and risk attitude over two attributes. One problem of the approach presented by O'Brien and Sculpher is that it uses the variance of the incremental cost-effectiveness ratio, which will often not be defined.

A paper by Meltzer (2001) uses expected utility to model the choice how much of a specific treatment should be financed. In his model, all money not used for the specific treatment is used for non-medical consumption, and there is a fixed budget constraint. In contrast to our analysis, Meltzer goes on to assume perfect insurance at the population level, for both effects and costs. Then, he finds optimality conditions in terms of the ratio of expected costs to expected effects, similar to those we find for a risk neutral decision maker.

This raises the question whether it is reasonable to assume that decision makers are risk averse and whether they will be risk averse for health costs, health effects, or both. The effects of health programs accrue to a small group of people, for whom it may

affect a large part of their utility whether or not a program is available. Thus, referring to Arrow & Lind (1970), it may be argued that decision makers who use the societal perspective are risk averse for health effects. For other decision makers, e.g. HMO managers or insurers, the argument proposed by Arrow and Lind at the end of their article may apply so that they are also risk averse for costs.

Thus, it becomes clear that in order to achieve optimal budget allocation much information is required: information on costs, effects and size for each health care program and information on goals, preferences and risk attitude of decision makers. While it may not be feasible to collect all the required information on costs, effects and size, the fact that information is required about goals, preferences and risk attitude of decision makers should be seen as an advantage of the described approaches. At this moment, decision makers already weigh information on total costs, total effects and the amount of uncertainty when making decisions, and the approaches in this chapter compel decision makers to make this explicit. It is not the intention that the model prescribes what the decision makers should decide. Rather, the model has two applications. First, it compels decision makers to be explicit about their preferences and informs them about the consequences of certain choices. Second, decisions can be analysed afterwards, and fed back to decision makers to confront them with the risk attitude that is reflected by their choices.

Other attributes, such as equity and ethics, may of course play a role in the decision-making process as well. An advantage of the expected utility approach is that the function may easily be extended to include such attributes. As such, this approach presents a very broad general framework for budget allocation, that better describes reality than the currently assumed model of maximising health effects for a fixed budget.

Acknowledgements

The work of Ms. Al was funded through the Pfizer Health Economics Fellowship Program. The constructive comments of Han Bleichrodt, Mark Koopmanschap and Xander Koolman are gratefully acknowledged.

Appendix

When the goal is to minimise $\sum_{i=1}^{n} d_i \mu_{c_i}$ subject to $\mu^e - z_{0.95}\sigma^e \geq L$

where $z_{0.95}$ denotes the 95th percentile of the standard normal distribution,

$$\sigma^e = \sqrt{\sum_{i=1}^{n} d_i^2 \sigma_{c_i}^2} \quad \text{and} \quad \mu^e = \sum_{i=1}^{n} d_i \mu_{c_i}$$

the first order conditions for optimality are given by:

$$\begin{cases} \dfrac{\mu_{c_i}}{\mu_{c_i}} + \dfrac{2z_{0.95}\sigma_{c_i}}{\mu_{c_i}} \geq \dfrac{1}{\lambda} & \rightarrow \quad d_i = 1 \\ \dfrac{\mu_{c_i}}{\mu_{c_i}} < \dfrac{1}{\lambda} & \rightarrow \quad d_i = 0 \\ else & \rightarrow \quad 0 < d_i < 1 \end{cases}$$

The critical ratio $1/\lambda$ is found as the shadow price belonging to the constraint that. $\mu^e - z_{0.95}\sigma^e \geq L$. If the effects are more uncertain, the variance in effects is larger, and consequently this constraint becomes stricter. The result is a critical ratio that, for the same aspiration level, is higher for larger uncertainty. Note that in this case, a higher critical ratio implies that fewer programs are implemented because the first order conditions are now stated in terms of ratios of expected effects to costs.

CHAPTER 8

Discussion

In this thesis, several methods for dealing with uncertainty in cost-effectiveness studies have been discussed: statistical methods, methods that may be used in modelling studies and methods to explicitly incorporate uncertainty in the decision-making process.

Statistical methods for dealing with uncertainty

In chapter 2, a general approach was discussed to assess the uncertainty surrounding the incremental cost-effectiveness ratio (ICER) in situations where data have been collected at the patient level in a randomised clinical trial. This approach included the calculation of a 95% probability ellipse and the concept of a so-called C/E-acceptability curve was introduced. This latter curve defines for each predefined limit on the ICER the probability that the ICER found in the study is acceptable. Using this curve, a 95% one-sided interval can be derived by finding at which limit the probability that the ICER is acceptable is 95%.

This acceptability curve was presented as an alternative for the confidence interval based on Taylor series that was proposed by O'Brien et al. (1994). Shortly after publication of the acceptability curve method, Wakker & Klaassen (1995) critiqued this method. They stated "It should, however, be understood that acceptability curves are not confidence intervals in a formal statistical sense ...". Further, they presented a formal statistical approach for calculating a confidence interval for the ICER. The method they proposed works as follows: Find a one-sided 97.5% confidence interval for costs and effects separately, with an upper limit for costs and a lower limit for effects. The ratio of these limits gives the upper limit of a 95% one-sided confidence interval for the ICER. Unfortunately, this interval is *at least* a 95% confidence interval, which means that the interval may be much wider than strictly necessary, especially when few data are available. This is, for instance, the case for the data from chapter 2; the interval as proposed by Wakker & Klaassen is ($-\infty$, 224,514] compared with ($-\infty$, 42,000] in chapter 2. Thus, whilst the interval of Wakker & Klaassen is easier to calculate than the acceptability curve, there may be situations where its usefulness for a decision maker is questionable. About the same time, Sacristan et al. (1995) published a paper using Fieller's theorem to calculate a confidence interval for the ICER. In 1954, Fieller presented a formula for the confidence interval of two normally distributed variables. This formula has interesting properties, as there may be situations where the interval is ($-\infty$, ∞), which means that the data are consistent with all possible hypotheses about the ICER. Furthermore, an interval might have the form of ($-\infty$, -1000] $\cup$ [1000, ∞), which will typically happen if the additional costs differ significantly from zero, while the effects in both groups are approximately equivalent.

In 1996, Chaudhary & Stearns also proposed the use of Fieller's theorem but, as an alternative, they also proposed a simulation method that does not make assumptions about the distribution of costs and effects, i.e. the bootstrap method (see Efron & Tibshirani, 1993). In their paper, they illustrated both methods with data on 300 observations per group. For their data, Fieller's method and the bootstrap method produced similar results. The same was found in a study by Polsky et al. (1997) where Fieller's method was compared with the bootstrap method, the Taylor series method

and the so-called box method (O'Brien et al., 1994). In this comparison, Fieller's method and the bootstrap method performed equally well, and much better than the other two methods.

Despite that since the introduction of the acceptability curve other (more formal) methods for quantifying uncertainty have been published, the acceptability curve is widely used (e.g. Gray et al., 2000; Moayyedi et al., 2000; Griffin et al., 2001; Chancellor et al., 2001). An advantage of the acceptability curve over confidence intervals is that it does not focus on 95% confidence, but basically gives an infinite number of x% one-sided confidence intervals (for each value x, the corresponding limit on the ICER can be found in the curve). As Briggs & Fenn (1998) wrote: "We believe that confidence surfaces on the cost-effectiveness plane, summarised in the form of cost-effectiveness acceptability curves, give greater guidance for decision-makers than confidence intervals around cost-effectiveness ratios". Several authors have pointed out that the acceptability curve does not have the interpretation of a confidence interval within the frequentist framework (Wakker & Klaassen, 1995; Heitjan et al., 1999). However, as also mentioned in chapter 4 (and by Briggs, 1999; Heitjan et al., 1999), within the Bayesian framework, the acceptability curve does present posterior confidence intervals under the assumption of a non-informative prior distribution. Furthermore, Löthgren & Zethraeus (2000) showed that the acceptability curve also has a net benefit definition, in that it represents the probability that the net benefits estimator ($NB(\lambda) = \lambda\mu_{\Delta E} - \mu_{\Delta C}$, with λ the limit on the ICER) is positive. With this interpretation, the curve may be used for testing the null hypothesis that the net benefit is non-positive. Consequently, it may be stated that the acceptability curve has both an interpretation within the frequentist and the Bayesian framework.

In chapter 2 the acceptability curve was drawn based on simulated values from a bivariate normal distribution for costs and effects but, alternatively, it may also be drawn based on the simulated values from a bootstrap procedure. Furthermore, it can be used when outcomes are simulated in a probabilistic sensitivity analysis, as was done in chapter 5. It should be noted that in the latter case the curve does not have an interpretation in a frequentist framework, when parameters and their distribution are for instance postulated based on expert opinion. However, as this may be seen as defining a prior distribution for the parameter, within the Bayesian framework curves based on a probabilistic sensitivity analysis do have an interpretation. The curve was also used in a paper by Hunink et al. (1998) in which they use it to present the results of a Markov model that was used to simulate individual patient data. In their paper, Hunnink et al. also emphasise another advantage of the acceptability curve compared with a confidence interval for an ICER. The problem with presenting ICERs is that information is lost. A negative ICER can indicate that there are increased effects plus cost savings or reduced effects plus additional costs. While in the first situation the new treatment would be accepted, in the latter it would be rejected. The same is true for a positive ratio. Assuming that R is the maximum acceptable ratio, in case of increased effects the ICER should be smaller than R to be acceptable, while in case of a loss of effects the ICER should be larger than R. When using an acceptability curve as in chapter 2, i.e. subdivided into curves for each quadrant of the cost-effectiveness plane,

no important information is lost. Clearly, the acceptability curve is a valuable addition to the methods available to describe uncertainty around cost-effectiveness ratios.

