PEPIJN VEMER

Dealing with Differences

Different populations, data sources and countries in HTA modelling

Dealing with Differences

Different populations, data sources and countries in HTA modelling

Pepijn Vemer

FUNDING:
The studies in this thesis were financed by the Dutch Ministry of Health (ch 2), Takeda Pharmaceuticals (ch 3), Netherlands Organisation for Health Services Research (ZonMW, ch 4&5), Pfizer (ch 6) and iMTA (ch 7&8).
Vemer P.

Design and printing: Optima Grafische Communicatie, Rotterdam. *www.ogc.nl*
Foto cover: Guus van Os

ISBN: 978-94-6169-533-8

Dealing with Differences

Different populations, data sources
and countries in HTA modelling

Omgaan met Verschillen

Verschillende populaties, data bronnen
en landen in HTA modelleren

Proefschrift
ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus
Prof.dr. H.A.P. Pols
en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op
vrijdag 3 oktober 2014 om 11.30 uur
door

Pepijn Vemer
Geboren te Enter

Promotor:
Prof.dr. M.P.M.H. Rutten-van Mölken

Other members:
Prof.dr. C.A. Uyl-de Groot
Prof.dr. F. Rutten
Prof.dr. M.J. Postma

PROPOSITIONS

Propositions accompanying this thesis.

1. Patient heterogeneity and parameter uncertainty should both be considered in cost-effectiveness Markov models, but separately (this thesis).
2. When only direct evidence is available, evidence from different sources is best synthesized using a frequentist random effects model (this thesis).
3. When indirect evidence is also available, combining direct and indirect evidence is preferred over only using direct evidence (this thesis).
4. Transferability factors can be ordered by their impact on the cost-effectiveness (this thesis).
5. The national threshold value of a QALY has wrongly been disregarded as an important transferability factor (this thesis).
6. Several advanced statistical methods exist to calculate country-specific cost-effectiveness results based on multinational trials, but they have not been used on a wider scale yet, while simpler, naïve methods are still routinely employed (this thesis).
7. Mathematical formulas, even making no sense within the presented context, may add credibility when added to your abstract. (Eriksson K. The nonsense math effect. Judgment and Decision Making, 2012;7(6):746-9)
8. Although a good presentation can never save bad research, a bad presentation can certainly ruin good research.
9. Scientific discovery doesn't begin with "Eureka!", but rather with "Huh. That's funny. That can't be right." (Paraphrased from Isaac Asimov)
10. Inequality can never be overcome by more inequality.
11. The societal impact of donating half a liter of blood every ten weeks during the past six years, is higher than that of this thesis.

TABLE OF CONTENTS

Chapter 1

Introduction

A simple concept; many complications

1.1 A SIMPLE CONCEPT

At its heart, health technology assessment (HTA) is very simple. It compares two or more alternative courses of action, often pharmaceutical interventions, in terms of both their costs and health outcomes.[1] One of the interventions will have better health outcomes, for example fewer number of exacerbations, longer survival or a better quality of life. This usually comes at an extra cost, often in the way of a higher price for the intervention. HTA makes this exchange between costs and effects explicit.

The idea that costs are an important element to take into account, does not come naturally to many health care workers. Doctors, nurses, and other health care workers do everything they can to help patients improve their lives. The interventions these patients need are provided in a large part by companies developing and producing the necessary drugs and devices. Health care scientists and epidemiologists try to make sense of what constitutes health, what illness is and how disease is spread. Their focus is purely on the patient: what does he or she need? Choices between treatment options are usually a consideration between availability, possible side effects, and patient characteristics. If a new medication comes on the market, doctors are often eager to treat patients with this newest treatment option.

With the focus on the patient in front of them, health care workers usually do not look beyond the operating room or treatment room. An oncologist wants to treat all patients to the best of his or her ability, no matter the costs of the intervention. Budgetary constraints are not, and should not, be part of the decision making process of a health care worker when dealing with an individual patient. Cost considerations should be taken into account at a more aggregate level in the clinical guidelines, written by their organizations. In this way, HTA separates health care workers from these concerns in their daily practice, which are in the public and political domain.

1.2 SCARCITY

Where health care workers are trained to focus on the patient in front of them, it is the task of policy makers and politicians to look beyond individual concerns. Health care workers may want access to the newest treatment option, but the money that will be used to pay for the latest treatment may very well have a better use somewhere else. As the then chairman of the National Institute for Health and Clinical Excellence (NICE) Professor Sir Michael Rawlins said in an interview in 2008: 'We have a finite amount of money for healthcare, and if you spend money one way you can't spend it in another.'[2] This could very well be within the health care sector, where the money could be put to use for another patient. It

could also be used for a public campaign to quit smoking or to build a new hospital wing, to maintain forests or to build a new museum.

The fact that we need to make a choice between all these options, is due to the basic notion of scarcity in economics. Money, like any other good, is finite and can only be spend once. In health care, the money is usually paid by the general public in the form of taxes (the UK's National Health Service for example) or insurance premiums (The Netherlands). As such, "society" (that is: all of us) pays for health care and we may expect doctors to be careful in how they spend it. This scarcity asks for an explicit valuation of all treatment options and their costs. In other words: HTA.

1.3 HEALTH TECHNOLOGY ASSESSMENT

HTA as a field is intended to provide a bridge between the world of research and the world of decision-making.[3] HTA is the toolbox which helps to keep a societal perspective when making decisions about new treatments, without intervening with the day-to-day decision that health care workers have to make. Using the tool box of HTA, the outcomes are often presented as an incremental cost-effectiveness ratio (ICER) or the incremental net monetary benefit (INMB). The ICER is the ratio between the additional health outcomes of one treatment option over another (the comparator), divided by the difference in costs. By comparing this ICER to an (implicit or explicit) threshold, policy makers can deem the intervention to be cost-effective compared to the comparator when the ICER is below this threshold. If the ICER is above the threshold, the intervention cannot be considered cost-effective compared to the comparator. The height of this threshold should be a public choice. The INMB is the difference in health outcomes, valued in monetary terms, minus the monetary costs. For the valuation of health outcomes the threshold value mentioned above is used. If the INMB is positive, the new intervention has more value to society than costs, and can thus be considered cost-effective compared to the comparator. If the INMB is negative, the new intervention will cost more than the societal benefits, and the intervention cannot be considered cost-effective compared to the comparator.

Health outcome measures that are often used are the number of years that a patient lives, or the number of events (heart attacks, hospitalizations) a patient experiences. Often, one would want to also include quality in life, leading to a measure called the quality adjusted life year (QALY). With several interventions, the difference between these outcomes can then be used to say which one is "better". Costs are commonly categorized in costs that are directly related to the intervention or disease, and indirect societal costs or savings.[4] Direct medical costs include for example the costs of medication. Direct non-medical costs include for example travel, informal care and patient time. Indirect non-medical costs may include productivity losses and consumption changes. Finally, indirect medical

costs are medical costs that may be incurred by living longer.[5] Including these costs in life years gained make cost estimates more conservative and accurate, but are not used in most published CE-studies.

HTA is an active field internationally and has seen continued growth fostered by the need to support management, clinical, reimbursement and other policy decisions. It has also been advanced by the evolution of evaluative methods in the social and applied sciences, including clinical epidemiology and health economics.[6] Health policy decisions are becoming increasingly important as the opportunity costs from making wrong decisions continue to grow, both in the number of in terms of wasted resources and opportunities for health gain forgone.[7]

In recent years, the field of HTA has seen rapid advances in the field of statistical techniques, which allow more realistic and complex healthcare models to be simulated more rapidly.[8] Some of these advances have helped in the calculation of parameter estimates and measurement of uncertainty. Techniques have also arisen to identify, quantify and handle differences between groups of patients, data sources and countries. This has led to an increased interest in these differences. This thesis focuses on some of these techniques.

1.4 THIS THESIS

At its heart, health technology assessment is very simple. However, HTA is facing many methodological challenges calling for more complexity in the analyses. Several of these challenges are addressed in this thesis, all revolving around the use of HE decision models. We start with an example of such a HE decision model in chapter 2, which calculates the long term health economic effects of the reimbursement of smoking cessation treatments.

The first methodological issue we approach is what happens when there are different populations for which we need to make a decision. In chapter 3, we show that heterogeneity, caused by differences in patient characteristics, can and should be analyzed separately from the measures of uncertainty used in HE models. Unfortunately, heterogeneity is often either ignored completely in practice, or analyzed together with parameter uncertainty, without taking into account the fundamental difference between the two. We show that this may lead to the wrong policy decision.

Next, we turn to the issue of different data sources. Often, parameters in a HE model come from a variety of sources, for example several trials. Each individual estimate is usually different from the estimate from the other sources, which may be due to either sampling error, or genuine differences between the trials. In order for this information to be useful, these estimates need to be combined into a single estimate for each parameter. The way to do this is called meta-analysis. In chapters 4 and 5, we compare several different forms of meta-analysis, using a simulation study. The methods are compared

with respect to their effect on the HE outcomes. When only information is used from a head-to-head comparison between the interventions under investigation, the methods used are called direct meta-analysis. Four methods of direct meta-analysis are compared in chapter 4. When the relative efficacy between two interventions is obtained through a common comparator, there is an indirect comparison. Evidence from both direct and indirect comparisons can also be included in a network meta-analysis. Four methods of mixed treatment comparison are compared in chapter 5.

In the next three chapters we turn to the methodological issues that arise from differences between countries. The field of health economics that deals with these issues is called transferability. In chapter 6 we show how the outcome for a single model can be different, depending on the country for which the model's parameters have been used. We investigate how much of the differences can be explained by different sets of parameters. In chapter 7, we show that the country-specific willingness-to-pay for a QALY is often ignored in the transferability discussion, although it does have a large influence on the transferability of the outcomes. Transferability issues are best dealt with as early as possible. It is possible, and in fact advantageous, to deal with differences between countries alongside a randomized controlled trial (RCT). In chapter 8, we compared how recent, large RCTs dealt with these differences in practice and showed that several advanced statistical techniques have been available for some time to calculate country-specific CE results. However, they have not been used on a wide scale yet, while simpler, naïve methods are still routinely employed.

We finish with a discussion in chapter 9, of some of the outcomes from the research presented in this thesis, and some further issues within HTA.

1.5 LITERATURE

[1] Drummond M, Sulpher M, Torrance G, O'Brien B, Stoddart G. Methods for the Economic Evaluation of Health care Programmes. 3rd ed. Oxford: Oxford University Press; 2005.

[2] Hinsliff G. Health chief attacks drug giants over huge profits. 2008; Available at: http://www.theguardian.com/uk/2008/aug/17/pharmaceuticals.nhs. Accessed 12/26, 2013.

[3] Battista RN. Towards a paradigm for health technology assessment. In: Peckham M, Smith R, editors. The Scientific Basis of Health Services London: BMJ Publishing Group; 1996.

[4] Rappange DR, van Baal PH, van Exel NJ, Feenstra TL, Rutten FF, Brouwer WB. Unrelated medical costs in life-years gained: should they be included in economic evaluations of health-care interventions? Pharmacoeconomics 2008;26(10):815-830.

[5] van Baal PH, Feenstra TL, Hoogenveen RT, de Wit GA, Brouwer WB. Unrelated medical care in life years gained and the cost utility of primary prevention: in search of a 'perfect' cost-utility ratio. Health Econ 2007 Apr;16(4):421-33.

[6] Menon D, Marshall D. The Internationalization of Health Technology Assessment. International Journal of Technology Assessment in Health Care 1996;12(1):45-51.

[7] Sculpher MJ, Claxton K, Drummond M, McCabe C. Whither trial-based economic evaluation for health care decision making? Health Econ 2006 Jul;15(7):677-687.

[8] Weinstein MC. Recent developments in decision-analytic modelling for economic evaluation. Pharmacoeconomics 2006;24(11):1043-53.

Chapter 2

If you try to stop smoking, should we pay for it?

The cost–utility of reimbursing smoking cessation support in The Netherlands

P. Vemer, M.P.M.H. Rutten-van Mölken, J. Kaper, R.T. Hoogenveen, C.P. van Schayck, T.L. Feenstra

Previously published in Addiction, 2010, 105, 1088–1097
doi:10.1111/j.1360-0443.2010.02901.x

ABSTRACT

Background Smoking cessation can be encouraged by reimbursing the costs of smoking cessation support (SCS). The short-term efficiency of reimbursement has been evaluated previously. However, a thorough estimate of the long-term cost–utility is lacking.

Objectives To evaluate long-term effects of reimbursement of SCS.

Methods Results from a randomized controlled trial were extrapolated to long-term outcomes in terms of health care costs and (quality adjusted) life years (QALY) gained, using the Chronic Disease Model. Our first scenario was no reimbursement. In a second scenario, the short-term cessation rates from the trial were extrapolated directly. Sensitivity analyses were based on the trial's confidence intervals. In the third scenario the additional use of SCS as found in the trial was combined with cessation rates from international meta-analyses.

Results Intervention costs per QALY gained compared to the reference scenario were approximately €1,200 extrapolating the trial effects directly, and €4,200 when combining the trial's use of SCS with the cessation rates from the literature. Taking all health care effects into account, even costs in life years gained, resulted in an estimated incremental cost–utility of €4,500 and €7,400, respectively. In both scenarios costs per QALY remained below €16,000 in sensitivity analyses using a life-time horizon.

Conclusions Extrapolating the higher use of SCS due to reimbursement led to more successful quitters and a gain in life years and QALYs. Accounting for overheads, administration costs and the costs of SCS, these health gains could be obtained at relatively low cost, even when including costs in life years gained. Hence, reimbursement of SCS seems to be cost-effective from a health care perspective.

2.1 INTRODUCTION

The World Health Organization Framework Convention on Tobacco Control is the first negotiated global health treaty[1]; most countries in the world have implemented some sort of smoking cessation policy. On a European level, the European Union (EU) has developed a policy to decrease tobacco use based on legislative policies, support for smoking prevention and cessation activities, mainstreaming tobacco control into other policies and ensuring that achievements also have an impact outside the EU region.[2] These international efforts notwithstanding, smoking policies still remain largely a national matter.

The United Kingdom has an elaborate programme of free cessation support in specialized cessation clinics within the setting of the National Health Service (NHS).[3,4] Other countries that reimburse some form of cessation support include New Zealand, Australia and the United States.[5] However, the reimbursement is often incomplete. For instance, in Australia, bupropion is reimbursed, but nicotine replacement therapy (NRT) is not.[5] In the United States, a small majority of states includes reimbursement of smoking cessation therapy in the Medicaid package, but Medicare does not usually cover smoking cessation support (SCS). Private insurances vary in their coverage.[6] France and Germany do not reimburse SCS[7], while Italy provides partial reimbursement.

Tobacco control policy in The Netherlands aims to reduce smoking prevalence to 20% in 2010.[8] Smoking prevalence is declining in The Netherlands, but the decline has slowed recently and in 2007 28% of the Dutch population still smoked. Additional efforts are required to reach the goal. A new policy might be the broad reimbursement of SCS via the obligatory health care insurance. A Cochrane review from 2005 on the effects of reimbursement of SCS concluded that there is some evidence that complete reimbursement leads to higher quit rates than partial or no reimbursement, but also that more research was necessary.[9] The Cochrane review by Reda and colleagues summarizes studies on financial incentives, among others full reimbursement of the costs of SCS to smokers.[10] They concluded that full financial interventions directed at smokers could increase the proportion quitting, quit attempts and utilization of pharmacotherapy by smokers.

In The Netherlands, SCS is reimbursed only partly at present and pharmacological SCS is not reimbursed at all. The health insurance board (CVZ) has advised the Dutch Ministry of Health to reimburse an integrated smoking cessation programme, consisting of a combination of behavioural counselling and pharmacotherapy.[11] Following this advice, the Ministry of Health intends to start reimbursement of integrated smoking cessation support in 2011.[12]

A randomized controlled trial was performed to investigate the effects of such a reimbursement policy in the Dutch region of Friesland[13,14] in May 2002. The trial included smokers above the age of 18 who were representative of the Dutch population with respect

to age (40 in the trial and 43 in the general population) and gender (55% male, both in the trial and in the general Dutch population).[15,16] Smokers were assigned randomly to either an intervention or a control group. For a period of 6 months, smokers in the intervention group were offered reimbursement for NRT, bupropion and behavioural counselling. They received a leaflet with a description of the type of SCS for which reimbursement was available, and information on how to receive the reimbursement. No reimbursement was offered to the control group. The trial was designed as a naturalistic implementation study.

The trial found that the number of participants using SCS was higher in the intervention group than in the control group. The difference was 6.7 percentage points (%pt), with a 95% confidence interval (CI) of 3.8–9.5. The total number of smokers attempting to quit was also higher, but this difference of 2.6%pt (95% CI -2.0 to 5.0) was not significant. The self-reported abstinence rate after 12 months was again significantly higher in the intervention group than in the control group (difference = 2.7%pt, 95% CI 0.5–4.9). These results are in line with international studies, as reviewed recently.[10]

Recently, a pilot study was carried out in The Netherlands to investigate the feasibility of large-scale implementation of reimbursement.[17] The results of that study showed that reimbursement improves the use of cessation support. For example, 24.8% of respondents used bupropion compared to 4.1% in the general population: a six-fold increase. Use of varenicline was nine times higher and nicotine patches 2.7 times. At the end of the 6-month test period, one-third of the participants who were interviewed indicated that they had stopped smoking. The 6-month self-reported abstinence rate as found in the Friesland trial was substantially lower than that measured in the pilot study; this may be explained by the lack of randomization and selection of participants in the pilot. Participants indicated that they wanted to quit at the start of the pilot study and the abstinence rate was calculated as a percentage of these respondents. Therefore, the current study was based upon the cessation rates in the Friesland trial.

The following research question was addressed: what is the long-term cost–utility of reimbursing smoking cessation strategies? We aimed to answer this question by extrapolating the above-mentioned trial data using a dynamic population model. The model calculated the effect of increased quit rates on the number of smokers, quality adjusted life years (QALYs) and costs, and the long-term cost–utility, accounting for relapse rates and delays in health effects among former smokers. Moreover, the cost–utility ratios included costs in life years gained (LYG) to provide a complete estimate of all health care related consequences of implementing reimbursement.