Chapter 3 presented a simulation method that enables the calculation of sample sizes for economic evaluations. This method uses a backward procedure, in the sense that for various sample sizes, the power is estimated and thus an estimate of the required sample size for a certain power is found. After the article on which chapter 3 is based was published, several algebraic methods for sample size calculation have been published (Briggs & Gray, 1998b; Laska et al., 1999; Willan & O'Brien, 1999). An advantage of these methods over the simulation method is that they are easier to apply. However, the simulation method also has some advantages. First, it is easy to change the hypothesis being tested, i.e. it is, for instance, possible to use a different threshold ICER in the area where the additional effectiveness is negative. This would reflect a willingness-to-accept a loss of effects that differs from the willingness-to-pay for gain of effects (see O'Brien & Gafni, 1996). Second, it is relatively easy to extend the simulation procedure such that instead of simulating costs and effects per patient from one overall bivariate distribution, values are simulated from components that are easier to specify. For instance, instead of simulating whether or not a patient is an event-free survivor, it is simulated whether the patient had balloon angioplasty, bypass surgery, a myocardial infarction, etc., each with well-known costs. This method also automatically generates the correlation between costs and effects.

When dealing with an economic evaluation that is 'piggy backed' to a phase III clinical trial, basing the sample size on economic data as well as clinical data may be deemed unethical as it will often increase the number of patients that has to be subjected to an experimental treatment. When the goal of a study is specifically to demonstrate cost-effectiveness after superior efficacy has been established, it may again be considered unethical to subject patients to a control treatment that has been proven to be less effective. Consequently, the number of situations where it is possible to base the sample size of a study on cost-effectiveness considerations will be limited, also because of the large number of parameters that need to be specified. As indicated in chapter 3, besides information on expected effects and their standard deviations, the same information for costs and the correlation between costs and effects is required. The conclusion of chapter 3 was, therefore, that given that such information is often incomplete there is a danger that these sample size calculations may be manipulated such that projected patient enrolment is suggested to be both sufficient and operationally feasible. The question remains, of course, whose interest would be served by such manipulations.

In chapter 4, a Bayesian approach to cost-effectiveness studies was presented, consisting of a univariate approach for the analysis of the differences in costs and effects and a multivariate approach for the analysis of the ICER. These approaches were applied to datasets of two clinical trials which compared both the costs and effects of stent implantation versus balloon angioplasty. In the multivariate analysis, several prior distributions were defined for the analysis of the second trial and the influence

of these priors on the posterior distribution was studied. The point estimates of the ICER varied from NLG 16,269 per additional event-free survivor to NLG 21,984. If the limit that society puts on the ICER is about NLG 30,000, the probabilities that the ICERs are acceptable differ markedly (65%, 80%, 90%), possibly leading to different decisions.

The multivariate model presented in chapter 4 relies on a normal approximation of the mean of a sample of costs and effects, which may not always be realistic. A recent paper by O'Hagan & Stevens (2001a) describes the full model for the same type of data as in chapter 4, i.e. a dichotomous distribution for effects and a log normal distribution for the costs. The posterior distribution was evaluated using Monte Carlo Markov chain simulation, since a simple algebraic representation of the posterior distribution is no longer available when the full model is used. In their analysis O'Hagan & Stevens used interesting informative prior distributions. For the probability of 'success' in both treatment groups a positive correlation was assumed, indicating that higher effectiveness in one group is also likely to imply higher effectiveness in the other group. Furthermore, for the variances of the costs a prior distribution was postulated that gives weak prior information about individual variances but such that the ratios of variances are likely to be reasonably close to unity.

The interest of health economists in Bayesian methods may partly be explained by the many papers that have recently been published advocating the use of Bayesian methods in biostatistics. The paper by Lilford & Braunholtz (1996) had as subtitle "A paradigm shift is overdue". At this moment, it appears that this paradigm shift may actually be taking place. One indicator is a recently published article in The Economist (2000), with the title "In praise of Bayes". This article describes not only the differences between the Bayesian and frequentist paradigm, but also cites a statistician (at the pharmaceutical company Pfizer), who describes how Bayes may be used in dose-allocation trials. The idea is that, instead of analysing the results at the end of the trial, patients' responses will be evaluated during the trial and where necessary the doses will be adjusted. With such an approach, more patients will receive (near) optimal doses, and consequently fewer patients may be required for such studies.

As discussed in chapter 4, it is the issue of using prior information that can be seen as the largest obstacle in a broader use of Bayesian methods. It should be realised, however, that this problem is not unique to Bayesian statistics: a frequentist meta-analysis also requires some judgement on which data to include. It seems that the time has come to shift from discussing the advantages of using Bayes to actually applying the methods in cost-effectiveness studies in order to gain experience, allowing us to update our prior beliefs about the possibilities of Bayes.

It is important to note that in all three chapters on statistical methods for dealing with uncertainty it was assumed that the average costs and effects would approximately follow a normal distribution, for large enough sample sizes, based on the Central Limit Theorem. A recent article by O'Hagan & Stevens (2001b) makes clear that such an assumption should not be made too easily. It is well known that the sample average is not a robust estimator for the population mean, which means that if the underlying distribution is very skewed, we are likely to overestimate the population mean when

using this normal approximation. O'Hagan & Stevens also show that using bootstrap simulation instead is not a solution because the bootstrap method is also based on approximation. When the sample size is relatively small and skewness is large, a modelling approach as presented in chapter 4 (univariate analysis) should be used. However, as also mentioned in the discussion of chapter 4, it is important to check the adequacy of the fit of the model to the data. Especially with a Bayesian approach, testing hypotheses about different models (model refers here both to the likelihood function and the prior distribution) is fairly straightforward, with no requirement about hypotheses being nested and no limit on the number of hypotheses considered simultaneously.

The example presented by O'Hagan & Stevens (2001b) clearly shows that for a very skewed distribution, even a sample size of 75 is too small to rely on the normal approximation. However, in the case studies presented in chapters 2 and 4 this problem does not occur. In chapter 4, the estimated average costs using the log normal model are similar to the sample mean. Similarly, the costs in chapter 2 are far less skewed than in the example given by O'Hagan & Stevens; thus, even with the small sample sizes assuming normality does not have major consequences (assuming log normality leads to very similar results as those presented in chapter 2). Nevertheless, there are many situations with large skewness, where formulating a parametric model (i.e. a probability distribution) for the data should be preferred over normal approximation or bootstrapping.

Dealing with uncertainty in modelling studies

Chapters 5 and 6 presented examples of dealing with uncertainty in modelling studies. In chapter 5 the cost-effectiveness of a fixed dose combination of diclofenac/misoprostol compared with diclofenac monotherapy in patients with rheumatoid arthritis was assessed. Diclofenac is a non-steroidal anti-inflammatory drug (NSAID) and misoprostol is a gastroprotective agent that reduces the probability of NSAID-induced gastro-intestinal side-effects. A model was used to convene data from literature and databases, and expert opinion. The baseline result of the study was NLG 4,179 per life-year gained for the use of the combination drug, which may be considered cost-effective. Parts of the analysis were a univariate and a probabilistic sensitivity analysis. In the univariate analysis, all variables were varied (one at a time) by plus and minus 20% and results were considered not sensitive to changes if the outcome changed by less than 10%. This criterion is of course arbitrary, and it adds to the problem of interpreting the results of such univariate analyses. The results presented in chapter 5 identify those parameters for which the results are most sensitive. However, as it is not clear how relevant the upper and lower limits of ± 20% are, it is also not clear how relevant the ordering is of the results. Another approach would have been to use limits (where possible) derived from the confidence intervals that were also used in the probabilistic analysis. As a criterion for sensitivity a threshold ICER might be used: if changing a parameter to its upper or lower limit causes the outcome to exceed the threshold, implying a different decision, the outcome is called sensitive to changes in the parameter. However, such a criterion is not ideal,

as it dichotomises the idea of sensitivity. Assuming that realistic distributions have been defined (where possible based on actual trial data), the results of a probabilistic analysis are easier to interpret: they indicate how likely all possible outcomes are. This is an important advantage above other approaches such as univariate analysis, where outcomes are calculated for various input values without any indication how likely these outcomes are.

Performing a probabilistic sensitivity analysis has become easier in recent years, as dedicated software is now available for such analysis, e.g. the decision analysis program DATA (*www.treeage.com*) and the risk analysis program @RISK (which is used together with a spreadsheet program; *www.palisade.com*). An interesting feature of @RISK is that it also allows for an assessment of the influence of each individual input parameter on the outcome, adjusting for all other input parameters. Given the information that it provides and the relative ease of its use, it is recommended that modelling studies include a probabilistic analysis in their sensitivity analysis.

As discussed in chapter 3, not all uncertainties in health care are of a statistical nature. This became apparent shortly after the publication of the article on which chapter 5 is based. In 1996, the Ministry of Health, Welfare and Sport introduced new pricing regulations for pharmaceuticals, allowing no price that would exceed the average price of four neighbouring countries. As a result the price of diclofenac decreased by 50%, whilst the price of the fixed dose combination of diclofenac/misoprostol remained unchanged. Such an extreme scenario had not been included in the sensitivity analysis. Because of this, a new cost-effectiveness study for diclofenac/misoprostol was done (see Al, 2000; Brouwers et al., 2000). In this new study, other comparators were also included since clinical practice has also changed in the last few years. Nowadays, many patients using NSAIDs long-term receive proton pump inhibitors as gastroprotective agent (Herings et al., 2000). As expected, the new study showed that treatment with diclofenac/misoprostol was no longer cost-effective compared with NSAID for the whole rheumatoid arthritis population, but only for patients at medium to high risk for developing gastro-intestinal side-effects.

In chapter 6 a modelling study was presented that assessed the costs and effects of lung transplantation. Actual data were available for the situation concerning transplantation, and an extensive simulation model was used to construct a control group without transplantation and to extrapolate the available data to the future. The analysis showed that the costs per quality-adjusted life-year (QALY) gained would be NLG 167,000 (after discounting by 5% per year). The model was subsequently used to assess additional scenarios: a donor scenario in which a larger number of donor lungs available was assumed (NLG 135,000 per QALY gained) and a policy scenario with a restricted patient inflow (NLG 144,000 per QALY gained). Based on the observations during the study period, the number of donor lungs available was estimated at 17 per year; thus, this estimate is surrounded by a certain amount of statistical uncertainty. However, the availability of donor lungs may also be influenced by decision makers (e.g. through legislation) and physicians (e.g. through changes in donor treatment and in selection criteria used for donor lungs). A separate study (Geertsma & ten Vergert,

1995) showed that with concerted action a maximum of 27 donor lungs per year might be achieved. The same applies for the patient inflow, the definition of stricter inclusion criteria might restrict the number of patients eligible for screening per year, thus decreasing the inflow beyond the statistical limits.

The model used in the present study is a micro-simulation model, which means that data are simulated at the patient level: for each patient a history is created, with time spent in each phase, until death. It might have been possible to use a Markov model instead, where patients move from one state (in this study: phase) to another on fixed moments. However, because of the large variation of time spent in each phase, it would have been necessary to use a cycle of 1 day, or 1 week at most. Given that the total evaluation period was 40 years and that the transition probabilities are time-dependent, a micro-simulation model was in fact easier to build and work with. Moreover, it would have proven very difficult to build the interaction between donor lungs and waiting list into a Markov model, while it is straightforward in the micro-simulation model.

When using a model for scenario analyses, it is recommended that scenarios are defined before developing the model. When the scenarios are defined afterwards, they may concern aspects that were not included in the model, making it difficult to address them without altering the structure of the model. Furthermore, when defining the scenarios the availability of data should be kept in mind. In the lung transplantation study, it would have been interesting to check whether cost-effectiveness is more favourable for certain patient groups. However, the small number of patients per group would have made the results highly unreliable.

Dealing with uncertainty in decision making

The purpose of developing methods to describe or reduce uncertainty, such as those presented in this thesis, is to facilitate decision making under uncertainty. However, these methods present only one part of the equation, as it is also necessary to describe preferences for possible outcomes before a decision can be made. In chapter 7 attention was given to risk attitude and preferences of decision makers. It has been argued that taking uncertainty into account in budget allocation decisions leads to different decisions than under certainty. A numerical example made it clear that under uncertainty a trade-off is made between the ratio of costs to effects and the degree of uncertainty. When a decision maker is risk averse, programs with small ratios of costs to effects but large variances may be abandoned in favour of programs with higher ratios of costs to effects but smaller variances.

The model presented in chapter 7 is a static model, in the sense that the budget of one year is allocated without regards for the years thereafter. This is an unrealistic situation because in practice, budget allocation in one year may influence the budget allocation in succeeding years. Thus, a logical next step in the development of budget allocation models is to adjust the current model to a stochastic dynamic model.

When decision makers have difficulty specifying the utility function, a risk premium on the riskless discount rate might be used to adjust the value of the mean outcome. This is what is proposed by Arrow & Lind (1970), and also suggested as an alternative

by Ben-Zion & Gafni (1983). Such risk premium should reflect how much risk is involved, i.e. it should be large if the variance is large and small if the variance is small (assuming risk aversion). Though this method may seem easier to use in practice, eliciting a value for a risk premium in such a way that it accurately reflects risk attitude may in practice be rather complicated. Furthermore, adjusting the riskless discount rate for risk assumes that our uncertainty is completely time-related, i.e. that we are sure about the outcome today, and become more and more uncertain as time progresses. The expected utility approach does not relate uncertainty to time, and therefore seems more pertinent to the type of uncertainty that has been discussed in this thesis and is generally discussed in health economics.

It should be noted that there are several (competing) theories on whose risk attitude (utility function) should be used. Welfarists will argue that for each individual his or her own risk attitude should be used, whereas extra-welfarists suggest that the risk attitude of society as a whole should be used. Importantly, they all agree that someone's risk attitude should be included in the decision-making process.

Another reason (besides risk attitude) why information on uncertainty is relevant to the decision maker are the so-called 'sunk costs'. If decision makers disregard uncertainty and base their decision only on the point estimate of the ICER, a wrong decision may be made. If the decision was to introduce a new program, and it is later found that the actual ICER is much higher, the decision maker might want to switch back to the old situation. However, switching is not without cost. Examples of sunk costs are the non-recoverable investment costs, such as the alteration of a building or equipment specific to the new technology. While some of these costs (such as the costs of special equipment) are usually taken into account in an economic evaluation, some are not. Not taking such cost categories into account is often a consequence of the choice to present the results independently from the initial situation. Examples of costs not taken into account are implementation costs, such as the hiring and training of personnel, and costs related to the loss of goodwill among the health care professionals as a result of frequent changes in treatment strategies. Another example are the "political costs". In practice, decision makers make decisions on which medical technology to add to the benefits package covered by social health insurance but it proves to be much more difficult to decide, based on cost-effectiveness arguments, to withdraw a currently available technology from the social insurance coverage. Presumably, the "political costs" of a policy reversal are very high. Sunk costs are often neglected as they are often intangible and rare (since we only switch if we have a highly reliably estimate of the ICER). If, on the other hand, no attention is given to the uncertainty surrounding the ICER, the probability of a policy reversal might be high, meaning that sunk costs can become very relevant.

Claxton (1999b) discussed this idea of sunk costs in the context of deciding whether or not to collect additional data in order to reduce uncertainty. If all sunk costs that may be involved in a decision can be quantified, it is possible to calculate whether or not a new program should be introduced based on the current information and expected results of a new study. However, even if the decision maker decides not to

start a new trial it is still likely that new information will become available through other channels (literature, study initiated by a hospital, etc.). Clearly, the later such information is expected to become available, the smaller the influence of the sunk costs on the current decision. A decision maker may not be able to control which information becomes available at what time, but can (and should) consider what the likelihood is of making the wrong decision by changing to the new technology. Confidence intervals and acceptability curves can provide that information. By combining his/her knowledge or judgement (in case the costs can not be quantified) about the magnitude of the sunk costs related to making the wrong decision with the likelihood of that event occurring, the decision maker will then decide if the ICER found in the current study should result in a switch from the old to the new technology.

An alternative method for dealing with uncertainty in economic evaluation is based on the real option approach (Dixit & Pindyck, 1994; Smit, 1996; Merton, 1998). This approach originates from the field of finance, where it has been successfully applied to explain why businesses do not behave as they are expected to on the basis of the net present value of an investment. Palmer & Smith (2000) discussed the idea of using option values in economic evaluations in health care.

Similar to the valuation of financial options, the real option approach assigns values to the options the decision maker holds. Real options hold value as they allow the decision maker to adapt and revise decisions in response to new information.

The real option approach assumes three features for a decision: first, the ability to delay the decision to introduce the new technology; second, that there is uncertainty about the future state of the world (which will be reduced over time); and finally that decisions involve sunk costs (i.e. the decision is at least partly irreversible). If any of the three features does not apply, the option has no value, and can therefore be discarded. Several types of options may be distinguished, two of which are the option to expand and the option to defer.

The option to expand may be relevant when a decision to implement a new technology acquires an expansion option, i.e. a possibility to expand at relatively low costs. A new technology may, for example, yield a small negative net benefit that because of great variation could well be coincidental. However, implementing this technology on a small scale acquires the option to expand when new evidence proves that the new technology is cost-effective. If the value of this option is higher than the expected loss of fully implementing the program, then it is wise to implement the program on a small scale.

The option to defer takes into account the opportunity costs of making a commitment now, and thereby giving up the option of waiting for new information. The option has value because it enables the decision maker to respond to new information at low cost. This new information will be valuable if returning on a decision involves sunk costs and the likelihood of this occurring is large. In case the new technology has a positive net benefit, but the value of the option to defer is larger, it may not be wise to implement the new technology at the current time. The value of the option is, for instance, related to the magnitude of uncertainty (the larger, the

higher the value of the option), the expected informational content (the larger, the higher the value of the option), the speed with which the new information will become available (the quicker, the higher the value of the option) and the discount rate (the lower, the higher the value of the option).