2.2 METHODS

2.2.1 Three policy scenarios

We simulated the long-term outcomes for three scenarios using the Chronic Disease Model (CDM), with the reimbursement policy in place for half a year, as was the case in the original trial. Three policy scenarios were defined. The first scenario, denoted as the reference scenario, considers current practice without reimbursement of cessation support. The initial distribution over all smoking classes and the 1-year smoking class transition probabilities were estimated from survey data from the Dutch Foundation on Smoking and Health (STIVORO).[18] This is a yearly representative national survey on smoking, including questions on current and past smoking status. Initiation rates of never smokers, quit rates of current smokers and rates of relapse for former smokers were estimated using the 2001, 2002 and 2003 surveys.[19]

Scenarios 2 and 3 are different versions of what would happen if SCS were being reimbursed by the obligatory basic health care insurance. As was the case in the original trial, smokers pay for the SCS themselves and will receive the money back from their insurers. The scenarios were defined by changes in costs and in quit rates; that is, the number of people moving from a state of 'current smoker' to 'former smoker' in one 12-month cycle of our model. Scenario 2 will be referred to as the trial-based reimbursement scenario, and it remained close to the original empirical data on quit rates and costs.[13,14] Because the CIs around the prices and effects of SCS from the original trial were relatively wide, we also included a third scenario. This scenario combined the use of SCS as observed in the trial with estimates of the effectiveness and costs of SCS from published literature.[20-22] It has higher intervention costs and lower effectiveness. This scenario will be referred to as the literature-based reimbursement scenario.

Year 1 of our model run, starting with 6 months of the reimbursement policy, corresponds to 2006. The outcomes were estimated after 20 years and for a life-time horizon. All input prices were in 2005 Euros, using the internationally accepted harmonized index of consumer prices—all items.[23] All outcomes have been discounted back to 2005. Costs were discounted at 4% and effects were discounted at 1.5%, according to Dutch guidelines for pharmacoeconomic evaluations.[24] We adopted a health care perspective. Sensitivity analyses were carried out to analyse the effect of different discount rates, the length of the period that the reimbursement policy will be in place and the uncertainty around the costs and effects of reimbursement.

2.2.2 Intervention effects

Two main effects are to be expected from reimbursement of SCS: more quit attempts and a higher use of cessation aids. As a result we expect more successful quitters. Yearly quit rates in the 'reference scenario' were defined by gender and 5-year age classes and

Table 2.1: Costs per smoker and quit rates used in the trial-based reimbursement scenario. [13,14].[a]

	Control group	Intervention group	Difference (95% CI)
Costs of SCS	€3.87	€14.06	€10.19
Overhead costs			
Application for reimbursement		€1.20	€1.20
Identifying smokers and informing them about the reimbursement policy		€7.32	€7.32
Total costs per smoker in the trial-based reimbursement scenario	€3.87	€22.58	€18.71 (8.82–33.90)
Quit rates			
12-month assessment, prolonged abstinence, self-reported	2.8%	5.5%	2.7%pt (0.5–4.9%pt)

[a] CI: confidence interval; SCS: smoking cessation support; %pt: percentage point.

ranged from 4.4% for men aged 15–19 years to 11% for women older than 85 years.[25] For the 'trial-based reimbursement scenario', the higher quit rate was based directly on the original trial data (table 2.1). Six months after the end of the reimbursement period, at the 12-month assessment, 2.8% of the control group and 5.5% of the intervention group reported being abstinent for at least 6 months. This difference was statistically significant. In the trial-based reimbursement scenario, we calculated the additional number of successful quitters by adding the absolute difference in quit rates, +2.7%pt, to the quit rates of the reference scenario. The 95% CI around this difference was used in the sensitivity analyses.

To calculate a quit rate, in the 'literature-based reimbursement scenario' the increased use of SCS that was observed in the trial was multiplied with efficacy estimates for these interventions based on meta-analyses from published Cochrane reviews (table 2.2). For instance, of 634 smokers in the control group, five (0.8%) used NRT only, while of the 632 smokers in the intervention group 15 (2.4%) used NRT only. The additional use of NRT is therefore 2.4 - 0.8% = 1.6%. The abstinence rate after 12 months from using NRT is 13.5%[20], which means that the extra use of NRT will cause, on average, 13.5 x 1.6% = 0.2% additional successful quitters. The total expected increase in successful quitters over all interventions is 1.1%, which is added to the quit rates of the reference scenario. Alternative treatments such as homeopathy or acupuncture were not included in the reimbursement scheme, and were used less frequently in the trial intervention group than in the control group. Because there is no evidence-based effect of these treatments, we used the efficacy of 'intensive counselling' for 'intensive counselling plus alternative SCS' and the effect of 'placebo' for 'alternative SCS'. Using the 95% CIs of the published abstinence rates and of the estimated additional numbers of users, we calculated a 95% CI for the effect in the literature-based reimbursement scenario, which was used in the sensitivity analyses.

Table 2.2: Use of smoking cessation support (SCS) in the Friesland trial and the predicted increase in successful quitters and costs in the literature-based reimbursement scenario.[a]

	Additional use (% smokers) [13,14]	Abstinence rates [13,14,20–22,39]	Price per user [20,39]	Cost per smoker (95% CI)	% Increase in the number of successful quitters (95% CI)
NRT	1.6%	13.5%	€186	€2.98	0.2%
NRT + IC	0.9%	22.0%	€394	€3.55	0.2%
NRT + bupropion + IC	0.2%	18.4%	€573	€1.15	0.0%
Bupropion	1.1%	13.3%	€180	€1.98	0.2%
Bupropion + IC	2.1%	18.4%	€377	€7.92	0.4%
IC	1.0%	13.9%	€198	€1.98	0.1%
IC + alternative SCS	-0.2%	13.9%	€253	-€0.50	-0.0%
Alternative SCS	-0.2%	9.3%	€55	-€0.11	-0.0%
Overhead costs				€18.94	
Application for reimbursement				€1.20	
Identifying smokers and informing them about the reimbursement policy				€7.32	
Additional total costs per smoker in the literature-based reimbursement scenario				€27.46 (9.11–61.45)	
Increase in quit rate per smoker in the literature-based reimbursement scenario					1.1%pt (0.6%–1.5%pt)

[a] CI: confidence interval; SCS: smoking cessation support; NRT: nicotine replacement therapy; IC: intensive counselling; %pt: percentage point.

2.2.3 Intervention costs

For the trial-based reimbursement scenario, costs per smoker and the corresponding CI were calculated bottom-up based on the resource use observed during the trial (see table 2.1). We included the costs of counseling and use of pharmacotherapy. Overhead costs consisted of the costs of identifying the smokers, sending all eligible smokers a letter and leaflet describing the SCS for which reimbursement were available and the way in which they could apply for reimbursement, plus administration costs of reimbursement. The costs of SCS that were used in the literature-based reimbursement scenario are given in table 2.2. These costs were estimated bottom-up based on recommended resource use and unit costs.[20,25] Costs per quit attempt were multiplied by the additional use of the SCS to calculate the additional cost per smoker. Returning to the earlier NRT example, 1.6% additional users multiplied with the costs of a quit attempt using NRT (€186) results in a cost of €2.98 per smoker. The same overhead costs as in the trial-based reimbursement

scenario were included. A CI was based upon the uncertainty surrounding additional use of SCS and minimum and maximum estimates for resource use.[20,25]

2.2.4 The CDM

To estimate the long-term effects of smoking cessation, the RIVM (National Institute for Public Health and the Environment) CDM was used. This model has been described extensively by Hoogenveen and colleagues.[26] The CDM is a Markov-type state-transition model that describes the effects of epidemiological risk factors on morbidity and mortality from 28 chronic diseases in the Dutch population, and several risk factors, including smoking. It is a population-based, dynamic model that accounts for changes over time in the demographics of the population and the prevalence of risk factors. It includes realistic time-lags between the moment of smoking cessation and effects on the incidence of smoking-related disease and takes into account the fact that successful quitters can relapse into smoking. The CDM has been used in relation to smoking for future projections of risk factor and disease prevalence numbers[15,27,28], cost–effectiveness analyses[29] and estimates of healthy life expectancy.[30] The CDM relates smoking to increased incidence rates of 14 smoking related chronic diseases, including coronary heart disease, chronic obstructive pulmonary diseases (COPD) and several types of cancer. The incidence rates of smoking-related diseases are increased in current smokers as well as in former smokers, with the relative risks of former smokers declining as a function of time since cessation.[19] More details on the model and model inputs are presented in[16,19,30] which are freely available on the internet and in Appendix A2.

2.3 RESULTS

2.3.1 Effects on smoking prevalence

In year 1 of our projections (2006), the percentage of smokers in The Netherlands was approximately 28.2%, which amounts to 3.8 million smokers.[31] This percentage diminishes slowly over time. Due to the reimbursement policy, the number of smokers decreases faster. At the end of the first year, the number of smokers in the trial-based reimbursement scenario is 0.7%pt lower than in the reference scenario and after 20 years (2025) is 0.18%pt. In the literature-based reimbursement scenario the percentage of smokers was 0.3%pt lower than in the reference scenario after 1 year and 0.08%pt lower after 20 years. Differences in the reference scenario diminish over time, as reimbursement was assumed to finish after half a year.

2.3.2 QALYs gained

Figure 2.1 shows the difference in QALYs per year in the different scenarios compared to the reference scenario. The trial-based reimbursement scenario raised the number of QALYs with a maximum of 24,00 per year in 2036. At this time-point, the largest proportion of the smokers who quit during the intervention were reaching an age where they would have developed one of the smoking-related diseases included in the CDM. The smokers who lived at the time of the reimbursement policy die consecutively and the health benefits of the intervention were mainly gone by the year 2095. In the literature-based reimbursement scenario, health benefits were lower than in the trial-based reimbursement scenario. The number of QALYs gained reached a peak in 2036 at 1,000 QALYs per year.

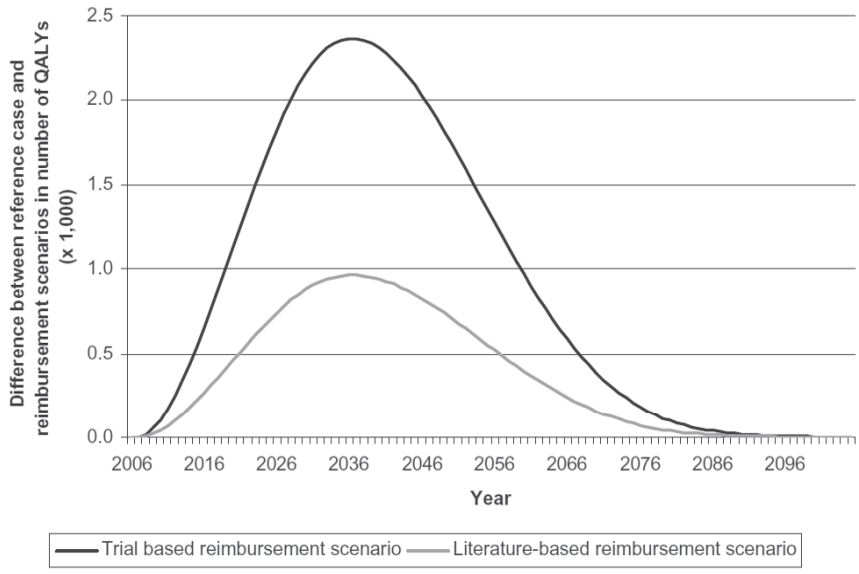

Figure 2.1: Difference in QALYs per year in the trial-based and literature-based reimbursement scenarios compared to the reference scenario.[a]

[a] QALY: Quality adjusted life year

2.3.3 Effects on health care costs

The effects on health care costs are shown in figure 2.2 for the trial-based scenario. The annual costs of smoking-related diseases were lower in the intervention scenarios than in the reference scenario during the first decades. The difference between the trial-based reimbursement scenario and the reference scenario reached a peak in 2029, with annual smoking-related costs being €12.0 million lower in the trial-based reimbursement scenar-

io; savings decreased thereafter. From 2059 onwards the annual smoking-related health care costs were higher in the trial-based reimbursement scenario than in the reference scenario. This can be explained by a greater number of people in the cohort still alive in the trial-based reimbursement scenario. The same pattern was seen for the literature-based reimbursement scenario, with maximum cost savings of €4.9 million in 2029.

In the extra life years gained by successful quitters, additional costs are generated for diseases unrelated to smoking, such as dementia or hip fractures in old age.[29] This increases the costs of all scenarios (figure 2.2). In the trial-based reimbursement scenario, annual total health care costs, including these unrelated health care costs, were lower in the first years than in the reference scenario, up to a maximum of -€3.4 million in 2018. From 2026 onwards, the total health care costs in the trial-based reimbursement scenario were higher than in the reference scenario, with a maximum cost-difference of €32.6 million in 2051. The literature-based reimbursement scenario showed a similar pattern, with cost savings up to -€1.4 million in 2018 and higher costs from 2026 onwards, with a maximum of €13.3 million in 2051.

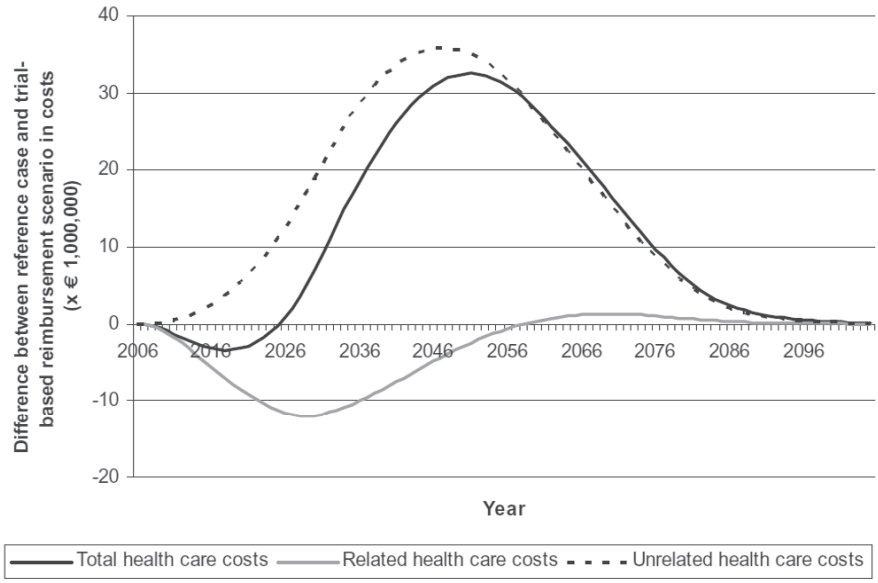

Figure 2.2: Difference in smoking-related, unrelated and total health care costs per year in the trial-based scenario compared to the reference scenario

2.3.4 Cost–utility

Two different incremental cost–utility ratios (ICUR) were computed (table 2.3). The first ratio is the intervention costs per QALY gained; the second ratio relates total costs from a health care perspective to QALYs gained.

Table 2.3: Difference in costs, LYs gained, QALYs gained and ICUR after 20 years and 100 years, as compared to the reference scenario.[a,b]

	Trial-based reimbursement scenario		Literature-based reimbursement scenario	
Time horizon	20	100	20	100
Costs (x €1 million)				
Smoking-related health care costs	-62.2	-130.7	-25.3	-53.3
Unrelated health care costs	37.9	305.4	15.4	124.4
Intervention costs	68.1	68.1	93.5	93.5
Total costs	43.9	242.9	83.6	164.6
Life years (LY) gained (x 1,000)	9.2	67.7	3.7	27.6
ICUR: intervention costs per LYG	7,450	1,010	25,080	3,390
ICUR: total costs per LYG	4,790	3,590	22,420	5,970
QALYs gained (x 1,000)	11.2	54.6	4.6	22.2
ICUR: intervention costs per QALY gained	6,100	1,250	20,530	4,200
ICUR: total costs per QALY gained	3,930	4,450	18,360	7,400

[a] LY: life year; QALY: Quality adjusted LY; ICUR: Incremental cost-utility ratio

[b] Costs discounted at 4%, LYs and QALYs discounted at 1.5%.

2.3.5 Sensitivity analysis

In our main analysis we discounted costs at 4% and outcomes at 1.5%.[24] When using a discount rate of 0% for both costs and outcomes, the life-time ICUR was €13,300 per QALY for the trial-based reimbursement scenario and €15,100 per QALY for the literature-based reimbursement scenario, relating total costs to QALYs gained. The respective ICURs after 20 years of the two reimbursement scenarios were €2,200 and €13,900 per QALY. Discounting both costs and outcomes at 4% decreases the net present value of the number of QALYs compared to the base analyses. Because the effects of reimbursement are found mainly far in the future, the LYG and QALYs gained are reduced considerably when using this higher discount rate, leading to less favourable ICURs. The life-time ICUR was €9,100 per QALY for the trial-based reimbursement scenario and €15,100 per QALY for the literature based reimbursement scenario. The respective ICURs after 20 years were €5,500 and €25,700 per QALY.

From a policy viewpoint, it is more realistic to assume that the policy will be in place for a longer period than the trial's implementation period of 6 months. Therefore, we have

also calculated the outcomes for a scenario where the intervention is in place for 4 years, assuming the quit rates to remain constant over this period. Four years were considered an appropriate period because it is the time span between elections for the Dutch parliament. With the intervention in place for a 4-year period, the difference in total costs between the reference scenario and our trial-based scenario was €823 million. The total number of QALYs gained was 204,000. Dividing both, the resulting ICUR of €4,040 per QALY was slightly better than in the main analysis. In our literature-based scenario the difference in total costs amounted to €540 million, with 84,000 QALYs gained, resulting in an ICUR of €6,400 per QALY, which was also slightly better than in the main analysis. At the end of the reimbursement period, the number of smokers in the trial-based reimbursement scenario was 1.7%pt lower than in the reference scenario. In the literature-based reimbursement scenario the number of smokers was 0.7%pt lower than in the reference scenario, using a 4-year reimbursement period.

Based on the confidence ranges presented in tables 2.1 and 2.2, the effect of uncertainty on intervention costs was analysed. Using a life-time horizon, the ICUR in the trial based scenario varied between €3,800 and €5,500 per QALY and the ICUR in the literature-based reimbursement trial varied between €4,600 and €12,600 per QALY. Uncertainty about the effectiveness of the intervention was also analysed based on the 95% CIs (see

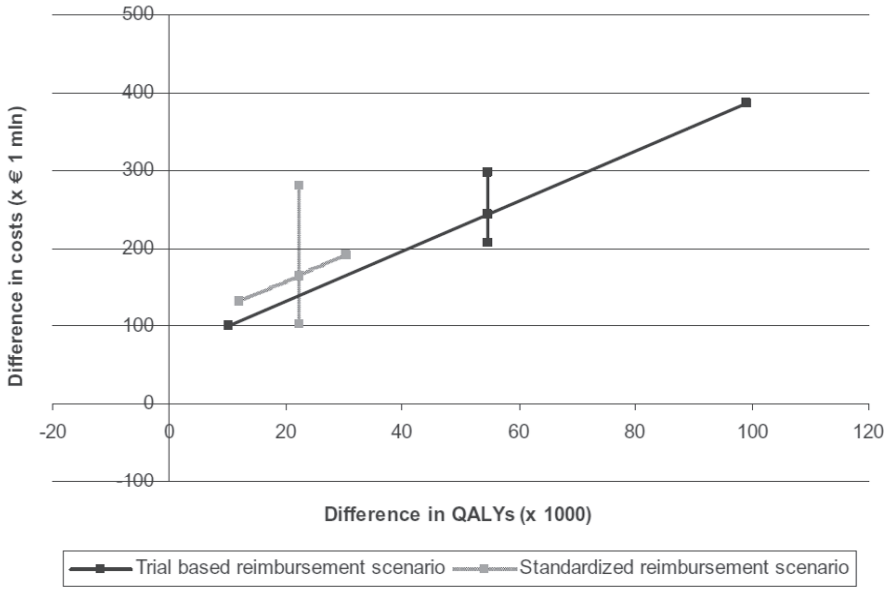

Figure 2.3: CE-plane that shows the sensitivity of the trial-based and literature-based reimbursement scenarios for changes in costs and effectivity.[a]

[a] CE: Cost-effectiveness; QALY: Quality adjusted life year.

tables 2.1 and 2.2). The ICUR of the trial-based reimbursement scenario varied between €3,900 (maximum effectiveness) and €9,900 (minimum effectiveness) per QALY using a life-time horizon. The ICUR of the literature-based reimbursement scenario varied between €6,300 and €10,900 per QALY using a life-time horizon.

Figure 2.3 shows the cost–utility plane (CU-plane) with the results of the sensitivity analyses of intervention costs and effectivity. The lines show by how much the minimum and maximum estimates of total effects on health care costs and QALYs differed. Uncertainty in effectiveness affected the estimates of both costs and QALYs, and therefore these lines were not horizontal.

2.4 DISCUSSION

This study showed that full reimbursement of the costs of smoking cessation support to smokers is a cost-effective way to contribute to a reduction in the percentage of smokers. Reimbursement for a period of half a year led to a decrease in the percentage of smokers between 0.3%pt and 0.7%pt, compared to no reimbursement. Having the policy in place for 4 years would lower the percentage of smokers at the end of the reimbursement period between 0.7%pt and 1.7%pt. This is a contribution towards the goal of the Dutch government of reducing the percentage of smokers to 20%. The findings support the advice by the health insurance board CVZ to the Dutch Ministry of Health to reimburse an integrated smoking-cessation programme.[11] However, reimbursement alone will probably not suffice, even if in reality this will be in place for a longer period of time, given that the smoking rate in the first year of our model (2006) was still 28%. In 2008 the Dutch smoking rate was 27%.[31] The Ministry of Health intends to begin reimbursement by 2011.

The number of life years that can be gained with a reimbursement scheme that is available for half a year was 68,000 in our trial-based scenario and 28,000 in our literature-based reimbursement scenario. When life years are adjusted for health-related quality of life, total gains were 55,000 and 22,000 QALYs, respectively. Because of extra health care costs in these life years gained, total costs of reimbursement exceed the total cost of no reimbursement after an initial period of savings in health care costs.[32] Relating costs to health gains, the costs per QALY gained for a life-time horizon were found to be about €4,500 in our trial-based estimate and €7,400 in our literature-based estimate. These low ratios reflect the large health gains thatmay be obtained from smoking cessation. Our results were robust to uncertainty in intervention costs and in effectiveness showing that, like most smoking cessation policies, reimbursement was a cost-effective intervention.[10,33]

Other evaluations of reimbursement or free supply of cessation support exist[5,34-36], but none of them was based directly on a randomized controlled trial allowing free choice of support methods and showing the actual increase in the use of cessation support. Also,

none of these evaluations included overhead costs related to reimbursement. Our study showed that these additional costs are well balanced by the health gains obtained from more successful quit attempts.

The analyses included costs in life years gained. Compared to estimates that account only for savings in costs of care for smoking-related diseases (e.g.[22,37]), our current estimates are more conservative and complete. Excluding the costs in the life years gained, as is performed frequently in other studies, and focusing upon intervention costs minus savings in costs of smoking-related diseases, the intervention was cost saving in the trial-based reimbursement scenario and had an ICUR of less than €2,000 in the literature-based reimbursement scenario (data not shown).

Sensitivity analyses showed that the time horizon and discount factors used had a considerable influence on the outcomes. The health effects of cessation need some time to become apparent, implying that for a time horizon of 20 years, the incremental cost–utility ratios (ICURs) were less favourable than for the main estimates, which used a life-time horizon so that all health effects were included.

The reimbursement period in our study was only 6 months, following the actual trial. Of course, when a reimbursement policy is implemented, it will usually remain available for a longer time-period. Because the trial had only 6 months of reimbursement, we do not have empirical data on what will happen when the policy remains in place. International trials all had a follow-up time of less than 12 months.[10] The smokers most eager to quit may have tried to quit first, resulting in a decline in the number of quit attempts over time. Smokers and medical professionals may also become more familiar with the policy, which might lead to an increase in quit attempts over time. With no reliable data available, in our sensitivity analysis on the length of reimbursement period we have assumed no change in quit rates compared to the first 6 months.

A longer implementation period had only a small effect on cost-effectiveness ratios, even though the total number of QALYs gained and the difference in total costs were much larger. This is because both costs and health effects increased proportionally. The small changes observed were due to a difference in discounting between costs and outcomes, which is the current Dutch standard for cost-effectiveness analyses.

The clinical trial by Kaper et al. focused upon short term costs and effects. It found that if society was willing to pay €1,000 (€10,000) for an additional 12-month quitter, the probability that reimbursement for SCS would be cost-effective was 50% (95%). The study also included a quick extrapolation to long-term effects, based on American data. They estimated an effect of 10.6 QALYs gained in the control group, amounting to 62.000 QALYs for all smokers in The Netherlands. This is close to, but higher than, our estimate of the trial-based reimbursement scenario. The intervention costs per QALY gained as estimated by Kaper et al. were €1,802, using discount rates of 4% for both costs and outcomes, which was lower than our estimates of the cost–utility of the two scenarios.

They concluded that if society was willing to pay €18,000 per QALY, an accepted Dutch cut-off point for preventive interventions[38], the probability that reimbursement for SCS was cost-effective was 95%. Our current estimates, based on a much more sophisticated model, confirmed the cost–utility of reimbursement.

A recently performed pilot study investigated the feasibility of large-scale implementation of reimbursement in The Netherlands.[17] It estimated the costs of implementing reimbursement for a 6-month period in the whole of The Netherlands. The estimate varied between €14.0 million and €22.7 million, based upon the pilot study costs per participant and the number of Dutch smokers who indicated that they wanted to quit smoking within 6 or 12 months. This is substantially lower than our calculation of the intervention costs. Part of this difference is explained because the pilot study ignored the overhead costs of reimbursement.

In the trial-based scenario we used the self-reported quit rates after 12 months. The biochemically validated abstinence rates after 12 months were also significantly higher in the intervention group compared to the control group. We based our scenario on self-reported quit rates, as the parameter estimates on smoking in the CDM were also based on questionnaires. Risk differences based on the biochemically validated cessation rates were very similar, with a somewhat narrower CI.

The main limitation of this study is that we used the results of a single trial. The number of quit attempts per type of SCS was relatively limited, as was the geographical coverage. In order to assess the degree of certainty around the outcomes we performed sensitivity analyses using the CIs around costs and quit rates, and simulated an alternative scenario based on systematic reviews and meta-analyses for costs and effectiveness of smoking cessation. Because of the small regional differences regarding smoking in The Netherlands, we feel that the outcomes of this study are applicable to the whole of the Netherlands. Even though the outcomes of the original trial were in line with those of other similar studies[10], the results may, however, not be applicable to countries with different health care systems or at different stages of the smoking epidemic. For instance, the important role of general practitioners (GPs) in the Dutch health care system was reflected by the success of smoking cessation support delivered by primary care. This may be less applicable to countries with a less prominent role for GPs. Also, a country with much higher smoking prevalence might gain a larger reaction from the intervention than in our study. Similarly, a country with much lower prevalence might notice a smaller impact on quit rates.

However, the advantage of using a trial is that effects on quit rates could be estimated with a comparable control group over a 12-month follow-up time. Such an experimental setting is usually unavailable for evaluation of policy measures.

2.5 CONCLUSION

Reimbursement of smoking cessation support via the obligatory health care insurance in the Netherlands will result in fewer smokers and more QALYs. Reimbursement seems a cost-effective way to contribute to a reduction in the percentage of smokers.

2.6 LITERATURE

[1] World Health Organization (WHO). WHO Framework Convention on Tobacco Control. Geneva: WHO FCTC; 2005.

[2] European Union Public Health Information System (EUPHIX). European Union Public Health Information System. Bilthoven: Directorate-General for Health and Consumer Protection and National Institute for Public Health and the Environment (DG-SANCO and RIVM); 2008.

[3] National Institute for Clinical Excellence (NICE). Brief Interventions and Referral for Smoking Cessation in Primary Care and Other Settings. London: NICE; 2006.

[4] National Institute for Clinical Excellence (NICE). Smoking Cessation Services in Primary Care, Pharmacies, Local Authorities and Workplaces, Particularly for Manual Working Groups, Pregnant Women and Hard to Reach Communities. London: NICE; 2008.

[5] Bertram M. Y., Lim S. S., Wallace A. L., Vos T. Costs and benefits of smoking cessation aids: making a case for public reimbursement of nicotine replacement therapy in Australia. Tob Control 2007; 16: 255–60.

[6] Professional Assisted Cessation Therapy (PACT). Reimbursement for Smoking Cessation Therapy—A Healthcare Practitioner's Guide. Hackensack, NJ: PACT; 2003.

[7] Nguyen-Kim L., Or Z., Paris V., Sermet C. The politics of drug reimbursement in England, France and Germany. 2005; 99: http://www.irdes.fr/EspaceAnglais/Publications/IrdesPublications/QES099.pdf (accessed 9 April 2010).

[8] Ministerie van Volksgezondheid, Welzijn en Sport(VWS). Kiezen voor Gezond Leven 2007–2010 [Choosing for healthy living]. The Hague: VWS; 2006.

[9] Kaper J., Wagena E. J., Severens J. L., Schayck C. P. Healthcare financing systems for increasing the use of tobacco dependence treatment. Cochrane Database Syst Rev 2005; 1: CD004305.

[10] Reda A. A., Kaper J., Fikrelter H., Severens J. L., van Schayck C. P. Healthcare financing systems for increasing the use of tobacco dependence treatment [update]. Cochrane Database Syst Rev 2009; 2: CD004305.

[11] Kroes M. E., Mastenbroek C. G. Stoppen-metrokenprogramma: te verzekeren zorg! [Smoking Cessation Programs: Care to be Insured!]. Report number 290065312009. Diemen, the Netherlands: College van Zorgverzekeringen; 2009.

[12] Ministerie van Financiën. Dutch National Budget [Rijksbegroting], Art 41. The Hague: Ministerie van Financiën; 2009.

[13] Kaper J., Wagena E. J., Willemsen M. C., van Schayck C. P. Reimbursement for smoking cessation treatment may double the abstinence rate: results of a randomized trial. Addiction 2005; 100: 1012–20.

[14] Kaper J. Smoking Cessation Treatment and Its Reimbursement, the Costs and Effects. Maastricht: Datawyse/Universitaire Pers Maastricht; 2006.

[15] Mulder I., van Genugten M. L. L., Hoogenveen R. T., de Hollander A. E. M., Bueno-de-Mesquita H. B. The impact of smoking on future pancreatic cancer: a computer simulation. Ann Oncol 1999; 10: 74–8.

[16] Hoogenveen R.T., deHollander A. E. M., van Genugten M. L. L. The chronic diseases modelling approach. Report number 266750001, May 1998. Bilthoven: RIVM. Available from www.rivm.nl/bibliotheek/rapporten/266750001.html (accessed 21 January 2010).

[17] van den Berg B., Soethout J. E. Proefimplementatie 'Stoppen met roken'. Resultaat begeleidend onderzoek. [Pilot implementation Study 'Smoking Cessation': Results of accompanying research.]. Report number 1663, March 2009. Amsterdam: Regioplan. Available from http://www.regioplan.nl/media/pdf/id/513/file_name/1663-Proefimplementatie+stoppen+met+roken (accessed 21 January 2010).

[18] Dutch Foundation on Smoking and Health (STIVORO). Roken, de harde feiten: Volwassenen 2003 [Smoking, the hard facts: Adults 2003]. The Hague: STIVORO; 2004.

[19] Hoogenveen R. T., van Baal P. H., Boshuizen H. C., Feenstra T. L. Dynamic effects of smoking cessation on disease incidence, mortality and quality of life: the role of time since cessation. Cost Eff Resour Alloc 2008; 6: http://www.resource-allocation.com/content/6/1/1 (accessed 16 March 2010).

[20] Feenstra T., Van Baal P., Hoogenveen R., Vijgen S., Stolk E., Bemelmans W. Cost-Effectiveness of Interventions to Reduce Tobacco Smoking in The Netherlands. An Application of the RIVM Chronic Disease Model. Report 260601003. Bilthoven: National Institute for Public Health and the Environment (RIVM); 2005.

[21] Lancaster T., Stead L. F. Individual behavioural counselling for smoking cessation [Review]. Cochrane Database Syst Rev 2008; 3: CD001292. DOI: 10.1002/14651858. CD001292.pub2.

[22] HoogendoornM.,Welsing P., Rutten-van MölkenM. P.M. H. Cost-effectiveness of varenicline compared with bupropion, NRT and nortriptyline for smoking cessation in the Netherlands. Curr Med Res Opin 2008; 24: 51–61.

[23] Organization for Economic Cooperation and Development (OECD). OECD Statistics. Paris: OECD; 2009.

[24] College voor Zorgverzekeringen. Richtlijnen Voor Farmaco-Economisch Onderzoek; Evaluatie En Actualisatie [Guidelines for pharmacoeconomic research: evaluation and actualization]. Diemen, the Netherlands: College voor Zorgverzekeringen; 2008.

[25] Feenstra T. L., Hamberg-van Reenen H. H., Hoogenveen R. T., Rutten-van MölkenM. P.M. H. Cost-effectiveness of face-to-face smoking cessation interventions: a dynamic modeling study. Value Health 2005; 8: 178–90.

[26] Hoogenveen R. T., van Baal P. H., Boshuizen H. C. Chronic disease projections in heterogeneous ageing populations: approximating multi-state models of joint distributions by modelling marginal distributions. Math Med Biol 2010; 27: 1–19.

[27] Feenstra T. L., van Genugten M. L., Hoogenveen R. T., Wouters E. F., Rutten-van Mölken M. P. M. H. The impact of aging and smoking on the future burden of chronic obstructive pulmonary disease: a model analysis in the Netherlands. Am J Respir Crit Care Med 2001; 164: 590–6.

[28] Struijs J. N., van Genugten M. L., Evers S. M., Ament A. J., Baan C. A., van den Bos G. A. Modeling the future burden of stroke in the Netherlands: impact of aging, smoking, and hypertension. Stroke 2005; 36: 1648–55.

[29] van Baal P. H., Feenstra T. L., Hoogenveen R. T., deWit G. A., Brouwer W. B. Unrelated medical care in life years gained and the cost utility of primary prevention: in search of a 'perfect' cost–utility ratio. Health Econ 2007; 16: 421–33.

[30] van Baal P. H., Hoogenveen R. T., deWit G. A., Boshuizen H. C. Estimating health-adjusted life expectancy conditional on risk factors: results for smoking and obesity. Popul Health Metrics 2006; 4: 14.

[31] Dutch Foundation on Smoking and Health (STIVORO). Website: STIVORO for a Smoke Free Future. The Hague: STIVORO; 2008.

[32] van Baal P. H., Polder J. J., de Wit G. A., Hoogenveen R. T., Feenstra T. L., Boshuizen H. C. et al. Lifetime medical costs of obesity: prevention no cure for increasing health expenditure. PLoS Med 2008; 5: e29.

[33] Ronckers E. T., Groot W., Ament A. J. Systematic review of economic evaluations of smoking cessation: standardizing the cost-effectiveness. Med Decis Making 2005; 25: 437–48.

[34] Fellows J. L., Bush T., McAfee T., Dickerson J. Cost effectiveness of the Oregon quitline 'free patch initiative'. Tob Control 2007; 16: i47–52.

[35] Salize H. J., Merkel S., Reinhard I., Twardella D., Mann K., Brenner H. Cost-effective primary care-based strategies to improve smoking cessation: more value for money. Arch Intern Med 2009; 169: 230–6.

[36] Joyce G. F., Niaura R., Maglione M., Mongoven J., Larson-Rotter C., Coan J. et al. The effectiveness of covering smoking cessation services for medicare beneficiaries. Health Serv Res 2008; 43: 2106–23.

[37] Bolin K., Mörk A., Willers S., Lindgren B. Varenicline as compared to bupropion in smokingcessation therapy—cost–utility results for Sweden. Respir Med 2008; 102: 699–710.

[38] Casparie A., van Hout B. A., Simoons M. L. Guidelines and costs [in Dutch]. Ned Tijdschr Geneeskd 1998; 142: 2075–7.

[39] Kaper J., Wagena E. J., van Schayck C. P., Severens J. L. Encouraging smokers to quit: the cost effectiveness of reimbursing the costs of smoking cessation treatment. Pharmacoeconomics 2006; 24: 453–64.