The option approach adjusts the ICER for both the degree of uncertainty and the reversibility of a decision. It is not immediately clear to what extent the option value approach and expected utility approach describe the same properties in decision making under uncertainty. In the set of axioms of expected utility, no notion of irreversibility occurs. Therefore, it would be interesting to compare both approaches to see if it is necessary and possible to combine both approaches.

Areas for further research

In this thesis, several sources of uncertainty have been discussed. One that has not yet been addressed is missing/censored data. When data are missing, uncertainty about the parameters of interest increases. When the data are missing completely at random, it is not difficult to adjust the estimates for that. However, if missingness depends on observed or unobserved data, special methods are required to adjust both the point estimate and the estimate of the variance. This problem of missing data is not new, and many methods for dealing with it are known (see Little & Rubin, 1987; Schafer, 1997). In health economics, the problem of dealing with missing data has not yet received much attention. When longitudinal data on quality-of-life are missing from a certain time point, it may be plausible to assume that patients have become too ill to fill out the questionnaire, and therefore it may not be assumed that the missingness is ignorable. The same is true for costs; it is possible that patients dropped out of a study due to serious side-effects causing high costs, which remain unobserved. However, it is unlikely that all methods that work well for quality-of-life data will work equally well for cost data, given that costs are usually very skewed, whereas quality-of-life data are often approximately normal. Both for costs and effects it is important to investigate how different methods for dealing with missing data perform for different types of missingness. This should provide some guidance for analysts in their choice between available methods.

Chapter 7 deals with the influence of uncertainty in decision making. From a theoretical point of view, reasons may be given why decision makers should take uncertainty about costs and effects into account when making their decisions. However, we have little knowledge of whether decision makers actually do take uncertainty into account and if so, in what way. These questions might be answered through an empirical study among decision makers.

Epilogue

In the introduction it was stated that it is unrealistic to assume that all uncertainty can be eliminated. However, this should not prevent us from making optimal use of all methods available to first reduce uncertainty and then describe the remaining uncertainty. Unfortunately, it is clear that these goals have not yet been reached in most economic evaluations of health care technologies; however, this is not surprising because most methodological developments in this area are quite recent. Hopefully, the coming years will see a wide dissemination of methods for dealing with uncertainty and a growing understanding of preferences over uncertain decisions.

References

Agro, KE, Bradley, CA, et al. (1997). Sensitivity analysis in health economics and pharmacoeconomics; an appraisal of the literature. *PharmacoEconomics* 11, 75-88.

Al, MJ (2000). Kosten en effecten van behandeling met Arthrotec. Rotterdam, Institute for Medical Technology Assessment.

Armstrong, CP and Blower, AL (1987). Non-steroidal anti-inflammatory drugs and life threatening complications of peptic ulceration. *Gut* 28, 527-532.

Arrow, K and Lind, A (1970). Uncertainty and the evaluation of public investment decisions. *American Economic Review* 60, 364-378.

Ballegooijen, M van, Habbema, JDF, et al. (1992). Preventive pap-smear: balancing costs, risks and benefits. *British Journal of Cancer* 65, 903-933.

Bardhan, KD, Hawkey, CJ, et al. (1995). Lansoprazole versus ranitidine for the treatment of reflux oesophagitis. *Alimentary Pharmacology & Therapeutics* 9, 145-151.

Barradell, LB, Whittington, R, et al. (1993). Misoprostol: pharmacoeconomics of its use as prophylaxis against gastroduodenal damage induced by nonsteroidal anti-inflammatory drugs. *PharmacoEconomics* 3, 140-170.

Ben-Zion, U and Gafni, A (1983). Evaluation of public investment in health care: is the risk relevant? *Journal of Health Economics* 2, 161-165.

Berge Henegouwen, GP van and Bijlsma, JWJ (1991). Therapeutic policy for NSAID-induced ulcers [in Dutch]. *Nederlands Tijdschrift voor Geneeskunde* 135, 504-508.

Birch, S and Donaldson, C (1987). Applications of cost-benefit analysis to health care: departures from welfare economic theory. *Journal of Health Economics* 6, 211-225.

Bjork, S (1992). *EuroQol Conference Proceedings, Lund, October 1991.* IHE Working Paper, Lund.

Black, WC (1990). The CE-plane: a graphic representation of cost-effectiveness. *Medical Decision Making* 10, 212-240.

Boer, MJ de, Hoorntjes, JCA, et al. (1994). Immediate coronary angioplasty versus intravenous streptokinase in acute myocardial infarction: left ventricular ejection fraction, hospital mortality and reinfarction. *Journal of the American College of Cardiology* 23, 1004-1008.

Boer, MJ de, Hout, BA van, et al. (1995). A cost-effective analysis of primary coronary angioplasty versus thrombolysis for acute myocardial infarction. *American Journal of Cardiology* 76, 830-833.

Boer, WJ de and Mannes, GPM (1993). *Behandelingsprotocol longtransplantatie.* University Hospital Groningen, Groningen.

Bolten, W, Melo Gomes, JA, et al. (1992). The gastroduodenal safety and efficacy of the fixed combination of diclofenac and misoprostol in the treatment of osteoarthritis. *British Journal of Rheumatology* 31, 753-758.

Box, GEP and Tiao, GC (1973). *Bayesian inference in statistical analysis.* Addison-Wesley, Reading, Mass.

Bratley, P, Fox, BL, et al. (1987). *A guide to simulation.* Springer Verlag, New York.

Breslow, N (1989). Biostatistics and Bayes. *Statistical Science* 5, 269-298.

Briggs, AH (1999). A Bayesian approach to stochastic cost-effectiveness analysis. *Health Economics* 8, 257-261.

Briggs, A and Fenn, P (1998). Confidence intervals or surfaces? Uncertainty on the cost-effectiveness plane. *Health Economics* 7, 723-740.

Briggs, A and Gray, A (1998a). The distribution of health care costs and their statistical analysis for economic evaluation. *Journal of Health Services Research & Policy* 3, 233-245.

Briggs, AH and Gray, AM (1998b). Power and sample size calculations for stochastic cost-effectiveness analysis. *Medical Decision Making* 18 suppl., S81-S92.

Briggs, A and Sculpher, M (1995). Sensitivity analysis in economic evaluation: a review of published studies. *Health Economics* 4, 355-371.

Briggs, AH, Wonderling, DE, et al. (1997). Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation. *Health Economics* 6, 327-340.

Brophy, JM and Joseph, L (1995). Placing trials in context using Bayesian analysis; GUSTO revisited by reverend Bayes. *Journal of the American Medical Association* 273, 871-875.

Brouwers, JRBJ, van Roon, EN, et al. (2000). Kosteneffectiviteit van NSAID's. Altijd gastroprotectie? *Pharmaceutisch Weekblad* 135, 172-175.

Carlin, BP and Louis, TA (1996). *Bayes and emperical Bayes methods for data analysis.* Chapman & Hall, London.

Centraal begeleidingsorgaan voor de intercollegiale toetsing (1998). Behandeling en preventie van coronaire hartziekte door verlaging van de plasmacholesterolconcentratie. Utrecht.

Chancellor, JVM, Hunsche, E, et al. (2001). Economic evaluation of celecoxib, a new cyclo-oxygenase 2 specific inhibitor, in Switzerland. *PharmacoEconomics* 19, Suppl. 1, 59-75.

Chaudhary, MA and Stearns, SC (1996). Estimating confidence intervals for cost-effectiveness ratios: an example from a randomized trial. *Statistics in Medicine* 15, 1447-1458.

Claxton, K (1999a). Bayesian approaches to the value of information: implications for the regulation of new pharmaceuticals. *Health Economics* 8, 269-274.

Claxton, K (1999b). The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics* 18, 341-364.

Cohen, J (1997). *Statistical power analysis for the behavioral sciences.* Academic Press, New York.

Commonwealth of Australia (1992). *Guidelines for the pharmaceutical industry on preparation of sub-missions to the Pharmaceutical Benefits Advisory Committee: including submissions involving economic analyses.* Australian Government Printing Office, Canberra.

COTG (Central Organisation on Tariffs in Health Care) (1995). Vademecum [in Dutch]. Alphen aan de Rijn.

De Nederlandse Lever Darm Stichting (1992). Spijsverteringsziekten in Nederland, 1970-1989. Breukelen.

Dixit, AK and Pindyck, RS (1994). *Investment under uncertainty.* Princeton University Press, Princeton, New Jersey.

Doubilet, P, Begg, CB, et al. (1985). Probabilistic sensitivity analysis using Monte Carlo simulation. *Medical Decision Making* 5, 157-177.

Drummond, MF, Bloom, BS, et al. (1992). Issues in the cross-national assessment of health technology. *International Journal of Technology Assessment in Health Care* 8, 671-682.

Drummond, MF, O'Brien, B, et al. (1997). *Methods for the evaluation of health care programmes.* Oxford University Press, Oxford.

Duan, N, Manning, WG, et al. (1983). A comparison of alternative methods for the demand for medical care. *Journal of Business & Economic Statistics* 1, 115-126.