A2 APPENDIX

A2.1 The Chronic Disease Model

To estimate the long term effects of smoking cessation, the RIVM Chronic Disease Model (CDM) was used. This model has been extensively described by Hoogenveen and colleagues.[1] The CDM is a state-transition Markov-type model that was designed to describe the effects of epidemiological risk factors on morbidity and mortality from several chronic diseases in the Dutch population. It includes 28 chronic diseases and several risk factors amongst which smoking, Body Mass Index, and physical inactivity. In modeling diseases explicitly, the structure of the model is similar to the Prevent model[2] and the Quit Benefits model.[3,4] An important difference with the Prevent model is that different risk factor classes are modeled. In comparison with the Quit Benefits Model the CDM includes more diseases and a finer structure of age categories, allows for co-morbidity and has the ability to track health effects over a longer period. The CDM has been used in relation to smoking for future projections of risk factor and disease prevalence numbers prevalence numbers[5-7], cost effectiveness analyses[8] and estimates of healthy life expectancy.[9]

The model describes the life course of cohorts in terms of transitions between risk factor classes and transitions between disease states over time. Risk factors and diseases are linked through relative risks of disease incidence.[10] The parameters used in the model are the 1-year probabilities of each transition between model states. The main model outcome variables are numbers of incident and prevalent cases and numbers of deaths, specified by disease, age and gender.

Demographic data such as all cause mortality rates and initial population numbers were available from Statistics Netherlands.[11] To estimate incidence, prevalence and mortality rates in the general population, three types of data sources were used: general practitioner registrations for non-cancer diseases, national cancer registries, and cohort studies for diabetes.[12,13] To compute health effects in terms of quality-adjusted life years (QALYs), the CDM used quality of life weights derived from the Dutch Burden of Disease Study[9,14,15], to adjust life years lived with a disease, with 1 reflecting full health and 0 a quality of life equal to death. Health care costs per patient per year are based on the Dutch Costs of Illness study.[8,16]

The CDM relates smoking to increased incidence rates of 13 smoking-related chronic diseases, i.e. coronary heart disease (acute myocardial infarction (AMI) and other coronary heart disease), congestive heart failure, stroke, chronic obstructive pulmonary diseases (COPD), diabetes, and cancer of the lung, stomach, larynx, oral cavity, esophagus, pancreas, bladder and kidney. The incidence rates of smoking-related diseases are increased in current smokers as well as in former smokers, with the relative risks of former smokers declining from the risk of a smoker immediately after stopping smoking to that of a never smoker as a function of time since cessation.[17] Smoking specific all cause mortality rates

were used in the model in combination with the disease specific excess mortality rates. More details on model inputs are presented by Van Baal et al.[9]

A2.2 Literature

[1] Hoogenveen R.T., van Baal P.H., Boshuizen H.C. Chronic disease projections in heterogeneous ageing populations: approximating multi-state models of joint distributions by modelling marginal distributions. Math Med Biol 2010; 27: 1-19 .

[2] Barendregt J. PREVENT: the technical background. Public Health Models: Tools for Health Policy making at National and European level. Nijkerk: Callenbach; 1999.

[3] Hurley S.F., Matthews J.P. The Quit Benefits Model: a Markov model for assessing the health benefits and health care cost savings of quitting smoking. Cost Eff Resour Alloc 2007; 5.

[4] Barendregt J.J., Bonneux L., van der Maas P.J. The health care costs of smoking. N Engl J Med 1997; 337: 1052-7 .

[5] Feenstra T.L., van Genugten M.L., Hoogenveen R.T., Wouters E.F., Rutten-van Mölken M.P.M.H. The impact of aging and smoking on the future burden of chronic obstructive pulmonary disease: a model analysis in the Netherlands. Am J Respir Crit Care Med 2001; 164: 590-6 .

[6] Struijs J.N., van Genugten M.L., Evers S.M., Ament A.J., Baan C.A., van den Bos G.A. Modeling the future burden of stroke in The Netherlands: impact of aging, smoking, and hypertension. Stroke 2005; 36: 1648-55.

[7] Mulder I., van Genugten M.L.L., Hoogenveen R.T., de Hollander A.E.M., Bueno-de-Mesquita H.B. The impact of smoking on future pancreatic cancer: a computer simulation. Ann Oncol 1999; 10: 74-8 .

[8] van Baal P.H., Feenstra T.L., Hoogenveen R.T., de Wit G.A., Brouwer W.B. Unrelated medical care in life years gained and the cost utility of primary prevention: in search of a 'perfect' cost-utility ratio. Health Econ 2007; 16: 421-33 .

[9] van Baal P.H., Hoogenveen R.T., de Wit G.A., Boshuizen H.C. Estimating health-adjusted life expectancy conditional on risk factors: results for smoking and obesity. Popul Health Metr 2006; 4: 14 .

[10] Lambert J.D. Numerical methods for ordinary differential systems: the initial value problem. Chichester: Wiley; 1991.

[11] Statistics Netherlands. ; 2008.

[12] Lancaster T., Stead L.F. Individual behavioural counselling for smoking cessation (Review). Database of Systematic Reviews 2008; Art. No.: CD001292. DOI: 10.1002/14651858. CD001292.pub2.

[13] Hoogendoorn M., Welsing P., Rutten-van Mölken M.P.M.H. Cost-effectiveness of varenicline compared with bupropion, NRT and nortriptyline for smoking cessation in the Netherlands. Curr Med Res and Op 2008; 24: 51-61 .

[14] Melse J.M., Essink-Bot M.L., Kramers P.G., Hoeymans N. A national burden of disease calculation: Dutch disability-adjusted life-years. Dutch Burden of Disease Group. Am J Public Health 2000; 90: 1241-7 .

[15] Stouthard M.E.A., Essink-Bot M.L., Bonsel G.J., et al. Disability Weights for Diseases in The Netherlands. Department of Public Health. Erasmus University Rotterdam; 1997.

[16] Slobbe L.C.J., Kommer G.J., Smit J.M., Groen J., Meerding W.J., Polder J.J. Kosten van ziekten in Nederland 2003. 2006.

[17] Hoogenveen R.T., van Baal P.H., Boshuizen H.C., Feenstra T.L. Dynamic effects of smoking cessation on disease incidence, mortality and quality of life: The role of time since cessation. Cost Eff Resour Alloc 2008; 6.

Chapter 3

Not simply more of the same

Distinguishing between patient heterogeneity and parameter uncertainty

P. Vemer, L.M.A. Goossens, M.P.M.H. Rutten-van Mölken

Accepted, Medical Decision Making

Acknowledgements: The authors wish to thank participants of the LoLa HESG meeting in May 2013 for their useful comments.

ABSTRACT

In cost-effectiveness (CE) Markov models, heterogeneity in the patient population is not automatically taken into account. We aimed to compare methods of dealing with heterogeneity on estimates of CE, using a case study in COPD. We first present a probabilistic sensitivity analysis (PSA) in which we sampled only from distributions representing parameter uncertainty. This ignores any heterogeneity. Next, we explored heterogeneity by presenting results for subgroups. The next method samples parameter uncertainty simultaneously with heterogeneity in a Single Loop PSA. Finally, we distinguish parameter uncertainty from heterogeneity in a Double Loop PSA, by performing a nested simulation within each PSA iteration. Point estimates and uncertainty differed substantially between methods. The incremental CE ratio (ICER) ranged from €4,900 to €13,800. The Single Loop PSA led to a substantially different shape of the CE-plane and an overestimation of the uncertainty compared with the other three methods. The CE-plane for the Double Loop PSA showed substantially less uncertainty and a stronger negative correlation between the difference in costs and the difference in effects than the other methods. This comes at the cost of higher calculation times. Not accounting for heterogeneity, Subgroup Analysis and the Double Loop PSA, can all be viable options, depending on the decision makers' information need. The Single Loop PSA should not be used in CE research. It disregards the fundamental differences between heterogeneity and sampling uncertainty and overestimates uncertainty as a result.

3.1 INTRODUCTION

Heterogeneity in cost-effectiveness (CE) models refers to true differences in outcomes between patients, which can be explained by differences in patient characteristics. Examples can be disease severity, age, presence of different biomarkers and many more.[1] To date, only limited attention is paid to patient heterogeneity and pharmacoeconomic guidelines provide hardly any methodological guidance on the subject.[2] CE models that follow individual patients, such as discrete event simulation or micro-simulation, automatically account for patient heterogeneity, as each individual patient can be modelled with its unique characteristics. However, in the case of Markov models, which follow a cohort of patients over time, heterogeneity is not automatically taken into account.

Patients in Markov models are often assumed to be homogeneous, with parameter estimates obtained from aggregated data. The probabilistic sensitivity analysis (PSA) is performed by drawing from distributions which represent the uncertainty about the population average of parameters, like transition probabilities or utilities. If the model contains any variables representing patient characteristics, such as the starting age of the cohort, the models use a single point estimate. The CE outcomes for this "average patient" are then assumed to represent the entire patient population.

However, with non-linearity being the rule rather than the exception in Markov modelling[3], the incremental costs and effects for the average patient are not equal to the average over all patients.[4,5] This makes it incorrect to assume that these outcomes are applicable for heterogeneous populations. In order to obtain a correct CE estimate over the heterogeneous population as a whole, heterogeneity should be taken into account explicitly.

Heterogeneity has been handled in Markov models in three different ways. The first is to calculate outcomes for several different combinations of patient characteristics in subgroup analyses.[4] The comparison of subgroups allows for the exploration of the effect that differences between patients have on CE outcomes. For example, Bolin et al. calculated CE for smoking-cessation interventions for men and women separately.[6] These subgroups may provide useful insights. However, policy and reimbursement decisions are commonly made for an entire patient population, not subgroups. It is, for example, inconceivable that a policy maker concerned with the reimbursement of smoking cessation treatment will make a different decision for male and female smokers. Such a distinction is usually made only between disease severity classes, but this is only a part of the total existing heterogeneity. In addition, it is often difficult to determine in practice whether a difference between subgroups is genuine or simply reflects noise in the data.[3] Furthermore, even if subgroups are a viable option in decision making, an average patient will have to be used to represent these subgroups. Here too, as above, the CE outcomes may not be representative for all patients in the subgroup.

A second approach is to perform a PSA, which draws from all available distributions at the same time: probability distributions that reflect parameter uncertainty and frequency distributions of patient characteristics. The expected outcomes of this analysis reflects parameter uncertainty and patient heterogeneity in a heterogeneous population[5], but ignores the fundamental difference between the two.[1] This PSA does not correctly provide the distribution of the expected outcome reflecting parameter uncertainty, for a heterogeneous population.

In order to correctly separate parameter uncertainty and heterogeneity, the analysis requires a nested Monte Carlo simulation.[5] We called this method the Double Loop PSA and draw a number of individual patients within each PSA iteration. In this way we investigate sampling uncertainty, while still accounting for patient heterogeneity. The results of this analysis reflects parameter uncertainty in a heterogeneous population and will therefore lead to the required outcome.[5] In essence, this "Double Loop PSA" uses the existing Expected Value of Partial Perfect Information (EVPPI) methodology with a different goal.[7]

In this paper, we discuss the value of each of the methods for decision makers. We illustrate the differences between the approaches by applying them to a Markov model, used to assess the CE of a new treatment option for patients with Severe and Very Severe Chronic Obstructive Pulmonary Disease (COPD).

3.2 METHODS

3.2.1 Case Study

To illustrate the different ways of handling heterogeneity, we chose to use a case study in COPD. COPD is characterized by airflow limitation that is preventable and treatable but not fully reversible,[8] often accompanied by periods of increasing symptoms, called exacerbations. Patients are often treated with a long-acting β2-agonist (LABA]. A relatively new treatment option is roflumilast (Daxas®, Takeda Pharmaceuticals International GmbH). In our case study, we compared patients who used roflumilast in combination with LABA (ROFLU + LABA) with patients who used LABA alone.

This study was performed using a published Markov model, which was used to support reimbursement decision making on roflumilast in for example the United Kingdom (UK)[9-11], Switzerland[12], and Germany[13] The model was built in TreeAge Pro 2009 (TreeAge Software, Inc., Williamstown MA, USA) and adapted to the Netherlands. Where possible, the parameter estimates and frequency distribution of patient characteristics were obtained from the subgroup of patients in two one-year clinical trials of roflumilast (ROFLU) versus placebo, that had concomitant treatment with a long-acting β2-agonist (LABA subgroup).[14] The benefits of adding ROFLU to LABA were modelled by a reduction in exacerbations that require medical intervention, and by an initial improvement in lung function. The re-

duction of exacerbations was expressed as a risk ratio of 0.800 (95% confidence interval: 0,700-0,914).[14] Benefits of ROFLU were further modelled as a temporality improvement in FEV1 of 46 ml (28-64). This improvement was assumed to last for five years, after which the lung function in the LABA + ROFLU treatment arm was assumed to be the same as that in the LABA alone treatment arm.

The Markov model had a cycle length of 1 month and comprised three health states, each with different rates of exacerbations: Severe COPD, Very Severe COPD and Death (figure 3.1). COPD severity was based on the amount of air which can be forcibly exhaled from the lungs in the first second of forced exhalation, expressed as a percentage of the predicted normal value for a given patient's height, gender and age, the FEV1%pred.[8] Severe COPD was defined as 30% ≤ FEV1%pred < 50% and Very Severe COPD as FEV1%pred < 30%. Each cycle, patients could either stay in their current state, progress from Severe to Very Severe COPD or die. Transitions to less severe states were impossible. All patients started in the Severe COPD stage. During each cycle, patients could have a COPD exacerbation, which was either community-treated or required a hospital admission.

The background mortality is taken from Dutch life tables[15], which was adjusted with standardised mortality ratios (SMR) to reflect the expected increased mortality in the COPD population.[16] These SMR were obtained from comparing the mortality in the general population with the mortality among COPD patients in the Dutch population-based COPD policy model.[17,18] The case fatality rate (CFR) for hospital admissions was 7.7%.[19]

Costs (price level 2009) and utilities were assigned to the two COPD severity states. Costs and utility decrements were assigned to the exacerbations. Half cycle correction

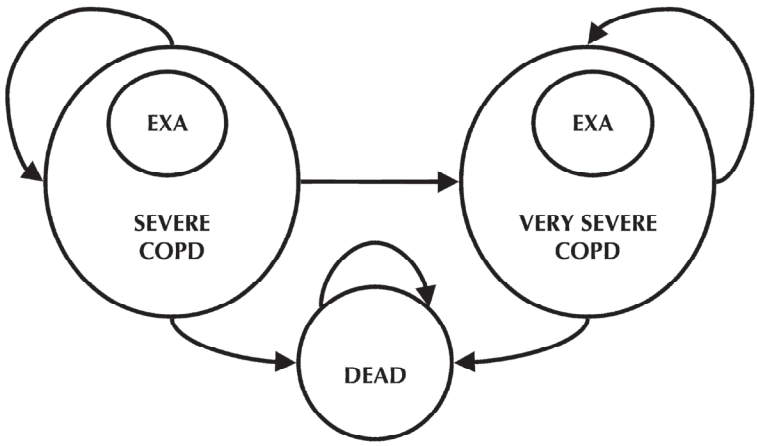

Figure 3.1: Markov model structure.[a]

[a] COPD = Chronic obstructive pulmonary disorder; EXA = exacerbation

was applied. A time horizon of 20 years was used. In accordance with Dutch guidelines, costs were discounted at 4% and health outcomes were discounted at 1.5%.[20] The CE analysis was performed from a societal perspective[20], which included all direct health care costs, the cost of productivity loss due to absence from work and the direct non-health care costs, in this case the travel and parking costs borne by patients and their families when visiting a health care provider.

3.2.2 Patient heterogeneity and sampling uncertainty

The model parameters that represent patient heterogeneity were gender, height, age at the start of treatment and the starting value for FEV1%pred (table 3.1). Parameter estimates and distributional assumptions were based on the trial data of the LABA subgroup, utilizing only patients who started with Severe COPD.[14] The proportion of males / females in the cohort was taken to be equal to the Dutch COPD patient population and was not sampled from a distribution.[21]

Gender was not considered part of the patient heterogeneity discussed here, as the model is calculated separately for male and female patients and the results are combined into a single CE outcome. As such heterogeneity due to gender is accounted for. Patient height was used, together with gender and age, to calculate the absolute FEV1, based on the cohort's FEV1%pred.[22] This is necessary, since the yearly lung function decline is modelled in absolute terms (52 ml).[23] Age at the start of treatment is measured in years, and influences the productivity costs, background mortality and the calculation of the absolute FEV1. The starting value of FEV1%pred impacts the progression of the cohort to

Table 3.1: Parameter estimates and distributional assumptions for patient heterogeneity

		Starting	Length in centimeters	
	Age[a]	FEV1%pred[b]	Female patients	Male patients
Disregard Heterogeneity	64	39.2%	161.7	171.3
Subgroup analyses				
Age 64, FEV1%pred 39%		Same as "Disregard Heterogeneity"		
Age 55, FEV1%pred 39%	55	39.2%	161.7	171.3
Age 73, FEV1%pred 39%	73	39.2%	161.7	171.3
FEV1%pred 31%, Age 64	64	31.0%	161.7	171.3
FEV1%pred 50%, Age 64	64	50.0%	161.7	171.3
Single Loop PSA	N(64, 9.1)	Trial data	N(161.7, 7.0)	N(171.3, 7.4)
Double Loop PSA		Same as "Single Loop PSA"		

[a] Based on average and standard error in the LABA subgroup.[14]

[b] Based on average in the long-acting β2-agonist (LABA) subgroup[14] and the inner boundaries for the Severe COPD stage.

a more severe stage. This in turn influences the probability to experience an exacerbation, and the resulting costs and mortality.

The model parameters for which we defined sampling uncertainty are mortality (uniform distributions), the probability that an exacerbation is treated in a hospital and utilities (beta distributions) and lung function decline, COPD exacerbation rates, utility reduction due to an exacerbation and cost parameters (all gamma distributions). An overview of all parameters can be found in the online appendix, including more detailed information on distributional assumptions.

3.2.3 Strategies to account for heterogeneity

First, we disregarded heterogeneity ("Disregard Heterogeneity") and performed a standard PSA with 2,000 random draws from the probability distribution of the model parameters. For the three patient characteristics investigated, height, age at start and starting FEV1%pred, we used the mean values of the cohort as a point estimate (table 3.1).

Next, we performed five subgroup analyses. The subgroups were cohorts defined by age, height and FEV1%pred and can be found in table 3.1. Since the effect of height on the outcomes is small, it is not used to define subgroups. For each of five subgroups of patients, we ran a PSA of 2,000 iterations. Younger patients live longer and will have more time to experience benefits, for example a higher number of future exacerbations which can be prevented. They will also experience the benefit of prevented productivity losses, until the Dutch pension age of 65. Because the exacerbation rate is higher in patients with Very Severe COPD than in patients with Severe COPD, patients with Very Severe COPD benefit more from the exacerbation-reducing effect of LABA + ROFL. It takes patients with a relatively high FEV1%pred more cycles to reach this stage, and thus experience these benefits, than the average group. In addition, LABA + ROFLU temporarily improves lung function in the intervention group, which means that patients in the intervention group stay on average about a year longer in the Severe COPD state than the patients in the control group.

The next approach was to perform a PSA with 2000 iterations, where we sampled from both the probability distributions representing sampling uncertainty and the frequency distributions representing heterogeneity at the same time ("Single Loop PSA"). Table 3.1 shows the distributional assumptions for the patient characteristics.

Finally, we performed a PSA, using a two-level sampling algorithm ("Double Loop PSA"). We used two nested levels of Monte Carlo sampling. First we drew values from probability distributions to reflect parameter uncertainty (outer loop). This set of values was kept constant when we drew values from frequency distributions (inner loop) to reflect heterogeneity in patients' height, age and FEV1%pred at start. Each of these draws within the inner loop can be interpreted as a patient with certain characteristics; the average over all patients in an inner loop as a "trial". We repeated the outer loop 2,000

times. Within each outer loop, we repeated the inner loop 30 times. In this way, we could account for both the patient heterogeneity and sampling uncertainty separately, yet at the same time. The result of this analysis is the average cost-effectiveness over all trials, for 2,000 draws from the distribution of the model parameters. It gives the distribution of the expected outcome in the heterogeneous population, reflecting the parameter uncertainty. To investigate the impact of the size of the inner and outer loop, we also performed a Double Loop PSA with an outer loop size of 300 and an inner loop size of 2,000.

3.2.4 Comparison of methods

We first compared the strategies with respect to the point estimates and 95% confidence intervals (CIs) of the outcomes. The point estimates and CIs of the difference in costs and quality adjusted life years (QALYs) were calculated as the average and the 2.5th and the 97.5th percentile from the PSA iterations. The point estimate for the incremental CE ratio (ICER) was calculated as the ratio of the two point estimates. Next, we visually compared the CE planes and described the differences. We then summarized the uncertainty from the CE planes in CE acceptability curves (CEAC). Finally, we discussed differences in calculation time, which was expected to differ greatly.

3.3 RESULTS

3.3.1 Point estimates of CE outcomes

Table 3.2 shows the point estimates of the results of the CE analysis for LABA + ROFLU versus LABA. It is clear that the CE results differ between methods. Disregard Heterogeneity produces an ICER of €10,700. Subgroup analyses show considerable differences between subgroups with lower ICERs in patients with more severe COPD and in younger patients. The ICERs range from €4,900 in the subgroup with FEV1%pred 31% and age 64, to €13,800 in the subgroup with FEV1%pred 50% and age 64. Both the Single Loop PSA and Double Loop PSA yield an ICER of €7,800 per QALY.

Table 3.2: CE results (average, 95% CI) of a long-acting β2-agonist (LABA) in combination with roflumilast versus LABA alone, using four different methods.

	Δ Costs (in €)	Δ QALYs	ICER
Disregard Heterogeneity	3,080 (1,760;3,940)	0.286 (0.147;0.438)	€10,700
Subgroup analyses[a]			
Age 64, FEV1%pred 39%	3,080 (1,760;3,940)	0.286 (0.147;0.438)	€10,700
Age 55, FEV1%pred 39%	2,200 (-760;4,050)	0.394 (0.132;0.713)	€ 5,600
Age 73, FEV1%pred 39%	2,100 (1,210;2,730)	0.165 (0.091;0.249)	€12,700
FEV1%pred 31%, Age 64	2,110 (440;3,210)	0.427 (0.251;0.619)	€ 4,900
FEV1%pred 50%, Age 64	3,370 (2,200;4,220)	0.245 (0.110;0.394)	€13,800
Single Loop PSA	2,350 (270;3,930)	0.299 (0.071;0.613)	€ 7,800
Double Loop PSA	2,330 (1,220;3,360)	0.300 (0.154;0.464)	€ 7,800

[a] Results for the subgroup "Age 64, FEV1%pred 39%" are by definition the same as for "Disregard Heterogeneity".

3.3.2 Confidence intervals

Table 3.2 also shows that the width of the 95% confidence intervals (CIs) differs considerably. The most uncertainty was found around both the difference in costs and QALYs for the subgroup with age 55 and FEV1%pred 39%, and the least uncertainty for the subgroup with age 73 and FEV1%pred 39%. There is considerably less uncertainty around the estimates in the Double Loop PSA, than in the Single Loop PSA.

3.3.3 Cost-effectiveness planes

The CE-planes in figure 3.2 show that disregarding heterogeneity (graph 2A) led to a roughly cone shaped scatter plot, with a focal point towards the top left, and spreading outwards towards the bottom right. The relationship between the difference in costs and the difference in QALYs is negative, meaning that when LABA + ROFLU is relatively more effective (more exacerbations prevented, slower disease progression), the extra costs of roflumilast are also compensated more by lower other health care costs.

When performing subgroup analyses (graph 2B), all subgroups also roughly show a cone shape. They differ in terms of size and position in the CE plane. The subgroup with age 73 and FEV1%pred 39% lies closest to the origin shows the least uncertainty. In this subgroup, no productivity costs can be gained by using a new medication. Due to a shorter life expectancy the difference in QALYs that can be gained from preventing exacerbations and postponing death is also lower.

The subgroup with age 55 and FEV1%pred 39% has the highest uncertainty and is the only one including net cost savings, which are due to productivity gains in these younger patients. The reduced progression to Very Severe COPD leads to considerable QALY gains. The subgroup with FEV1%pred 50% and age 64 lies closer to the y-axis of the graph, with a greater proportion of simulations showing a small difference in QALYs. It also further to

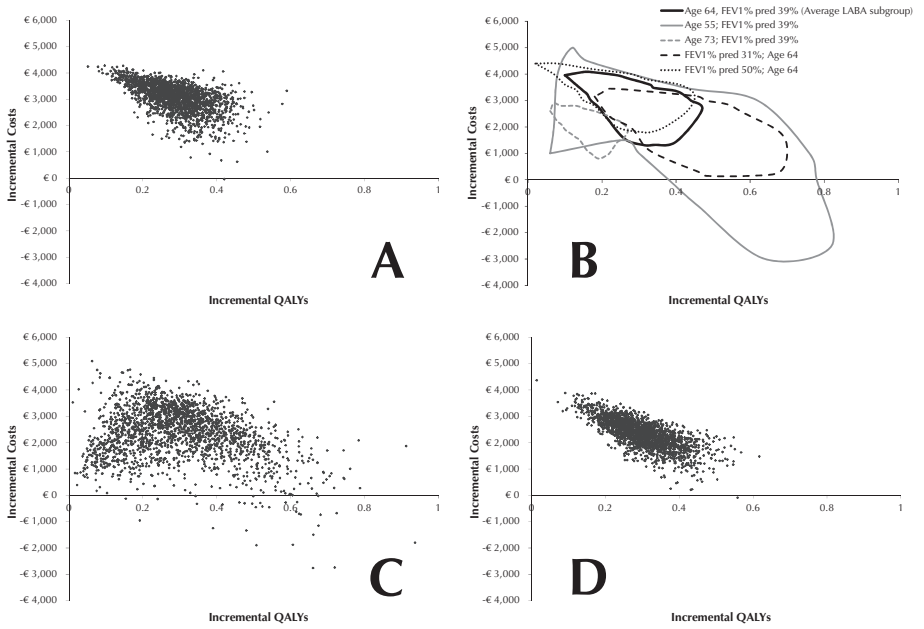

Figure 3.2: CE-plane of a long-acting β2-agonist (LABA) in combination with roflumilast versus LABA alone, comparing incremental costs to incremental QALYs, using four different methods.[a]

[a] A: Disregard Heterogeneity, 2,000 iterations; B: Subgroup analyses (approximations of 95% confidence regions shown for readability), 2,000 iterations for each subgroup; C: Single Loop PSA, 2,000 iterations; D: Double Loop PSA, outer loop 2,000, inner loop 30 iterations.

the top of the graph, with a greater proportion of simulations pointing to a high cost increase. In this group it takes relatively long for the cohort to reach the Very Severe disease stage, where real differences in costs and QALYs between LABA+ROFLU and LABA occur. Consequently the difference in health care costs and health outcomes between the two interventions are more heavily affected by discounting. The subgroup with FEV1%pred 31% and age 64 lies lower and more to the right, while the cohort defined by the average age and staring FEV1% predicted lies in the middle.

The Single Loop PSA (graph 2C), resulted in a differently shaped CE plane compared with the other methods, and is also much larger in size. It has a rough arch shape, which is explained when looking at the subgroup analyses. In essence, the Single Loop PSA is a collection of 1,000 subgroups, each with their own patient characteristics and therefore different point estimates. The left side of the arch consists mostly of people with Severe COPD and an age higher than 65, with patients with a lower starting FEV1% closer to the origin and patients with a relatively higher starting FEV1% higher in the graph. The gains from reducing exacerbations are less than in Very Severe COPD and no productivity costs can be gained by using ROFLU. The average age rises going left in the graph towards the origin, diminishing the extra costs and gains for LABA + ROFLU, since patients die earlier. Moving towards the right in the graph, we find patients with a lower age on average and

a lower starting FEV1%. Younger patients result in more productivity gains due to ROFLU and patients with a worse lung function lead to a larger difference in health care costs between the two arms due to discounting. This may offset the extra medication costs for ROFLU.

Performing a Double Loop PSA (graph 2D) produced a cone shape, which is more compact than with Disregard Heterogeneity (2A). The CE plane from the Double Loop PSA was situated slightly lower in the graph (less difference in costs) and showed a more clear negative relationship between the difference in costs and the difference in effects.

3.3.4 Double loop PSA: impact of loop size

To test whether a cohort size of 30 was appropriate, we have also performed a PSA with an outer loop size of 300 and an inner loop size of 2,000. The resulting point-estimate and shape of the CE-plane is shown in figure 3.3, with the original Double Loop PSA reproduced in figure 3.3A and the new cohort size in figure 3.3B. The shape and position of the two CE-planes are similar. The ICER with a lower cohort size is €7,500, which is very close to the ICER of the original Double Loop PSA of €7,800.

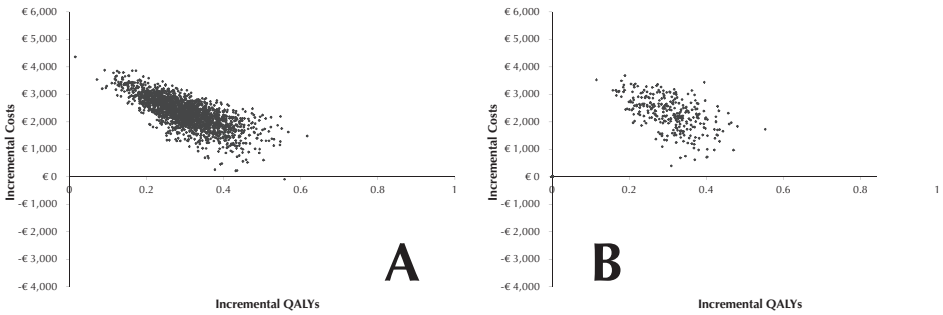

Figure 3.3: CE-plane of a long-acting β2-agonist (LABA) in combination with roflumilast versus LABA alone, comparing incremental costs to incremental QALYs, using the Double Loop PSA, with different inner and outer loop sizes.[a]

[a] A: Outer loop 2,000, inner loop 30 iterations; B: Outer loop 300, inner loop 2,000 iterations.

3.3.5 Policy decisions

The difference in the amount of uncertainty shown in the CE planes is summarized in the CEACs in figure 3.4 which show the percentage of PSA draws, where LABA + ROFLU can be considered cost-effective, for different threshold values of a QALY. The Double Loop PSA is much more certain at the three thresholds shown in table 3.3 than when disregarding heterogeneity or performing a Single Loop PSA. This is a result of relatively

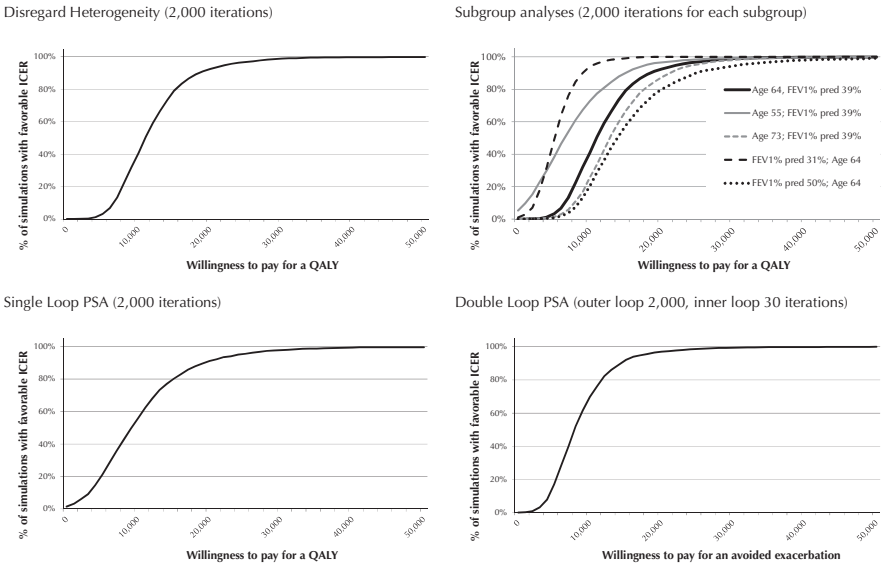

Figure 3.4: CEAC of a long-acting β2-agonist (LABA) in combination with roflumilast versus LABA alone, comparing incremental costs to incremental QALYs, using four different methods.

Table 3.3: Percentage of PSA draws where a long-acting β2-agonist (LABA) in combination with roflumilast is deemed cost-effective versus LABA alone, at different threshold values for a QALY, using four different methods.

	€10,000 / QALY	€20,000 / QALY	€40,000 / QALY
Disregard Heterogeneity	41%	92%	100%
Subgroup Analyses[a]			
Age 64, FEV1%pred 39%	41%	92%	100%
Age 55, FEV1%pred 39%	73%	97%	100%
Age 73, FEV1%pred 39%	26%	87%	100%
FEV1%pred 31%, Age 64	94%	100%	100%
FEV1%pred 50%, Age 64	20%	80%	98%
Single loop PSA	56%	91%	99%
Double loop PSA	70%	97%	100%

[a] Results for the subgroup "Age 64, FEV1%pred 39%" are the same as for "Disregard Heterogeneity".

low variation in incremental costs and effects compared to the Single Loop PSA and a lower point estimate of the ICER compared to Disregard Heterogeneity. If a policy maker would be interested in subgroup analyses it is clear the ROFLU has the highest probability to be considered cost-effective in the subgroup of very severe patients with a low FEV1% pred. and in patients with a longer life expectancy.

3.3.6 Calculation time

There were also differences in calculation times between the different methods. All calculations were done on the same dedicated computer. Both Disregard Heterogeneity and the Single Loop PSA took approximately 20 minutes to calculate, as did each of the subgroups in the subgroup analyses. With a total of five subgroups, the total analysis time is above 1.5 hours, plus the extra time in processing and interpretation. The Double Loop PSA had a much longer calculation time, close to 9 hours. The run-time of the Double Loop PSA with an outer loop of 300 and an inner loop of 2,000, was approximately 5 days.

3.4 DISCUSSION

Medical decision models are intended to inform physicians and policy makers, to aid in their choices in providing efficient health care. Since patients are by nature heterogeneous, these models should take this heterogeneity into account. In the case of Markov models, this is not straightforward. In this study, we have shown that there are several ways of dealing with heterogeneity and that the outcomes, and thus the policy decision may change when heterogeneity is handled differently. In practice, heterogeneity is often ignored. An average value of the patient population will then be used for any variables representing patient characteristics in the model. In addition to ignoring available evidence, the results are difficult to interpret since the "average patient" does not exist. It is also argued that ignoring patient heterogeneity will be costly in both monetary terms and health gains.[24,25]

An alternative is to define several subgroups of patients and to calculate the outcomes for each of these. This has been proposed as the best method for dealing with heterogeneity[4], although no comparison with the Double Loop PSA was made. Subgroup analyses did lead to insight in the differences between the different types of patients, but not all outcomes were useful for decision makers. If a decision maker wants to use the subgroup analyses for decision regarding specific subgroups, equity concerns are always an issue.[2,26] Patient heterogeneity in clinical characteristics, such as starting FEV1% in our example, may be acceptable for sub-group specific recommendations. Other input parameters, such as gender, race or in this case age, are not.[2]

Heterogeneity has sometimes been handled by combining it with parameter uncertainty in a PSA.[27-29] Using the Single Loop PSA, one is able to consider many different types of patients, without a significant increase in calculation time. However, because of a larger number of parameters being sampled, the amount of uncertainty increases and the shape of the CE-plane also changes compared to disregarding heterogeneity. The outcomes from the Single Loop PSA reflect both parameter uncertainty and patient heterogeneity, but it ignores the fundamental difference between the two. The expected outcome for the Single Loop PSA is correct for the population and equals the expected outcome for the

Double Loop PSA.[5] However, the distribution of the expected outcome, that reflects the uncertainty in which many decision makers are interested, is not correct. In order to correctly separate parameter uncertainty and heterogeneity, the analysis requires a nested Monte Carlo simulation.[5]

In the Double Loop PSA we drew a number of individual patients for each of the PSA draws and calculate the average CE outcomes. This method accounts sufficiently for heterogeneity, is easily interpretable and can be performed using existing software. It was not our main goal to minimize uncertainty around our CE outcomes, but the Double Loop PSA did lead to the smaller CIs. They are a correct reflection of the parameter uncertainty around the expected outcome, for this heterogeneous populations. This reduction occurred because each dot in the Double Loop PSA is an average of 30 patients. For each patient, the relationship between the difference in costs and the difference in effects is more of less the same, as we could see in the subgroup analyses, which all generally have a downward sloping cone shape. The Double Loop PSA is a combination of 2,000 of these subgroups, which leads to a more defined relationship between the difference in costs and effects. In the Double Loop PSA, extreme values of heterogeneity parameters can be drawn, but are averaged out in the inner loop, while in the Single Loop PSA, these extremes are shown.