Dutch Health Insurance Funds Council (1995). Farmacotherapeutisch Kompas 1995. Amstelveen.

Dyer, JS and Sarin, RK (1982). Relative risk aversion. *Management Science* 28, 875-885.

Eddy, DM, Hasselblad, V, et al. (1990). A Bayesian method for synthesizing evidence: the Confidence Profile Method. *International Journal of Technology Assessment in Health Care* 6, 31-55.

Edelson, JT, Tosteson, ANA, et al. (1990). Cost-effectiveness of misoprostol for prophylaxis against nonsteroidal anti-inflammatory drug-induced gastrointestinal tract bleeding. *Journal of the American Medical Association* 264, 41-47.

Eerste Kamer der Staten-Generaal (1995-1996). Law on the organ donation (in Dutch). 22358, nr. 46b.

Efron, B and Tibshirani, RJ (1993). *An introduction to the bootstrap.* Chapman and Hall, New York.

Enckevort, PJ van, Koopmanschap, MA, et al. (1997). Lifetime costs of lung transplantation: estimation of incremental costs. *Health Economics* 6, 479-489.

Eversmeyer, W, Poland, M, et al. (1993). Safety experience with nabumetone versus diclofenac, naproxen, ibuprofen, and piroxicam in osteoarthritis and rheumatoid arthritis. *American Journal of Medicine* 95 (suppl 2A), 10S-18S.

Farah, D, Sturrock, RD, et al. (1988). Peptic ulcer in rheumatoid arthritis. *Annals of Rheumatic Disease* 47, 478-480.

Fieller, EC (1932). The distribution of the index in a normal bivariate population. *Biometrika* 24, 428-440.

Fieller, EC (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society Series B* 16, 175-185.

Fisher, CJ, Slotman, GJ, et al. (1994). Initial evaluation of human recombinant interleukin-1 receptor antagonist in the treatment of sepsis syndrome: A randomized, open-label, placebo-controlled multicenter trial. *Critical Care Medicine* 22, 12-21.

Fries, JF (1991). NSAID gastropathy: the second most deadly rheumatic disease? Epidemiology and risk appraisal. *Journal of Rheumatology* 18 (suppl 28), 6-10.

Fries, JF, Williams, CA, et al. (1991). Nonsteroidal anti-inflammatory drug-associated gastropathy: incidence and risk factor models. *American Journal of Medicine* 91, 213-222.

Gabriel, SE, Jaakkimainen, RL, et al. (1993). The cost-effectiveness of misoprostol for nonsteroidal antiinflammatory drug-associated adverse gastrointestinal events. *Arthritis and Rheumatism* 36, 447-459.

Gafni, A and Birch, S (1993). Guidelines for the adoption of new technologies: a prescription for uncontrolled growth in expenditures and how to avoid the problem. *Canadian Medical Association Journal* 148, 913-917.

Geary, RC (1930). The frequency distribution of the quotient of two normal variates. *Journal of the Royal Statistical Society* 93, 442-446.

Geertsma, A and ten Vergert, EM (1995). Behoefte aan longtransplantatie, in E ten Vergert, ed., *Evaluatie-onderzoek longtransplantatie.* Groningen, Academisch Ziekenhuis Groningen.

Geis, GS (1992). Overall safety of Arthrotec®. *Scandinavian Journal of Rheumatology* Suppl. 96, 33-36.

Gelman, A, Carlin, JB, et al. (1995). *Bayesian Data Analysis.* Chapman & Hall, London.

Giercksky, KE, Huseby, G, et al. (1989). Epidemiology of NSAID-related gastrointestinal side effects. *Scandinavian Journal of Gastroenterology* 24 (suppl 163), 3-8.

Gnedenko, BV (1968). *The theory of probability.* Chelsea Publishing Company, New York.

Gordon, GS, Fisher, CJ, et al. (1992). Cost-effectiveness of treatment with Interleukin-1 receptor antagonist (IL-Ira) in patients with sepsis syndrome. *Clinical Research* 40, 254A.

Graham, DY, Agrawal, NM, et al. (1988). Prevention of NSAID-induced gastric ulcer with misoprostol: multicentre, double-blind, placebo-controlled trial. *Lancet* 2, 1277-1280.

Graham, DY, White, RH, et al. (1993). Duodenal and gastric ulcer prevention with misoprostol in arthritis patients taking NSAIDs. *Annals of Internal Medicine* 119, 257-262.

Gray, A, Raikou, M, et al. (2000). Cost effectiveness of an intensive blood glucose control policy in patients with type 2 diabetes: economic analysis alongside randomised controlled trial (UKPDS 41). *British Medical Journal* 320, 1373-1378.

Griffin, AD, Perry, AS, et al. (2001). Cost-effectiveness analysis of inhaled zanamivir in the treatment of influenza A and B in high-risk patients. *PharmacoEconomics* 19, 293-301.

Heitjan, DF, Moskowitz, AJ, et al. (1999). Bayesian estimation of cost-effectiveness ratios from clinical trials. *Health Economics* 8, 191-201.

Herings, RMC, Panneman, MJM, et al. (2000). Farmacotherapie in beweging. Utrecht, PHARMO Instituut.

Hilgeman, CN, Koopmanschap, MA, et al. (1992). *Costs of intensive care in The Netherlands (in Dutch).* Erasmus University Rotterdam.

Hillman, AL and Bloom, BS (1989). Economic effects of prophylactic use of misoprostol to prevent gastric ulcer in patients taking nonsteroidal anti-inflammatory drugs. *Archives of Internal Medicine* 149, 2061-2065.

Hosenpud, JD, Novick, RJ, et al. (1995). The registry of the international society for heart and lung transplantation; twelfth official report 1995. *Journal of Heart and Lung Transplantation* 14, 805-815.

Hout, BA van, Al, MJ, et al. (1994). Costs, effects and c/e-ratios alongside a clinical trial. *Health Economics* 3, 309-319.

Hout, BA van, Bonsel, GJ, et al. (1993). Heart transplantation in The Netherlands: costs, effects and scenarios. *Journal of Health Economics* 12, 73-93.

Hunink, MGM, Bult, JR, et al. (1998). Uncertainty in decision models analyzing cost-effectiveness: the joint distribution of costs and effectiveness evaluated with a nonparametric bootstrap method. *Medical Decision Making* 18, 337-346.

Jones, DA (1996). A Bayesian approach to the economic evaluation of health care technologies, in B Spilker, ed., *Quality of life and pharmacoeconomics in clinical trials, Second Edition.* Philadelphia, Lippincott-Raven Publishers: 1189-1196.

Jönsson, B and Haglund, U (1992). Cost-effectiveness of misoprostol in Sweden. *International Journal of Technology Assessment in Health Care* 8, 234-244.

Jorde, R, Burhol, PG, et al. (1988). Peptic ulcer bleeding in patients with and without dyspepsia. *Scandinavian Journal of Gastroenterology* 23, 213-216.

Kadane, JB (1995). Prime time for Bayes. *Controlled Clinical Trials* 16, 313-318.

Keeney, RL and Raiffa, H (1993). *Decisions with multiple objectives.* Cambridge University Press, Cambridge.

Knill-Jones, R, Drummond, M, et al. (1990). Economic evaluation of gastric ulcer prophylaxis in patients with arthritis receiving non-steroidal anti-inflammatory drugs. *Postgraduate Medical Journal* 66, 639-646.

Koning, HJ de, Ineveld, BM van, et al. (1991). Breast cancerscreening and cost-effectiveness; policy alternatives, quality of life considerations and the possible impact of uncertain factors. *International Journal of Cancer* 49, 531-537.

Koopmanschap, MA, Roijen, L van, et al. (1991). Kosten van ziekten in Nederland. Rotterdam, Erasmus University.

Larkai, EN, Smith, JL, et al. (1987). Gastroduodenal mucosa and dyspeptic symptoms in arthritic patients during chronic nonsteroidal anti-inflammatory drug use. *American Journal of Gastroenterology* 82, 1153-1158.

Laska, EM, Meisner, M, et al. (1997). Statistical inference for cost-effectiveness ratios. *Health Economics* 6, 229-242.

Laska, EM, Meisner, M, et al. (1999). Power and sample size in cost-effectiveness analysis. *Medical Decision Making* 19, 339-343.

Laupacis, A, Feeny, D, et al. (1992). How attractive does a new technology have to be to warrant adoption and utilisation? Tentative guide-lines for using clinical and economic evaluations. *Canadian Medical Association Journal* 146, 473-481.

Lilford, RJ and Braunholtz, D (1996). The statistical basis of public policy: a paradigm shift is overdue. *British Medical Journal* 313, 603-607.

Little, R and Rubin, D (1987). *Statistical analysis with missing data.* J. Wiley & Sons, New York.

Löthgren, M and Zethraeus, N (2000). Definition, interpretation and calculation of cost-effectiveness acceptability curves. *Health Economics* 9, 623-630.

Lwanga, SK and Lemeshow, S (1991). *Sample size determination in health studies: a practical manual.* World Health Organization, Geneva.

Makuch, R and Johnson, M (1989). Issues in planning and interpreting active control equivalence studies. *Journal of Clinical Epidemiology* 42, 503-511.

Manning, WG, Fryback, DG, et al. (1996). Reflecting uncertainty in cost-effectiveness analysis, in M Gold, J Siegel, L Russell and M Weinstein, eds., *Cost-effectiveness in health and medicine.* Oxford, Oxford University Press: 263, 264.

Martens, LL, Rutten, FFH, et al. (1989). Cost-effectiveness of cholesterol lowering therapy in The Netherlands: simvastatin versus cholestyramine.
American Journal of Medicine 87, 54S-58S.

McCarthy, DM (1989). Nonsteroidal antiinflammatory drug-induced ulcers: management by traditional therapies. *Gastroenterology* 96, 662-674.

McGuigan, JE (1994). Peptic ulcer and gastritis, in KJ Isselbacher, E Braunwald, JD Wilson, et al. eds., *Harrison's principles of internal medicine.* New York, McGraw-Hill.

Mellem, H, Stave, R, et al. (1985). Symptoms in patients with peptic ulcer and hematemesis and/or melena related to the use of non-steroid anti-inflammatory drugs. *Scandinavian Journal of Gastroenterology* 20, 1246-1248.

Melo Gomes, JA, Roth, SH, et al. (1993). Double-blind comparison of efficacy and gastroduodenal safety of diclofenac/misoprostol, piroxicam, and naproxen in the treatment of osteoarthritis. *Annals of Rheumatic Disease* 52, 881-885.

Merril, AS (1928). Frequency distribution of an index when both the components follow the normal law. *Biometrika* 20, 53-63.

Meltzer, D (2001). Addressing uncertainty in medical cost-effectiveness analysis. Implications of expected utility maximazation for methods to perform sensitivity analysis and the use of cost-effectiveness analysis to set priorities for medical research.
Journal of Health Economics 20, 109-129.

Merton, RC (1998). Applications of option-pricing theory: twenty-five years later.
The American Economic Review 88, 323-349.

Michel, BC, Al, MJ, et al. (1996). Economic aspects of treatment with captopril for patients with asymptomatic left vetricular dysfunction in The Netherlands.
European Heart Journal 17, 731-740.

Michel, BC, Bonsel, GJ, et al. (1992). Levertransplantatie: de effectiviteit op lange termijn. Rotterdam, Dept. of Public Health, Erasmus University.

Moayyedi, P, Soo, S, et al. (2000). Systematic review and economic evaluation of Helicobacter pylori eradication treatment for non-ulcer dyspepsia.
British Medical Journal 321, 659-664.

Mulder, CJJ, Dekker, W, et al. (1996). Lansoprazole 30 mg versus omeprazole 40 mg in the treatment of reflux oesophagitis Grade II, III and IVa (a Dutch multi-centre study). *European Journal of Gastroenterology & Hepatology* 8, 1101-1106.

Netherlands Central Bureau of Statistics (1993). Maandstatistiek van de Bevolking. December 1993. [Monthly bulletin of population. December 1993]. Den Haag, SDU/publishers, Netherlands Central Bureau of Statistics.

Niessen, LW, Dippel, DWJ, et al. (2000). Berekening van kosten na beroerte en kosteneffectiviteit van 'stroke units' en van secundaire preventie, zoals aanbevolen in de herziene CBO-richtlijn 'Beroerte' [Calculation of costs and cost effectiveness of stroke units and of secundary prevention in patients after a stroke, as recommended in the revised CBO guideline 'Stroke']. *Nederlands Tijdschrift voor Geneeskunde* 144, 1959-1964.

O'Brien, B and Gafni, A (1996). When do the "dollars" make sense? Towards a conceptual framework for contingent valuation studies in health care. *Medical Decision Making* 16, 288-299.

O'Brien, BJ and Sculpher, MJ (2000). Building uncertainty into cost-effectiveness rankings: Portfolio risk-return tradeoffs and implications for decision rules. *Medical Care* 38, 460-468.

O'Brien, BJ, Drummond, MF, et al. (1994). In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Medical Care* 32, 150-163.

OECD (1987). Financing and delivering health care: a comparative analysis of OECD countries. Paris.

O'Hagan, A and Stevens, JW (2001a). A framework for cost-effectiveness analysis from clinical trial data. *Health Economics* 10, 303-315.

O'Hagan, A and Stevens, JW (2001b). Assessing and comparing costs: how robust are the bootstrap and methods based on asymptomatic normality? *Health Economics* (submitted).

Palmer, S and Smith, PC (2000). Incorporating option values into the economic evaluation of health care technologies. *Journal of Health Economics* 19, 755-766.

Parrilo, JE, Parker, MM, et al. (1990). NIH Conference. Septic shock in humans: advances in the understanding of pathogenesis, cardiovascular dysfunction and therapy. *Annals of Internal Medicine* 113, 227-742.

Pocock, SJ (1983). *Clinical trials: a practical approach.* John Wiley & Sons, Chichester.

Polsky, D, Glick, HA, et al. (1997). Confidence intervals for cost-effectiveness ratios: a comparison of four methods. *Health Economics* 6, 243-252.

Pouvourville, Gd (1995). Preventing NSAID-induced gastrointestinal toxicity. Economic considerations, methodological problems and results. *PharmacoEconomics* 7, 99-110.

Ramsey, SD, Patrick, DL, et al. (1995a). The cost-effectiveness of lung transplantation; A pilot study. *Chest* 108, 1594-1601.

Ramsey, SD, Patrick, DL, et al. (1995b). Improvement in quality of life after lung transplantation: a preliminary study. *Journal of Heart and Lung Transplantation* 14, 870-877.

Rutten, FFH and Bonsel, GJ (1992). High cost technology in health care: a benefit or a burden? *Social Science and Medicine* 4, 567-577.

Sacristan, JA, Day, SJ, et al. (1995). Use of confidence intervals and sample size calculations in health economic studies. *The Annals of Pharmacotherapy* 29, 719-725.

Schafer, JL (1997). *Analysis of incomplete multivariate data.* Chapman and Hall, London.

Schubert, F (1993). History of the development of pharmacoeconomic guidelines, in F Schubert, ed., *Proceedings of the Canadian Collaborative Work-shop on Pharmacoeconomics,* Exerpta Medica: 1-2.

Schulman, KA, Glick, HA, et al. (1991). Cost-effectiveness of HA-1A monoclonal antibody for gram-negative sepsis; economic assessment of a new therapeutic agent. *Journal of the American Medical Association* 266, 3466-71.

Serruys, PW, Emanuelsson, H, et al. (1996). Heparin-coated Palmaz-Schatz stents in human coronary arteries. Early outcome of the Benestent-II pilot study. *Circulation* 93, 412-422.

Serruys, PW, van Hout, BA, et al. (1998). Randomised comparison of implantation of heparin-coated stents with balloon angioplasty in selected patients with coronary artery disease (BENESTENT II). *The Lancet* 352, 673-681.

Serruys, PW, de Jaegere, P, et al. (1994). A comparison of balloon-expandable stent implantation with balloon angioplasty in patients with coronary artery disease. *New England Journal of Medicine* 331, 489-495.

SIG Zorginformatie (1993). SIG Jaarboek Ziekenhuizen 1991. Utrecht.

Silverstein, FE, Graham, DY, et al. (1995). Misoprostol reduces serious gastrointestinal complications in patients with rheumatoid arthritis receiving nonsteroidal anti-inflammatory drugs. *Annals of Internal Medicine* 123, 241-249.

Smit, JTJ (1996). Growth options and strategy analysis. Rotterdam, Erasmus University.

Spiegelhalter, DJ, Freedman, LS, et al. (1993). Applying Bayesian ideas in drug development and clinical trials. *Statistics in Medicine* 12, 1501-1511.

St. Louis International Lung Transplant Registry (January 1997). report.

Stinnett, AA (1996). Adjusting for bias in C/E ratio estimates. *Health Economics* 5, 470-472.

Stinnett, AA and Paltiel, AD (1996). Mathematical programming for the efficient allocation of health care resources. *Journal of Health Economics* 15, 641-653.

Stinnett, AA and Paltiel, AD (1997). Estimating CE ratios under second-order uncertainty: The mean ratio versus the ratio of means. *Medical Decision Making* 17, 483-489.

Stucki, G, Johannesson, M, et al. (1994). Is misoprostol cost-effective in the prevention of nonsteroidal anti-inflammatory drug-induced gastropathy in patients with chronic arthritis? *Archives of Internal Medicine* 154, 2020-2025.

TenVergert, EM, Essink-Bot, ML, et al. (1998). The effect of lung transplantation on health related quality of life: a longitudinal study. *Chest* 113, 358-364.

TenVergert, EM, Geertsma, A, et al. (in press). The cost-effectiveness of the Dutch lung transplantation programme. *European Respiratory Journal.*

The Economist (2000). In praise of Bayes: September 30th, 111-112.

Thompson, SG and Barber, JA (2000). How should cost data in pragmatic randomised trials be analysed? *British Medical Journal* 320, 1197-1200.

Torrance, GW, Thomas, WH, et al. (1972). A utility maximization model for evaluation of health care programs. *Health Services Research* 7, 118-133.

Umeda, N, Miki, K, et al. (1995). Lansoprazole versus famotidine in symptomatic reflux oesophagitis: a randomized multicenter study. *Journal of Clinical Gastroenterology* 20 Suppl 1, 17-23.