Calculation time may be a burden, even with the relatively small cohort size of 30. On the other hand, a calculation time of 9 hours (one overnight calculation) is not a huge obstacle, in our opinion. Another drawback is that this method cannot be used to directly inform subgroup-specific policy decisions. When a policy maker wants or needs to consider subgroups, for example defined by disease severity classes, one solution may be to perform a Double Loop PSA within each subgroup. This however, will increase the computational burden.

We have chosen for a small sample of 30, because larger sample sizes rapidly increases the computation time. This so-called "M by N problem", where the calculation time increases due the number of inner calculations, is a major obstacle to performing PSAs in for example patient level models.[3,30,31] We tested to see whether our cohort size of 30 was appropriate for our needs, by also performing a Double Loop PSA with an outer loop size of 300 and an inner loop size of 2,000. The run-time of this model was approximately 5 days, which made it impractical in daily use. The resulting point-estimate and shape of the CE-plane were similar to Double Loop PSA with a smaller cohort size. We therefore concluded that 30 would be a good middle ground between accuracy and runtime. However, we acknowledge that a larger number of patients in the inner loop might improve results. Fortunately, since computational speed increases rapidly, it is likely that using faster, more modern computers would decrease the necessary time. Since the model was build, a newer version of TreeAge has also been released, which might also

increase computational speed. This means that a higher number of patients in the inner loop is becoming possible.

In the current model, the COPD health states were defined by lung function and the probability to get a COPD exacerbation was estimated for each state.[8] The model was constructed before the revision of the GOLD strategy document in 2011, in which a classification of COPD based on three characteristics, symptoms, lung function and exacerbation history, was proposed. The latter increases the number of variables used to describe the heterogeneity of a patient population and -when the possible combinations of these 3 characteristics are taken- the number of subgroups. This reinforces the need to separate heterogeneity from parameter uncertainty. However, applying the latest classification is unlikely to have a big impact our point estimate of the ICER in the double loop PSA, because the sample as a whole does not change, only the way it is classified.

The uncertainty around the standardized mortality ratio and the case fatality ratio was modelled with a uniform distribution. It could be argued that a more natural choice might be a lognormal or a beta distribution, respectively. Since the model was built to support the reimbursement dossier in several countries, it was decided not to make any changes to the provided distributions, to preserve consistency with these dossiers. It was expected that these changes would not have impacted our study results.

3.5 CONCLUSION

To conclude, we think that three of the methods discussed can be useful in CE research, each in different circumstances. When little or no heterogeneity is expected, or when it is not expected to influence the CE results, disregarding heterogeneity may be correct. In our case study, heterogeneity did have an impact. Subgroup analyses may inform policy decisions on each subgroup, as long as they are well defined and the characteristics of the cohort that define a subgroup truly represent the patients within that subgroup. Despite the necessary calculation time, the Double Loop PSA is a viable alternative which leads to better results and better policy decisions, when accounting for heterogeneity in a Markov model. The Single Loop PSA can only be used to calculate the point estimate of the expected outcome. It disregards the fundamental differences between heterogeneity and sampling uncertainty and overestimates uncertainty as a result.

3.6 LITERATURE

[1] Briggs AH, Weinstein MC, Fenwick EA, Karnon J, Sculpher MJ, Paltiel AD, and the ISPOR-SMDM Modeling Good Research Practices Task Force. Model Parameter Estimation and Uncertainty: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-6. Med Decis Making. 2012 Sep-Oct;32(5):722-32.

[2] Grutters JP, Sculpher M, Briggs AH, Severens JL, Candel MJ, Stahl JE, et al. Acknowledging patient heterogeneity in economic evaluation : a systematic literature review. Pharmacoeconomics 2013 Feb;31(2):111-123.

[3] Groot Koerkamp B, Weinstein MC, Stijnen T, Heijenbrok-Kal MH, Hunink MG. Uncertainty and patient heterogeneity in medical decision models. Med Decis Making 2010 Mar-Apr; 30(2):194-205.

[4] Briggs A, Scuplher M, Claxton K. Decision Modelling for Health Economic Evaluation. Oxford: Oxford University Press; 2006.

[5] Groot Koerkamp B, Stijnen T, Weinstein MC, Hunink MG. The combined analysis of uncertainty and patient heterogeneity in medical decision models. Med Decis Making 2011 Jul-Aug;31(4):650-661.

[6] Bolin K, Mörk A, Willers S, Lindgren B. Varenicline as compared to bupropion in smokingcessation therapy—Cost–utility results for Sweden. Respir Med 2008;102(5):699-710.

[7] Claxton K. Value of information analysis. In: Sculpher M, Briggs AH, Claxton K, editors. Course book: Advanced modelling methods for health economic evaluation Oxford; 2005.

[8] Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global strategy for the diagnosis, management and prevention of chronic obstructive pulmonary disease. Updated 2009. 2009; Available at: http://www.goldcopd.com/. Accessed April 21th, 2010.

[9] The cost-effectiveness of roflumilast in the management of severe COPD in the UK setting. (Poster PRS18). International Society for Pharmacoeconomics and Outcomes Research (ISPOR), 16th Annual International Meeting; May 21-25, 2011; ; 2011.

[10] Hertel N, Kotchie RW, Samyshkin Y, Radford M, Humphreys S, Jameson K. Cost-effectiveness of available treatment options for patients suffering from severe COPD in the UK: a fully incremental analysis. Int J Chron Obstruct Pulmon Dis 2012;7:183-199.

[11] Samyshkin Y, Kotchie RW, Mork AC, Briggs AH, Bateman ED. Cost-effectiveness of roflumilast as an add-on treatment to long-acting bronchodilators in the treatment of COPD associated with chronic bronchitis in the United Kingdom. Eur J Health Econ 2013 Feb 8.

[12] Samyshkin Y, Schlunegger M, Haefliger S, Ledderhose S, Radford M. Cost-effectiveness of roflumilast in combination with bronchodilator therapies in patients with severe and very severe COPD in Switzerland. Int J Chron Obstruct Pulmon Dis 2013;8:79-87.

[13] Nowak D, Ehlken B, Kotchie R, Wecht S, Magnussen H. Roflumilast in combination with long-acting bronchodilators in the management of patients with severe and very severe COPD. A cost-effectiveness analysis for Germany. Dtsch Med Wochenschr 2013 Jan;138(4):119-125.

[14] Calverley PM, Rabe KF, Goehring UM, Kristiansen S, Fabbri LM, Martinez FJ, et al. Roflumilast in symptomatic chronic obstructive pulmonary disease: two randomised clinical trials. Lancet 2009 Aug 29;374(9691):685-694.

[15] Statistics Netherlands (CBS). Statline Databank [Statline database]. 2012; Available at: http://statline.cbs.nl/. Accessed 01/12, 2012.

[16] Spencer M, Briggs AH, Grossman RF, Rance L. Development of an economic model to assess the cost effectiveness of treatment interventions for chronic obstructive pulmonary disease. Pharmacoeconomics 2005;23(6):619-37.

[17] Hoogendoorn M, Rutten-van Mölken MPMH, Hoogenveen RT, van Genugten MLL, Buist AS, Wouters EFM, et al. A dynamic population model of disease progression in COPD. Eur Respir J 2005;26(2):223-233.

[18] Hoogendoorn M, Hoogenveen RT, Rutten-van Molken MP, Vestbo J, Feenstra TL. Case fatality of COPD exacerbations: a meta-analysis and statistical modelling approach. Eur Respir J 2011 Mar;37(3):508-515.

[19] Royal College of Physicians, British Thoracic Society, British Lung Foundation. Report of the National Chronic Obstructive Pulmonary Disease Audit 2008: clinical audit of COPD exacerbations admitted to acute NHS units across the UK, 2008

[20] College voor Zorgverzekeringen. Richtlijnen voor farmaco-economisch onderzoek; evaluatie en actualisatie [Guidelines for pharmacoeconomic research: evaluation and actualization]. 2008.

[21] van Baal PH, Engelfriet PM, Hoogenveen RT, Poos MJ, van den Dungen C, Boshuizen HC. Estimating and comparing incidence and prevalence of chronic diseases by combining GP registry data: the role of uncertainty. BMC Public Health 2011 Mar 15;11:163.

[22] Crapo RO, Morris AH, Gardner RM. Reference spirometric values using techniques and equipment that meet ATS recommendations. Am Rev Respir Dis 1981;123(6):659-659-664.

[23] Scanlon PD, Connett JE, Waller LA, Altose MD, Bailey WC, Buist AS. Smoking cessation and lung function in mild-to-moderate chronic obstructive pulmonary disease. The Lung Health Study. Am J Respir Crit Care Med 2000;161:381-90.

[24] Coyle D, Buxton MJ, O'Brien BJ. Stratified cost-effectiveness analysis: a framework for establishing efficient limited use criteria. Health Econ 2003 May;12(5):421-427.

[25] Basu A, Meltzer D. Value of information on preference heterogeneity and individualized care. Med Decis Making 2007 Mar-Apr;27(2):112-127.

[26] Sculpher M. Subgroups and heterogeneity in cost-effectiveness analysis. Pharmacoeconomics 2008;26(9):799-806.

[27] Perman G, Rossi E, Waisman GD, Aguero C, Gonzalez CD, Pallordet CL, et al. Cost-effectiveness of a hypertension management programme in an elderly population: a Markov model. Cost Eff Resour Alloc 2011 Apr 5;9(1):4-7547-9-4.

[28] Anderson BR, McElligott S, Polsky D, Vetter VL. Electrocardiographic Screening for Hypertrophic Cardiomyopathy and Long QT Syndrome: The Drivers of Cost-Effectiveness for the Prevention of Sudden Cardiac Death. Pediatr Cardiol 2013 Sep 5.

[29] Bentley A, Gillard S, Spino M, Connelly J, Tricta F. Cost-utility analysis of deferiprone for the treatment of beta-thalassaemia patients with chronic iron overload: a UK perspective. Pharmacoeconomics 2013 Sep;31(9):807-822.

[30] Halpern EF, Weinstein MC, Hunink MG, Gazelle GS. Representing both first- and second-order uncertainties by Monte Carlo simulation for groups of patients. Med Decis Making 2000 Jul-Sep;20(3):314-322.

[31] Claxton K, Sculpher M, McCabe C, Briggs A, Akehurst R, Buxton M, et al. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. Health Econ 2005 Apr; 14(4):339-347.

[32] Rutten-van Mölken MPMH, Hoogendoorn M, Lamers LM. Holistic preferences for 1-year health profiles describing fluctuations in health: the case of chronic obstructive pulmonary disease. Pharmacoeconomics 2009;27(6):465-477.

[33] CVZ. Farmacotherapeutisch Kompas [Pharmacotherapeutic Compass]. 2010; Available at: http://www.fk.cvz.nl/. Accessed 02/25, 2010.

[34] College voor Zorgverzekeringen, Geneesmiddelen Informatie Project. GIPdatabank. 2010; Available at: http://www.gipdatabank.nl/. Accessed 2/25, 2010.

[35] Oostenbrink JB, Rutten-van Mölken MPMH, Monz BU, FitzGerald JM. Probabilistic Markov model to assess the cost-effectiveness of bronchodilator therapy in COPD patients in different countries. Value Health 2005 Jan-Feb;8(1):32-46.

[36] LMR. Landelijke LMR-informatie - Diagnosen [National medical registration Information - Diagnoses]. 2010; Available at: http://cognosserver.prismant.nl/cognos7/cgi-bin/ppdscgi.cgi?DC=Q&E=/Prisma-Landelijke-LMR/Landelijke+LMR-informatie+-+Diagnosen. Accessed 02/25, 2010.

[37] RIVM. Zorgbalans 2008 [Care balance 2008]. 2008; Available at: http://www.rivm.nl/vtv/object_document/o5274n32928.html. Accessed 02/25, 2010.

[38] Oostenbrink JB, Bouwmans CAM, Koopmanschap MA, Rutten FFH. Handleiding voor kostenonderzoek, methoden en richtlijnprijzen voor economische evaluaties in de gezondheidszorg (geactualiseerde versie 2004) [Guideline for cost research, methods and prices for economic evaluations in health care]. Diemen: College voor zorgverzekeringen (CVZ); 2004.

[39] Heijmans MJWM, Spreeuwenberg P, Rijken PM. Monitor zorg- en leefsituatie van mensen met astma en mensen met COPD. Ontwikkelingen en trends over de periode 2001 - 2004 [Monitor care and living situation of people with asthma and people with COPD. Developments and trends during 2001-2004]. 2005.

[40] Hoogendoorn M, van Wetering CR, Schols AM, Rutten-van Molken MP. Is INTERdisciplinary COMmunity-based COPD management (INTERCOM) cost-effective? Eur Respir J 2010 Jan; 35(1):79-87.

[41] NPCG. Patientenpanel chronisch zieken, Astma en COPD monitor [Patient panel chronically ill, Asthma and COPD Monitor]. 2010; Available at: http://www.nivel.nl/npcg/. Accessed 02/25, 2010.

[42] Koopmanschap MA, Rutten FFH, van Ineveld BM, van Roijen L. The friction cost method for measuring indirect costs of disease. J Health Econ 1995;14:171-89.

[43] Rutten-van Mölken MPMH editor. Van Kosten tot Effecten Een handleiding voor evaluatiestudies in de gezondheidszorg [From Costs to Effects. A manual for evaluation studies in health care]. 2nd ed. Maarssen: ELSEVIER gezondheidszorg; 2010.

[44] Calverley PM, Pauwels R, Vestbo J, Jones P, Pride N, Gulsvik A, et al. Combined salmeterol and fluticasone in the treatment of chronic obstructive pulmonary disease: a randomised controlled trial. Lancet 2003 Feb 8;361(9356):449-456.

[45] Buckingham RJ, Lowe D, Pursey NA, Roberts CM, Stone RA, On behalf of the National COPD Audit 2008 Steering Group. Report of The National Chronic Obstructive Pulmonary Disease Audit 2008: clinical audit of COPD exacerbations admitted to acute NHS units across the UK. 2008.

A3 APPENDIX

A3.1 Input data used in the model

Information on epidemiological data as used in the model, and efficacy of roflumilast, can be found in table A3.1.

A3.2 Utilities

EQ-5D utility weights for stable COPD, which can be found in table A3.2, were derived from the LABA-subgroup analysis.[14] Utility decrements from experiencing either a community-treated or a hospital-treated exacerbation were obtained from a Dutch study in which one-year COPD health profiles were valued by the general population using the time trade-off method.[32] The entire utility decrement is applied at the time of the exacerbation.

A3.3 Direct health care costs

Monthly costs of roflumilast were calculated using a daily price for the Netherlands provided by the manufacturer of € 1.38. The monthly costs of LABA was calculated as the weighted average of the monthly costs of long-acting β2-agonists and inhaled corticosteroids in separate devices and in fixed combinations, as published in the Pharmacotherapeutic Compass.[33] In both calculations, IMS panel data on the number of prescriptions of the various medications among COPD patients in the Netherlands were used as weights.[34], VAT of 6% and a Dutch mark-up on the price per prescription of € 7.28 to cover pharmacy-expenses (*receptregelvergoeding*) was included in the calculation of all medication costs. It was assumed that each patient received four prescriptions per year.

Direct health care costs of maintenance treatment and COPD exacerbations were calculated using data from Oostenbrink et al. who reported the costs of maintenance treatment by GOLD stage and the costs of non-severe and severe COPD exacerbations.[35] These costs were updated to the price level of 2009 using consumer price indices.[15] In addition, the length of hospital stay for a severe COPD exacerbation was adjusted from 11 days as reported by Oostenbrink et al. to 9.5 days as obtained from the most recent LMR data, calculated as the weighted average of the length of stay for ICD-9 codes 490, 491, 492, 494 and 496.[36] Furthermore, the costs of maintenance treatment as reported by Oostenbrink et al were adjusted by excluding the costs of theophylline and corticosteroids, because both medications were not allowed during the trials.[14]

A3.4 Direct non-health care costs

Direct non-health care costs are made up of travel costs to and from health care providers. Severe and Very Severe COPD patients were assumed not to visit a GP for their regular control visits, but rather a pulmonary specialist in a hospital. For stable severe COPD,

we have used two outpatient visits per year to the pulmonary specialist, and four for very severe COPD. A community-treated exacerbations required 0.34 visits to the pulmonary specialist and 0.66 visits to the GP.[35] Hospital-treated exacerbations required 1 inpatient admission, 0.82 visits to the pulmonary specialist and 0.70 visits to the GP.[35] The average distance to the hospital in the Netherlands is 7 km and 1.1 km to the GP.[37] Unit costs per kilometer of travel by car, public transportation and taxis, as well as starting tariffs for the taxi, were based on by standard tariffs.[38] It was assumed that one third of the travel for control visits by patients with Severe COPD was done by car, one third by public transport and one third by taxi. For Very Severe COPD and all exacerbations, the mix was half car, half taxi. Family of patients who experienced a hospital-treated exacerbations, were assumed to visit the patient every other day, which amounted to 5.5 visits.[35] Family was assumed to travel half of these visits by car and half by public transportation.

A3.5 Productivity costs

In the model it is assumed that 40% of the COPD patients younger than 55 years had paid employment. This is reduced to 20% at age 55 and to 0% at age 65. Given the age and gender distribution of the patients in the model this would lead to an overall percentage of COPD patients with paid employment of 14.1%.[39] This percentage was close to the percentage observed in the INTERCOM trial (14%), a Dutch randomized trial on the effects of an interdisciplinary community-based COPD program for patients with impaired exercise capacity.[40]

In the model productivity days were lost because of death or absenteeism during an exacerbation, among patient with paid employment. Death is caused by both background mortality and the case fatality of a hospital-treated exacerbation. Absenteeism was assumed to be either 7 days for a community-treated exacerbation and 21 days for a hospital-treated exacerbation.[41] In accordance to Dutch guidelines, productivity costs were calculated using the friction cost method.[20] This method accounts for the fact that almost all employees are replaceable in the work force.[42] The length of the period it takes for a person to be replaced (i.e. the friction period) is dependent on the time it takes to fill job openings. This was estimated to be 156 days in the Netherlands.[38] The productivity loss per day was calculated using the average productivity loss per hour for a male employee of € 31 and € 25.33 for a female employee.[43] Using the gender distribution from the LABA subgroup, the weighted average productivity loss is € 29.55 per hour.[14,38] We assumed a full work day of 8 hours.