Valkenburg, HA (1980). Reumatoide arthritis, in DE Grobbee and A Hofman, eds., *Epidemiologie van ziekten in Nederland.* Utrecht, Wetenschappelijke uitgeverij Bunge: 200-205.

Vandenbrouke, JP, Hazevoet, HM, et al. (1984). Survival and cause of death in rheumatoid arthritis: a 25-year prospective followup. *Journal of Rheumatology* 11, 158-161.

Varian, HR (1984). *Microeconomic analysis.* W.W. Norton & Company, New York.

Verdickt, W, Moran, C, et al. (1992). A double-blind comparison of the gastroduodenal safety and efficacy of diclofenac and a fixed dose combination of diclofenac and misoprostol in the treatment of rheumatoid arthritis. *Scandinavian Journal of Rheumatology* 21, 85-91.

Wakker, P and Klaassen, MP (1995). Confidence intervals for cost/effectiveness ratios. *Health Economics* 4, 373-381.

Weinstein, M and Zeckhauser, R (1973). Critical ratios and efficient allocation. *Journal of Public Economics* 2, 147-157.

Willan, AR and O'Brien, BJ (1996). Confidence intervals for cost-effectiveness ratios: an application of Fieller's theorem. *Health Economics* 5, 297-305.

Willan, AR and O'Brien, BJ (1999). Sample size and power issues in estimating incremental cost-effectiveness ratios from clinical trials data. *Health Economics* 8, 203-211.

Ziegler, EJ, Fisher, CJ, et al. (1991). Treatment of gram-negative bacteraemia and septic shock with HA-1A human monoclonal antibody against endotoxin. *New England Journal of Medicine* 324, 429-436.

Samenvatting

GEZONDHEIDSZORG-TECHNOLOGIEËN EVALUEREN: ZEKERE METHODEN VOOR ONZEKERE SITUATIES

Inleiding

Economische evaluaties van nieuwe medische technologieën (zoals therapeutische interventies, screeningsprogramma's, transplantaties etc.) hebben tot doel te bepalen wat de verhouding is tussen enerzijds de opbrengsten in termen van gezondheid en anderzijds de (extra) kosten van de nieuwe technologie. Deze economische evaluaties kunnen beleidsmakers ondersteunen bij het nemen van beslissingen over het al dan niet introduceren van een nieuwe technologie in de gezondheidszorg. Helaas geven de uitkomsten van dergelijke evaluaties niet altijd een duidelijk beeld van de balans tussen kosten en baten. Dit wordt onder andere veroorzaakt door de onzekerheid rond de uitkomst. Deze onzekerheid kan verschillende vormen aannemen; zo kan men onzeker zijn over de waarde van variabelen (denk hierbij aan de kans op ziekte, de waardering van kwaliteit van leven, de kosten van ziekenhuisopname etc.) die in de berekeningen gebruikt worden, maar er kan ook onzekerheid bestaan over hoe deze variabelen gecombineerd dienen te worden.

Over de waarde van een variabele kan bijvoorbeeld onzekerheid bestaan als deze geschat is op basis van een steekproef, of als deze door experts op basis van eigen ervaring is ingeschat. Ook kan er onzekerheid ontstaan als de waarde van een variabele geschat wordt op basis van extrapolatie van bekende gegevens. Dit gebeurt bijvoorbeeld als de resultaten van een tweejarige klinische studie worden vertaald naar de langere termijn of als gegevens over percentage cholesterolverlaging vertaald worden naar morbiditeit of sterfte.

Het is niet realistisch te verwachten dat alle onzekerheid geëlimineerd kan worden. Wel kan getracht worden methoden te ontwikkelen om zo goed mogelijk met deze onzekerheid om te gaan.

Een eerste stap in het omgaan met onzekerheid is het beschrijven van de mate van onzekerheid, bijvoorbeeld met behulp van een betrouwbaarheidsinterval. Vervolgens kan getracht worden de onzekerheid te verkleinen. Dit kan bijvoorbeeld voor de aanvang van een studie, door aan te geven welke onzekerheid maximaal acceptabel is en op basis hiervan de benodigde steekproefomvang te berekenen. Ook kan na afloop van een studie de onzekerheid rond de resultaten verkleind worden door de nieuwe gegevens te combineren met gegevens uit andere studies, bijvoorbeeld via een meta-analyse. Tot slot moet getracht worden beleidsmakers te informeren over hoe zij de onzekerheid kunnen meewegen in hun beslissingen. Al de hierboven genoemde methoden van omgaan met onzekerheid komen in dit proefschrift aan de orde.

Beschrijven van onzekerheid

In hoofdstuk 2, 5 en 6 worden verschillende methoden gepresenteerd om onzekerheid te beschrijven.

In hoofdstuk 2 wordt een methode gepresenteerd om de onzekerheid rond een incrementele kosten-effectiviteitsratio (IKER) te beschrijven in situaties waar gegevens over kosten en effecten verzameld zijn op patiëntniveau in het kader van een klinische studie. Eerst wordt de betrouwbaarheidsellips geïntroduceerd, welke een uitbreiding is van een betrouwbaarheidsinterval naar twee dimensies (kosten en effecten). Hierbij wordt verondersteld (verwijzend naar de Centrale Limiet stelling) dat de gemiddelde kosten en effecten per patiënt een bivariate normale verdeling volgen. Net als bij een betrouwbaarheidsinterval geldt dat naarmate de ellips kleiner is, men zekerder is van de schatting van de IKER. Wanneer er positieve additionele effecten en negatieve additionele kosten (ofwel besparingen) zijn, zal een IKER altijd als acceptabel beoordeeld worden. Als de effecten negatief, en de kosten positief zijn zal daarentegen de IKER als onacceptabel gezien worden. Wanneer positieve effecten gepaard gaan met extra kosten, of als negatieve effecten gepaard gaan met besparingen, zal het oordeel afhangen van de limiet die men stelt aan de IKER. Voor elke limiet die men veronderstelt kan nu de kans dat de IKER beneden deze limiet ligt geschat worden door simulaties uit te voeren. Uit de bivariate normale verdeling rond gemiddelde kosten en effecten wordt een groot aantal paren van kosten en effecten getrokken. Voor elk paar wordt gekeken of hun ratio, gegeven de veronderstelde limiet, acceptabel is. Deze procedure kan voor een groot aantal limieten herhaald worden, en de uitkomsten kunnen vervolgens in een 'acceptability curve' weergegeven worden. In deze curve staat op de x-as de limiet die men kiest, en op de y-as het percentage ratio's dat acceptabel is. Op deze wijze kan bijvoorbeeld een eenzijdig 95%-betrouwbaarheidsinterval afgeleid worden door te kijken voor welke limiet de kans dat de IKER acceptabel is 95% is. Overigens moet opgemerkt worden dat in de klassieke statistiek deze uitspraak niet gedaan kan worden: de werkelijke IKER is een 'door de natuur gegeven constante' en is dus wel of niet acceptabel. Alleen binnen de Bayesiaanse statistiek (welke in hoofdstuk 4 besproken wordt) is het toegestaan een kans op 'acceptabel zijn' toe te schrijven aan de IKER.

In hoofdstuk 5 wordt een voorbeeld van een modelleringsstudie gepresenteerd, waarin een zogenaamde gevoeligheidsanalyse wordt gebruikt om de onzekerheid rond de puntschatting van de IKER in kaart te brengen. Het betreft een studie die de kosten en effecten van behandeling met het combinatiepreparaat diclofenac/misoprostol vergelijkt met de kosten en effecten van diclofenac alleen in de behandeling van patiënten met reumatoïde artritis. Diclofenac is een niet-steroïdaal anti-inflammatoir medicijn dat pijnstillend werkt, en misoprostol is een medicijn dat de kans op gastro-intestinale bijwerkingen ten gevolge van het gebruik van diclofenac kan verminderen.

In een model werden de gegevens uit klinische studies, gegevens uit databases en expert-meningen bijeen gebracht. De centrale raming van de IKER kwam uit op NLG 4.179 per gewonnen levensjaar. De gevoeligheid van deze uitkomst voor veranderingen in de invoervariabelen werd zowel in een univariate als in een probabilistische gevoeligheidsanalyse bestudeerd. In een univariate analyse worden alle variabelen één voor één gevarieerd tussen bepaalde grenzen. In het algemeen is het moeilijk vast te stellen welke grenzen hiervoor gebruikt moeten worden en nog lastiger is het te beoordelen wanneer de uitkomst van de studie gevoelig voor veranderingen in een bepaalde variabele genoemd moet worden. In de probabilistische analyse wordt voor elke variabele een kansverdeling gedefinieerd. Vervolgens wordt een groot aantal maal uit elke verdeling een waarde getrokken waarmee het model opnieuw wordt doorgerekend. Op deze manier krijgt men niet alleen een idee wat de meest waarschijnlijke uitkomst is, maar ook wat de meest extreme uitkomsten (best denkbare en slechts denkbare) zijn, met een inschatting hoe waarschijnlijk deze extreme uitkomsten zijn. Met name dit laatste is een belangrijk voordeel van de probabilistische gevoeligheidsanalyse.