Table A3.1: Parameter estimates and distributional assumptions for treatment arms.[a]

Variable	Point estimate	Distribution	Source
FEV1 decline per year in the COPD population in litres	0.052	Gamma(mean 0.052, se 0.001)	23
FEV1 improvement due to ROFLU in litres			
LABA + ROFLU versus LABA	0.046	Normal(mean 0.046, s.e.0.009)	14
Estimated duration of FEV1 improvement in years	5	-	Assumption
Rate of COPD exacerbations per year			
Severe COPD (LABA)	1.606	Gamma(mean 1.606, se 0.113)	14
Very Severe COPD (LABA)	1.910	Gamma(mean 1.910, se 0.235)	14
Relative risk for all COPD exacerbations due to ROFLU			
LABA + ROFLU versus LABA	0.8	Lognormal(mean -0.223, s.e. 0.068)	14
Standardised mortality ratio (SMR)			
Severe COPD	2.779	Uniform(2.4-3.2)	17,18, Assumption uniform 85% - 115% of mean
Very Severe state COPD	3.572	Uniform(3.0-4.1)	17,18,44, Assumption uniform 85% - 115% of mean
Proportion of exacerbations that are hospitalized			
Severe COPD	0.142	Beta(mean 0.142, se 0.011)	14
Very Severe COPD	0.226	Beta(mean 0.226, se 0.017)	14
Case fatality rate for hospital admissions	7.7%	Uniform(0.054-0.100)	45, Assumption uniform 80% - 120% of mean.

[a] ROFLU = roflumilast, LABA = long acting β2-agonist, se = standard error.

Table A3.2: Parameter estimates and distributional assumptions for utilities for stable COPD and utility decrements for a COPD exacerbation.[a]

Variable	Point estimate	Distribution	Source
Utility for stable COPD			
Severe COPD	0.751	Beta(mean 0.751, se 0.007)	14
Very Severe COPD state	0.657	Beta(mean 0.657, se 0.011)	14
Annual utility reduction in case of 1 exacerbation per year			
Community-treated exacerbation	0.01	Gamma(mean 0.010, se 0.007)	32
Hospital-treated exacerbation	0.042	Gamma(mean 0.042, se 0.009)	32

[a] se = standard error

Table A3.3: Parameter estimates and distributional assumptions for direct healthcare and non-healthcare costs.[a]

Variable	Point estimate	Distribution	Source
Drug costs			
Monthly cost of LABA	€29.49	-	39,40
Monthly cost of roflumilast	€47.07	-	Takeda
Cost of one month COPD maintenance treatment			
Severe COPD	€33.78	Gamma(mean 33.78, se 3.1)	35
Very Severe COPD	€107.41	Gamma(mean 107.41, se 13.3)	35
Cost of COPD exacerbations			
Community-treated COPD exacerbations	€83.07	Gamma(mean 83.07, se 6.8)	35
Hospital-treated COPD exacerbations	€2,997.07	Gamma(mean 2,997.07, se 802.1)	35
Length of stay in days for a hospital-treated exacerbation[b]	9.5	-	36
Travel costs			
Per month for a patient with severe COPD.	€2.23	Gamma(mean 2.23, se 0.3)	37,43, assumption se 15% of mean.
Per month for a patient with very Severe COPD.	€6.22	Gamma(mean 6.22, se 0.9)	37,43, assumption se 15% of mean.
For a patient experiencing a community-treated exacerbation	€7.90	Gamma(mean 7.90, se 1.2)	37,43, assumption se 15% of mean.
For a patient experiencing a hospital-treated exacerbation	€35.71	Gamma(mean 35.71, se 5.4)	37,43, assumption se 15% of mean.
For the family of a patient experiencing a hospital-treated exacerbation	€25.50	Gamma(mean 25.50, se 3.8)	37,43, assumption se 15% of mean.

[a] ROFLU = roflumilast, LABA = long-acting β2-agonist, se = standard error

[b] Not a separate variable in the model, only used to calculate other cost variables

Table A3.4: Parameter estimates and distributional assumptions for productivity costs.[a]

Variable	Point estimate	Distribution	Source
Proportion of patients < 55 yrs old with paid employment	40%	Beta(mean 0.4, se 0.1)	39, assumptions
Proportion of patients 55-64 years old with paid employment	20%	Beta(mean 0.2, se 0.1)	39, assumptions
Proportion of patients > 64 years with paid employment	0	-	Assumption
Retirement age in years[b]	65	-	Assumption
Days of absenteeism due to a community-treated exacerbation[b]	7	Gamma(mean 7.0, se 1.1)	41, assumption, se 15% of mean.
Days of absenteeism due to a hospital-treated exacerbation[b]	21	Gamma(mean 21.0, se 3.2)	41, assumption, se 15% of mean.
Labor costs per hour[c]	€29.55	Gamma(mean 29.55, se 4.4)	14,43, assumption, se 15% of mean.
Duration of friction cost period in days	156	-	43

[a] se = standard error

[b] Not a variable in the model, only used to calculate other variables

[c] Using average productivity loss per hour, gender distribution from LABA subgroup and assuming a work day of 8 hours.

Chapter 4

A choice that matters?

A simulation study on the impact of direct meta-analysis methods on health economic outcomes

P. Vemer, M.J. Al, M. Oppe, M.P.M.H. Rutten-van Mölken

Previously published in PharmacoEconomics (2013) 31:719–730.
DOI 10.1007/s40273-013-0067-0

Acknowledgments: The authors wish to acknowledge the support of Prof. Lidia Arends, Ph.D. and Prof. Jos Kleijnen, Ph.D. for their valuable contributions to the design and execution of this study. The authors would also like to thank four anonymous reviewers for their valuable comments on the final manuscript.

ABSTRACT

Background Decision-analytic cost-effectiveness (CE) models combine many different parameters like transition probabilities, event probabilities, utilities and costs, which are often obtained after meta-analysis. The method of meta-analysis may affect the CE estimate.

Aim Our aim was to perform a simulation study that compares the performance of different methods of meta-analysis, especially with respect to model-based health economic (HE) outcomes.

Methods A reference patient population of 50,000 was simulated from which sets of samples were drawn. Each sample drawn represented a clinical trial comparing two fictitious interventions. In several scenarios, the heterogeneity between these trials was varied, by drawing one or more of the trials from predefined subpopulations. Parameter estimates from these trials were combined using frequentist fixed (FFE) and random effects (FRE), and Bayesian fixed (BFE) and random effects (BRE) meta-analysis. The pooled parameter estimates were entered into a probabilistic cost-effectiveness Markov model. The four methods of meta-analysis resulted in different parameter estimates and HE outcomes, which were compared with the true values in the reference population. Performance statistics were: (1) the percentage of repetitions that the confidence interval of the probabilistic sensitivity analysis covers the true value (coverage), (2) the difference between the estimated and true value (bias), (3) the mean absolute value of the bias (MAD) and (4) the percentage of repetitions that result in a statistically significant difference between the two interventions (statistical power). As the differences between methods could be due to chance, we repeated every step of the analysis 1,000 times to study whether differences were systematic.

Results FFE, FRE and BFE lead to different parameter estimates, but, when entered into the model, they do not lead to large differences in the point estimates of the HE outcomes, even in scenarios where we built in heterogeneity. Random effects methods do not necessarily reduce bias when heterogeneity is added to the trials, and may even increase bias in certain situations. BRE tends to overestimate uncertainty reflected in the CE acceptability curve.

Conclusion FFE, FRE and BFE lead to comparable HE outcomes. BRE tends to overestimate uncertainty. Based on this study, we recommend FRE as the preferred method of meta-analysis.

4.1 INTRODUCTION

In 2006, the Netherlands implemented a policy of conditional, temporary reimbursement of potentially innovative, but expensive hospital drugs.[1] Additional hospital funding is provided on the condition that outcomes research is performed to show further evidence of the value of the new drugs. The final reimbursement decision is made based on all evidence available, after 4 years. A systematic approach to aid decision making is called comprehensive decision modelling, in which available evidence is structured in a probabilistic decision-analytic model.[2, 3] Meta-analysis is one step in this process, and is used to combine all available evidence in model parameters. A wide range of model parameters need to be estimated, from transition probabilities to costs and utility values.[4]

Many different methods of meta-analysis exist, and many authors have compared them (e.g. [5–8]). They have shown that the choice of method can considerably affect parameter estimates. These comparisons concentrated on the impact of the method of meta-analysis on the estimate of a single treatment effect, for example a risk ratio (RR). However, in the probabilistic models used in economic evaluations we need to estimate many different parameters, including the baseline value of each model parameter in the comparator group. Altogether, the method of meta-analysis to obtain these parameters may considerably affect the final cost-effectiveness (CE) estimates.

Our group has previously investigated the effect of four different methods of meta-analysis on model-based CE estimates.[9] Although we found considerable differences, there was no way of knowing which of the methods was best, because we had no 'truth' to which we could compare our results. That is, we only had data from different samples of the total patient population, not of the population itself. To overcome this problem we performed a simulation study, in which we created a reference population, which reflected the value that should be obtained by the different methods. We then proceeded by drawing sets of samples from this population, mimicking sets of clinical trials, and combined these trial estimates. Each method of meta-analysis generated a separate set of pooled parameters. We filled a health economic (HE) model with these different sets of parameters and investigated whether there were systematic differences between the meta-analysis methods by comparing the outcomes of the sets of samples with the outcomes of the reference population. We were especially interested in the impact on the differences in costs and quality-adjusted life years (QALYs), the incremental CE ratio and the CE acceptability curve.

The available methods of meta-analysis can be divided into two groups, namely direct and indirect methods. Direct methods of meta-analysis combine evidence from trials that compare the two interventions of interest directly. In the absence of head-to-head studies, or with the availability of both direct and indirect evidence, indirect methods of meta-

analysis come into play. Methods of indirect meta-analysis are compared in chapter 5. We therefore focus here on direct meta-analysis methods.

4.2 METHODS

4.2.1 Simulation study

The simulation comprised several steps, shown in figure 4.1. In **step 1 (Create reference population)**, we simulated a reference patient population (n = 50,000), including individual patient-level disease progression using one of two fictitious treatments. The mean values of the parameters and HE outcomes, as calculated from the entire population, are reference values to which we compared the estimates of the meta-analyses. In other words, they represented the 'truth' and are referred to as reference parameters and reference outcomes. Parameters included transition probabilities, probabilities to experience an event, maintenance costs, utilities and costs and utility-decrements due to an event. HE outcomes included the total number of QALYs, life years (LY) and events, intervention and maintenance costs, and the incremental CE ratio (ICER).

In **step 2 (Trial selection)**, we sampled trials from the reference population, comparing the two treatments. For each of the trials we calculated the parameters that are needed as input for the HE model, called trial parameters. In **step 3 (Meta-analysis)**, we pooled

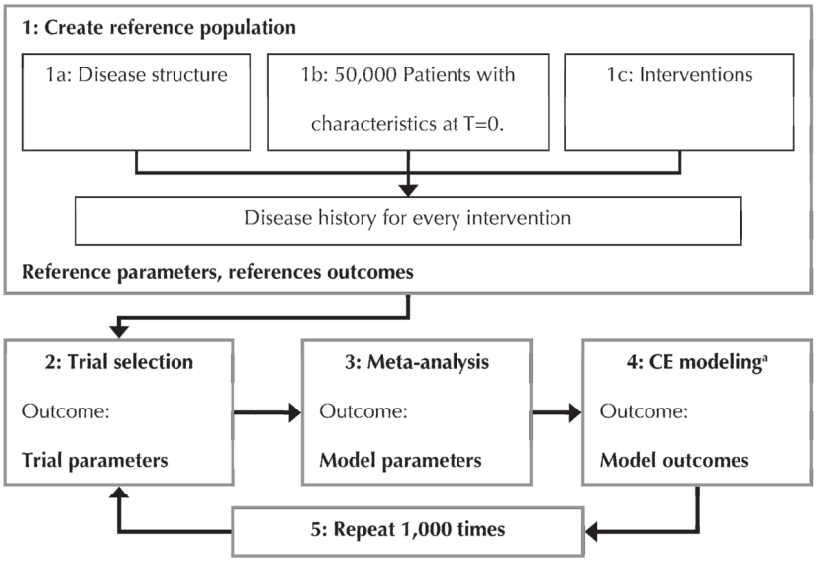

Figure 4.1: Design of the simulation study.
[a] HE: health-economic, CE: cost-effectiveness

the trial parameters using several methods of meta-analysis. These methods are explained in detail in paragraph 4.2.4. The combined estimates are called model parameters. For each model parameter, both mean and appropriate dispersion measures were calculated. We used a disease progression model in **step 4 (CE modeling)**, filled first with a set of model parameters obtained by the first method of meta-analysis. A probabilistic sensitivity analysis (PSA; 1,100 iterations) was run and the HE outcomes, called model outcomes, were collected. This process was repeated with model parameters obtained from each of the methods of meta-analysis.

Differences in model outcomes could be due to chance, i.e. the particular set of trials that was drawn. In order to study whether there was a systematic difference between the methods of meta-analysis, we repeated steps 2 to 4 in **step 5 (Repeat)**, further referred to as 1,000 repetitions.

4.2.2. Disease and model structure

The modeled disease was a progressive, chronic disease (figure 4.2), with events during which symptoms worsen considerably. The disease was simulated using a Markov model with four stages: moderate, severe and very severe disease, and death.

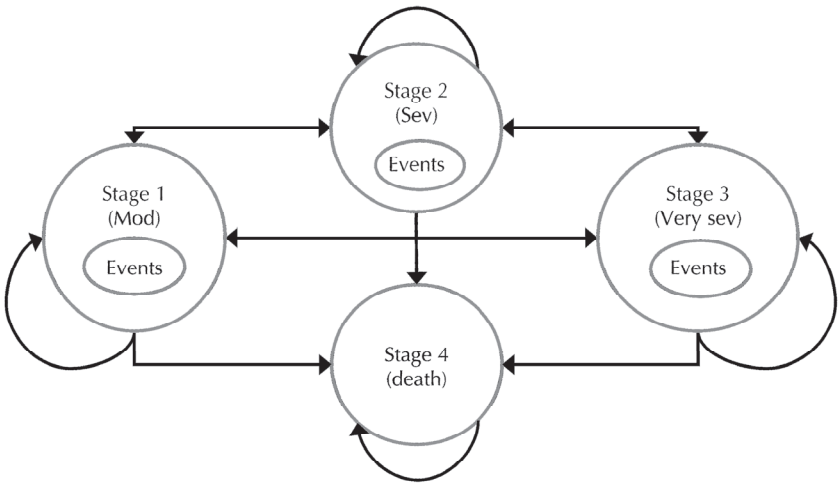

Figure 4.2: Markov model of the chronic disease.

For each patient in the reference population, we simulated their disease progression. We did this by first defining the reference disease progression (RDP), which can be thought of as the disease progression of an untreated, base-case patient. It consists of a set of distributions for each reference parameter (table 4.1). Next, these distributions were modified

Table 4.1: Characteristics of the simulated patient population.

Size simulated cohort	50,000
Starting disease stage	5/8 in moderate, 2/8 in severe and 1/8 in very severe
Gender	50% male, 50% female
Age in years	18 – 34; 35 – 64; 65+
	Determined by a random draw from a uniform distribution from 18 to 75
Developed/developing country.	50% from developed countries, 50% developing countries
Body Mass Index (BMI)	<25 (average or low); 25-30 (high); >30 (obese),
	Determined by a random draw from a normal distribution with mean 23 and standard deviation of 4.
Smoking status	30% smokers, 70% non-smokers

based on individual patient characteristics—gender, age, developed/developing country, body-mass index (BMI) and smoking status and interventions. These characteristics made it possible for us to add heterogeneity to trials in relevant scenarios, by sampling from sub-populations. In this manner we have simulated a heterogeneous population of individual patients.

How patient characteristics and interventions influenced the RDP is stated in appendix A4. In short, male patients have a higher probability to move to a worse disease stage than female patients. Older patients have a higher probability to move to a worse disease stage than younger patients; they have higher costs and a wider spread in quality-of-life weights. Patients from developing countries have lower maintenance costs than patients from developed countries. Patients with a higher BMI have a higher probability to move to a worse disease stage than patients with lower BMI; they also have a higher probability of an event, higher maintenance costs and lower quality of life. Patients who smoke have a higher probability to move to a worse disease stage and a higher probability of an event, than patients who do not smoke.

Interventions influence the RDP in the same manner as patient characteristics do. For each patient in the reference population, we simulated their disease progression twice: once receiving Usual Care and once receiving the New Intervention. Usual Care is a drug that decreases the probability of disease progression compared with the RDP, at € 60 per monthly cycle. New Intervention, the focus of the HE analysis, decreases the probability of disease progression, more so than Usual Care, plus it increases the probability of moving to a better disease stage and decreases the probability of an event. The costs were set at € 350 per monthly cycle. In the HE model, probabilities for the New Intervention are modeled as a RR, with the estimated probabilities for the Usual Care as a baseline.