In hoofdstuk 6 wordt een scenarioanalyse beschreven die werd uitgevoerd in het kader van de technology assessment van het Nederlandse longtransplantatie programma 1990-1995. Deze scenarioanalyse beschrijft een toekomstig programma dat gedurende 15 jaar zal worden uitgevoerd. Door een dergelijke periode te kiezen wordt bereikt dat er een stabiele in- en uitstroom van patiënten op de wachtlijst ontstaat, wat nog niet het geval was in de studieperiode. De gegevens uit de studieperiode werden gebruikt om kansen te schatten voor de groep met transplantatie, en op basis hiervan werd, met behulp van een microsimulatie-model, een controlegroep geconstrueerd. Vervolgens werden kosten, overleving en kwaliteit van leven in beide groepen met elkaar vergeleken. Uit de analyse volgde een schatting van de kosten per voor kwaliteit van leven (kvl) gecorrigeerd levensjaar van NLG 167.000 (na disconteren met 5% per jaar).

Vervolgens werd het model gebruikt om additionele scenario's door te rekenen, namelijk een donorscenario waarin een toename van het aantal beschikbare donorlongen werd verondersteld (NLG 135.000 per voor kvl gecorrigeerd levensjaar) en een beleidsscenario waarin de instroom van patiënten in het programma beperkt zou worden (NLG 144.000 per voor kvl gecorrigeerd levensjaar). Deze scenarioanalyses beschrijven de (extra) onzekerheid die kan ontstaan door externe invloeden op een variabele. Het aantal beschikbare donorlongen werd in de basisanalyse geschat op 17 per jaar, wat gebaseerd is op de geobserveerde aantallen per jaar. Dit betekent dat met behulp van statistische methoden een betrouwbaarheidsinterval rond dit aantal berekend kan worden. Echter, het aantal beschikbare donorlongen kan ook beïnvloed worden door beleidmakers (via wetgeving) en door artsen (bijvoorbeeld via de selectiecriteria voor donorlongen en de behandeling van de donor). Een deelstudie liet zien dat het mogelijk is met extra inspanning tot maximaal 27 beschikbare donorlongen per jaar te komen. Iets dergelijks geldt ook voor de instroom van patiënten. Door striktere inclusiecriteria te gebruiken kan deze instroom sterk beperkt worden, waarbij de instroom veel lager wordt dan de ondergrens van een betrouwbaarheidsinterval rond de waargenomen instroom.

Reduceren van onzekerheid

In de hoofdstukken 3 en 4 worden (statistische) methoden gepresenteerd die kunnen bijdragen aan het verkleinen van onzekerheid.

Hoofdstuk 3 beschrijft een simulatiemethode voor het bepalen van de optimale steekproefomvang in kosten-effectiviteitsstudies. Om een dergelijke berekening te maken is het nodig te specificeren welke mate van onzekerheid men maximaal acceptabel vindt, welke limiet er aan de IKER gesteld wordt, wat de verwachte kosten en effecten zijn, wat de verwachte spreiding rond de gemiddelde kosten en effecten is en wat de verwachte correlatie tussen kosten en effecten is. Ook moet vastgesteld worden welk onderscheidend vermogen vereist wordt, dat wil zeggen welke kans men wil hebben om te concluderen dat een behandeling kosteneffectief is als deze het ook werkelijk is. In de gepresenteerde studie is nagegaan hoe gevoelig de geschatte steekproefomvang is voor veranderingen in deze parameters, hetgeen geïllustreerd wordt aan de hand van twee klinische studies. Het blijkt onder andere dat de benodigde steekproefomvang groter wordt naarmate de limiet die aan de IKER gesteld wordt lager ligt, en naarmate kosten en effecten meer negatief gecorreleerd zijn.

In hoofdstuk 4 wordt beschreven hoe onzekerheid rond de resultaten van een studie verminderd kan worden door deze resultaten te combineren met gegevens uit bijvoorbeeld andere studies met behulp van zogenaamde Bayesiaanse methoden.

De methoden die worden gepresenteerd worden geïllustreerd aan de hand van gegevens uit twee opeenvolgende klinische studies. In beide studies werden de kosten en effecten vergeleken van stent implantatie versus 'dotteren' bij patiënten met angina pectoris.

Voor de analyse van de tweede studie werden drie a-priori verdelingen van kosten en effecten gespecificeerd, welke in toenemende mate gebaseerd waren op de resultaten uit de voorafgaande klinische studie. Door de a-priori verdelingen te combineren met de resultaten van de tweede studie werden drie a-posteriori resultaten gevonden. De puntschatting van de IKER varieerde van NLG 16.269 tot NLG 21.984 per additionele 'event-free' overlevende. Wanneer men veronderstelt dat de limiet op de IKER NLG 30.000 bedraagt, dan zijn de kansen dat de IKERs acceptabel zijn zeer verschillend (65%, 80%, 90%), wat betekent dat verschillende veronderstellingen omtrent de a-priori verdeling tot verschillende beslissingen over de introductie van stent implantatie zouden kunnen leiden.

Beslissen in onzekerheid

In hoofdstuk 7 wordt ingegaan op de vraag hoe beleidsmakers rekening kunnen houden met hun attitude ten aanzien van onzekerheid bij het nemen van beslissingen. Dit wordt gedaan vanuit het perspectief van budgetallocatie, dat wil zeggen dat de beleidsmaker een budget optimaal moet verdelen over mogelijke programma's. In de literatuur voor economische evaluaties in de gezondheidszorg wordt over het algemeen aangenomen dat de doelstelling is de gezondheidseffecten te maximaliseren voor een vast, gegeven budget. Op basis van deze doelstelling is het optimaal de gezondheidszorgprogramma's te rangschikken op hun ratio van totale kosten en totale effecten en vervolgens, beginnend met het programma met de kleinste ratio, net zo lang programma's te accepteren tot het gehele budget op is. Echter, in deze benadering wordt er geen rekening mee gehouden dat de eerste 1000 levensjaren die 'gekocht' worden hoger gewaardeerd worden dan de extra 1000 levensjaren op het moment dat men er al 20.000 heeft. En evenzo, men zal liever 1000 gulden uitgeven als het hele budget nog intact is dan wanneer dit de laatste 1000 gulden van het budget zijn. In hoofdstuk 7 wordt een alternatief model gepresenteerd, waarbij de beleidsmaker tot doel heeft een waarderingsfunctie over kosten en effecten te maximaliseren. Deze functie wordt gekarakteriseerd door de hierboven beschreven afnemende meeropbrengst. In deze benadering wordt nog geen rekening gehouden met onzekerheid. Deze notie van onzekerheid kan in het optimalisatieprobleem worden aangebracht door een logische extensie van de waarderingsfunctie te gebruiken, namelijk de verwachte nutsfunctie. In deze functie worden de preferenties van de beleidsmaker over onzekere beslissingen weergegeven. In het hoofdstuk wordt aan de hand van getallenvoorbeelden geïllustreerd wat het effect van de verschillende beschreven benaderingen op de budgetallocatie is.

Tot slot

Zoals eerder is aangegeven, is het niet realistisch te veronderstellen dat alle onzekerheid geëlimineerd kan worden. Dit mag natuurlijk geen beletsel zijn om optimaal gebruik te maken van de methoden die beschikbaar zijn voor het beschrijven en reduceren van onzekerheid. Het is echter duidelijk dat dit nog niet in alle economische evaluaties van medische technologieën gebeurt. Dit is niet verrassend, als men bedenkt dat de meeste methodologische ontwikkelingen pas de laatste paar jaar hebben plaatsgevonden. Hopelijk zullen de ontwikkelde methoden de komende jaren steeds vaker toegepast worden, en zal de kennis van preferenties over onzekere beslissingen toenemen.

Acknowledgments

I would like to begin by thanking my promotors Ben van Hout and Frans Rutten for all their work. Ben van Hout offered me a position at the iMTA 9 years ago, where he introduced me to the modelling of diseases and to sensitivity analysis. Later, he initiated our research on statistical methods in economic evaluations. I truly enjoyed the inspiring discussions that we would have, often on Friday evenings when his phone wouldn't ring. I am grateful to Frans Rutten for his crucial comments on this thesis. However, what I appreciated most was his quiet, implicit assumption that I would finish this thesis and his continued support in reaching this goal.

I also want to thank my current supervisor Els Grijseels, for overseeing the process of writing and finishing this thesis. She kept the train moving until it could not be stopped anymore.

I enjoyed the collaboration with the colleagues with whom I wrote some of the chapters of this thesis, Bowine Michel, Marc Koopmanschap, Xander Koolman, and Talitha Feenstra. It is amazing how much you learn by working with someone from another discipline.

I owe many thanks to Pfizer Inc. USA for offering me the Pfizer Health Economics Fellowship. This fellowship has allowed me to concentrate on methodological work, which has not only resulted in several chapters for this thesis, but will also be of great use in the coming years. Many people at Pfizer in Groton, CT and New York have made my stays in the USA unforgettable, and from time to time my mind still wanders off to those lovely summers there. One person I would like to mention is Joe Cappelleri. He supervised my work while I was in Groton, and I am grateful for everything I learned from him.

Finally, I would like to thank my family, friends, and colleagues for their support. I feel privileged to have so many people around me who will listen to me, advise me and encourage me in all my undertakings.