Changes to reference parameters were additive across patient characteristics and interventions. For example, a female patient aged 35–64 years who used the New Intervention

Table 4.2: Reference outcomes, per patient per 12 cycles/months – Mean (Standard deviation).[a]

Variables	Usual Care	New Intervention	Difference
QALYs	0.485 (0.232)	0.540 (0.231)	0.054
LYs	0.740 (0.328)	0.786 (0.313)	0.046
Intervention costs	€ 533 (€ 236.24)	€ 3,300 (€ 1,310)	€ 2,770
Maintenance costs	€ 3,260 (€ 2,080)	€ 3,070 (€ 1,810)	- € 180
Event costs	€ 2,330 (€ 2,610)	€ 1,260 (€ 1,780)	- € 1,070
Total costs	€ 6,120 (€ 4,340)	€ 7,630 (€ 3,830)	€ 1,520
Number of cycles in:			
Moderate disease	5.171 (3.750)	6.209 (3.965)	1.038
Severe disease	2.477 (2.512)	2.313 (2.507)	-0.164
Very severe disease	1.238 (1.850)	0.911 (1.554)	-0.327
Death	3.114 (3.937)	2.567 (3.751)	-0.547
Number of events	1.160 (1.259)	0.630 (0.856)	-0.530
Proportion surviving	49.9%	58.3%	8.4%pt
ICER, total costs per QALY			€ 28,020

[a] LY: Life year; QALY: Quality adjusted LY; ICER = Incremental cost-effectiveness ratio; %pt: percentage points

had a monthly probability to die in the very severe disease stage of 10 % (the probability within the RDP) – 2 % (modification for gender) + 4 % (modification for age) - 3 % (modification for New Intervention) = 9%.

Table 4.2 shows the reference outcomes when applying the two interventions to the entire patient population. They represent the 'truth' with which the outcomes of the meta-analyses were compared.

The structure of the HE model mirrors the disease progression in the reference population; in other words, there was no structural uncertainty. The time horizon of the HE model was 1 year and the cycle length 1 month. We assumed that data in the trials were collected each month during 1 year. We have not applied discounting. Simulation and modelling were performed using SAS 9.2 and WinBUGS 1.4.3.

4.2.3. Scenarios

The number and size of the trials sampled in step 2: Trial selection was varied in scenarios, as well as the amount of heterogeneity between trials. Heterogeneity in the meta-analysis literature is any kind of variability between different studies.[10] Trial heterogeneity is different from patient heterogeneity, which is the difference between patients that can be adequately explained by patient characteristics. Table 4.3 shows the different scenarios that were investigated. The last column of table 4.3 described the impact of the non-randomly drawn trials on the trial parameters. We will focus mainly on the three scenarios

Table 4.3: Overview of different scenarios in the simulation study.[a]

Scenario	Number of trials	Number of patients per treatment arm	Added heterogeneity with effect on disease progression
1	5	All trials 500	-
2	5	Trial 1 en 2: 500, trial 3: 100, trial 4: 250, trial 5: 1,000	-
3	10	All trials 250	-
4	5	All trials 500	Trial 5 has relatively old patients, more smokers and more obese patients, which leads to more rapid disease deterioration, higher probability of events, higher maintenance costs, lower quality of life.
5	5	All trials 500	Trial 2 has relatively young patients, which leads to slower disease deterioration Trial 4 has only patients from developing countries, which leads to lower maintenance costs Trial 5: the same as in scenario 4
6	5	Trials have different sample sizes, the same as in scenario 2	The same as in scenario 5
7	5	Trials have different sample sizes, the same as in scenario 2	Trials 2, 4 and 5 has relatively old patients, more smokers and more obese patients, which leads to more rapid disease deterioration, higher probability of events, higher maintenance costs, lower quality of life.

[a] Scenarios in shaded rows are discussed in main text. The other four scenarios are discussed in the discussion.

in shaded rows, namely 1, 4 and 7. The other scenarios will be discussed in the discussion section of this paper.

4.2.4. Methods of meta-analysis

In our study, we compared four widely used methods of meta-analysis: frequentist fixed effects (FFE), frequentist random effects (FRE), Bayesian fixed effects (BFE) and Bayesian random effects (BRE). The FFE and FRE were based on the Inverse Variance method, which can be used for meta-analysis of both continuous and dichotomous data.[11] The pooled effect estimate for the FFE is calculated as a weighted average of the individual study estimates, using the inverse of the squared standard error (s.e.) of the effect estimates as weights. Thus, studies with a smaller s.e., typically larger studies, are given more weight than studies with a larger s.e.. For the FRE, we used the DerSimonian-Laird method.[11] It was developed for situations where there is heterogeneity between study results, caused for example by differences in patient population or study design. It incorporates an estimate of the between-study heterogeneity into the weights. It is assumed that all studies are samples drawn from a pool of all possible studies, i.e. the population.[10] The goal is to

estimate the mean of this population. The true parameter value may be study-specific and can vary across studies.

Both the FFE and FRE assume that the weights are known. With little or no heterogeneity among the studies, the FFE and FRE will give identical results.[10] With heterogeneity present, confidence intervals will be wider for the FRE and claims of statistical significance will be more conservative. The point estimate of the parameter might also be different. We report the I^2-statistic as a measure of heterogeneity[12], which can be interpreted as the proportion of the total variation in the pooled estimates that is due to heterogeneity between studies. When the amount of between-trial heterogeneity increases compared with the within-trial variance, then the I^2 also increases. Higgins et al. provide a rough guide to the interpretation of I^2.[8] Above 30% "may represent moderate heterogeneity"; above 50% "may represent substantial heterogeneity".

The BFE method requires the data from the different trials, the definition of a prior for the parameter to be synthesized and a likelihood linking both.[9,13] We used a binomial likelihood function to model the total number of transitions, with a flat beta prior; and a normal likelihood function for all other parameters, with a flat normal prior centered on 0 and a precision of 1.0E-6. When specifying the BRE method, prior distributions need also be defined for the between-trial heterogeneity.[9,13,14] We used the inverse of a squared uniform distribution from 0 to 10. Other likelihoods and priors were as in the BFE. Before simulation started, we tested several priors and could find no meaningful differences.

Conceptually, confidence intervals in frequentist statistics and credibility intervals in Bayesian statistics have very different interpretations (see for example[15,16]). However, for convenience and legibility, we abbreviate both as CI. For each pooled parameter estimate, we report the mean and the 95% CI.

We performed meta-analysis on all baseline values (transition probabilities, utilities, etc) using data on the New Intervention. In addition, we performed meta-analysis on all effect measures (RR), using data on the difference between the New Intervention and Usual Care. Interested readers may request code on both the simulation study and the methods of meta-analysis from the corresponding author.

4.2.5. Comparing performance

When judging the performance of the methods of meta-analysis, we assumed that a researcher doing a meta-analysis aims to estimate the CE of the New Intervention compared with Usual Care in the entire patient population, not a specific subgroup. We further assumed that a researcher is unaware of the fact that heterogeneity, when present, was caused by sampling from subgroups (i.e. they do not know we deliberately built in heterogeneity). To the researcher, the heterogeneity might either be caused by random sampling or unobserved differences between the trials in terms of patient characteristics, setting or other elements that could affect the parameter estimates. These assumptions are made

because, if these differences in design are known to the researcher, either the trials would not be synthesized at all, or a way has to be found to control for these differences. Hence, these assumptions made it possible to judge the performance of the different methods of meta-analysis by comparing model parameters and outcomes with the reference values.

The statistical performance of the different methods was judged by calculating the coverage, bias, mean absolute deviation (MAD) and statistical power. Coverage is the percentage of all repetitions that the simulated CI covered the 'truth'. Since the coverage is based on 95% CIs, we expect that, if all trials are drawn randomly, the coverage should on average be close to 95%.[5] The observed coverage was compared to this benchmark. Assuming that we have an unbiased point estimate, if the coverage is below 95%, the model does not take into account all uncertainty. If the coverage is above 95%, it has accounted for too much uncertainty. In this study, if the coverage was smaller than 90%, we say the method underestimated uncertainty; if the coverage was higher than 98% the method overestimated uncertainty. Bias is expressed as the difference between the point estimate in the simulated data set and the true population value, averaged over all repetitions. The MAD is the average, over all repetitions, of the absolute value of the bias. The MAD indicates how far the estimated value was from the 'truth', regardless of whether it was too high or too low. For HE outcomes, we also calculated statistical power, expressed as the percentage of all repetitions where the simulated result yields a statistically significant difference between treatments.

4.3 RESULTS

4.3.1. Model parameters for one set of trials

Figure 4.3 compares the methods for scenarios 1, 4 and 7, using only the first repetition. From bottom to top, we compare the different meta-analysis models for the seven scenarios. Each dot represents the point estimate for the model parameter, in this case the transition probability from severe to very severe disease, and the bars the estimated CIs. At the bottom of the graph the 'true' reference parameter value, as found in the population, is pictured, with which each of the estimates needs to be compared. The results are illustrative for the other parameters. When five equally sized, large trials are randomly drawn from the same population (scenario 1), all methods lead to similar point estimates of the model parameters, but the BRE model has a much wider CI and a higher coverage. The difference in point-estimate between FFE and BFE is due to the different distributional assumptions: BFE assumes a binomial model, whereas FFE (implicitly) assumes a normal distribution.

In scenario 4 we added heterogeneity by drawing one of the trials from a less healthy population. The point estimates from the random effects (RE) models are further from the

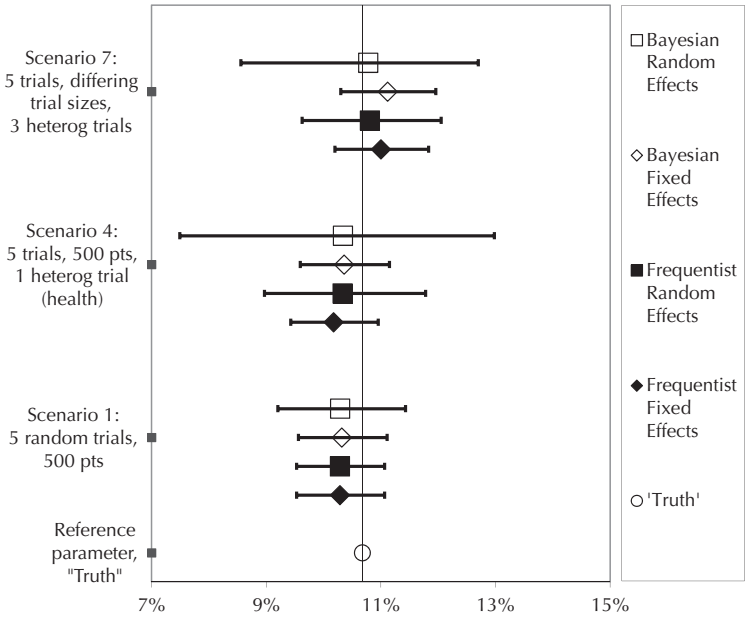

Figure 4.3: Meta-analysis on the transition from the severe to very severe disease stage, for three of the seven scenarios, for the Usual Care arm, for one repetition.[a]

[a] pts = number of patients per trial, equal in each arm. heterog = added heterogeneity by sampling from subpopulations

reference parameter. RE models assign a relatively greater weight to trials which outcomes differ from the rest. Due to the wider CIs, RE models are more likely to include the reference parameter value, but tend to overestimate uncertainty.

Varying trial sizes, with three trials from the same subgroup (scenario 7) leads to results comparable to scenario 4, where only one of the trials was drawn from this subgroup.

4.3.2. Model parameters for 1,000 repetitions

To investigate whether the results from the previous paragraph are due to chance, or if there are systematic differences, table 4.4 presents a summary of the performance indicators over 1,000 repetitions. It reports the number of model parameters out of 33 for which the performance indicators are below or above certain threshold values. First we look at the I^2, averaged over 1,000 repetitions. For many parameters and scenarios, the mean of the I^2 statistic does not exceed 30%, indicating no heterogeneity, even in scenarios where heterogeneity is built in. Some parameters show substantial heterogeneity, even if all trials are randomly drawn from the same population. The number of parameters with a mean I^2 below 30% decreases when the amount of heterogeneity increases and the number of parameters with a mean I^2 above 50% increases slightly.

Table 4.4: Summary of the result of meta-analysis on all parameters of the health-economic model. Means over 1,000 repetitions.

Total number of parameters for which:	Scenario 1	Scenario 4	Scenario 7
Total number of parameters	33	33	33
Parameters influenced by added heterogeneity	0	24	24
Mean I^2 < 30%: heterogeneity might not be important	27	27	22
Mean I^2 > 50%: substantial heterogeneity	4	6	6
Mean coverage < 90% (underestimation of uncertainty)			
Frequentist fixed effects method (FFE)	6	9	23
Frequentist random effects method (FRE)	6	0	21
Bayesian fixed effects method (BFE)	6	9	23
Bayesian random effects method (BRE)	0	0	0
Mean coverage > 98% (overestimation of uncertainty)			
FFE	11	7	3
FRE	12	13	4
BFE	12	10	4
BRE	26	32	24
Mean bias > 2%			
FFE	0	12	19
FRE	0	13	19
BFE	0	12	20
BRE	0	13	21
Mean MAD[a] > 5%			
FFE	0	3	16
FRE	0	5	16
BFE	3	6	17
BRE	5	9	17

[a] MAD: Mean Absolute Deviation

When equally sized trials are randomly drawn from the same underlying population (scenario 1), the number of parameters with mean coverage below 90% or above 98% is comparable for FFE, FRE and BFE. BRE, on the other hand, shows no underestimation of uncertainty in any of the parameters, and an overestimation in 26 of the 33 parameters. FFE and BFE have a tendency to underestimate uncertainty when heterogeneity is added (scenarios 4 and 7), as is illustrated by the increasing number of parameters with a coverage lower than 90%. It should be noted that an increase in bias and MAD also contributes to a lower coverage. In scenario 7, even the FRE model underestimates uncertainty for several parameters and the number of parameters where the uncertainty is overestimated decreases. BRE never underestimates uncertainty, and overestimates uncertainty for nearly

all of the parameters in all scenarios. In fact, the coverage is 100% in a large number of cases (not shown).

There are only small differences between methods in bias, with more bias in the scenarios with more added heterogeneity. There are, however, differences between the methods with respect to the MAD. The number of parameters where the MAD is larger than 5% is smaller for the FFE and FRE, than for the BFE and BRE methods, regardless of heterogeneity. The BRE method generally yields point-estimates that are further away from the true population value than the other methods. Using RE models in scenarios with heterogeneity does not necessarily reduce bias. They may even increase bias, especially when the trials that differ from the others all differ in the same direction (scenario 7).

Table 4.5: Health economic outcomes for three of the seven scenarios. Both intervention arms and the difference. Means and range from the 2.5th and 97.5th percentiles over 1,000 repetitions.[a]

Scenario	Scenario 1			Scenario 4			Scenario 7		
	Five randomly drawn, equally sized trials			Five equally sized trials; one trial drawn from a less health population			Five equally sized trials; three trials drawn from a less health population		
Intervention arm	New Int	Usual	Diff	New Int	Usual	Diff	New Int	Usual	Diff
Number of QALYs									
Truth	0.540	0.485	0.054	0.540	0.485	0.054	0.540	0.485	0.054
FFE	0.542	0.488	0.054	0.533	0.480	0.053	0.515	0.464	0.051
FRE	0.541	0.487	0.054	0.532	0.479	0.053	0.514	0.463	0.051
BFE	0.541	0.486	0.054	0.531	0.478	0.053	0.513	0.461	0.052
BRE	0.540	0.487	0.054	0.531	0.478	0.052	0.513	0.462	0.051
Number of LYs									
Truth	0.786	0.740	0.046	0.786	0.740	0.046	0.786	0.740	0.046
FFE	0.789	0.744	0.045	0.781	0.738	0.044	0.767	0.723	0.043
FRE	0.788	0.744	0.045	0.780	0.736	0.044	0.766	0.723	0.044
BFE	0.787	0.742	0.045	0.779	0.735	0.044	0.764	0.720	0.044
BRE	0.787	0.743	0.045	0.779	0.735	0.043	0.764	0.721	0.042
Total costs									
Truth	€ 7,633	€ 6,116	€ 1,517	€ 7,633	€ 6,116	€ 1,517	€ 7,633	€ 6,116	€ 1,517
FFE	€ 7,657	€ 6,140	€ 1,517	€ 7,652	€ 6,158	€ 1,494	€ 7,643	€ 6,167	€ 1,476
FRE	€ 7,653	€ 6,137	€ 1,515	€ 7,644	€ 6,152	€ 1,492	€ 7,639	€ 6,164	€ 1,475
BFE	€ 7,639	€ 6,126	€ 1,513	€ 7,627	€ 6,136	€ 1,490	€ 7,615	€ 6,139	€ 1,476
BRE	€ 7,650	€ 6,129	€ 1,522	€ 7,635	€ 6,145	€ 1,490	€ 7,627	€ 6,157	€ 1,470

[a] FFE: Frequentist fixed effects method; FRE: Frequentist random effects method; BFE: Bayesian fixed effects method; BRE: Bayesian random effects method; New Int: New Intervention; Usual: Usual Care; Diff: Difference between two intervention arms

4.3.3. Health-economic outcomes for 1,000 repetitions

Differences in model parameters may also lead to differences in HE outcomes. In table 4.5, we show the mean HE outcomes over 1,000 repetitions, for both interventions and the difference between them. In scenario 1, all HE outcomes are very close to the true population value. In scenario 7, we can see that the point-estimates are further from the truth than is the case in the other two scenarios, for all methods of meta-analysis. On average, the number of QALYs estimated in each of the treatment arms is around 5% below the true population value, and so is the difference in QALYs. In scenario 7, the fixed effects (FE) CIs (not shown) are comparable to those in scenario 1, but the RE CIs are much wider, especially for the BRE method.

Table 4.6 shows the coverage, bias and MAD, for the difference between the two intervention groups. In general we see that the coverage is larger in the RE methods, due to wider CIs which take heterogeneity into account. In addition, the Bayesian methods have higher coverage than the frequentist methods. Bias is generally low when no heterogeneity is included (scenario 1) and increases when it is (scenarios 4 and 7). The largest bias and MAD is found in the BRE method, for all outcomes in all scenarios. For the other three methods, bias and MAD are comparable. Despite the higher bias and MAD, the coverage of the BRE is still larger.

For the number of QALYs, events and total costs, statistical power (appendix A4) is 100% for all scenarios of FFE, FRE and BFE. It is slightly lower for the LYs for FFE, FRE and BFE, with a minimum of 96.7%. For the BRE method, the statistical power for LYs ranges from 17.5% (scenario 6) to 100% (scenario 3). It is generally lower when there are more trials drawn from a subgroup of patients and when there is a difference in sample size between the trials.

Table 4.6: Health economic outcomes for three of the seven scenarios. Means of coverage, bias and mean absolute deviance (MAD) of the difference between two interventions, over 1,000 repetitions.[a]

Scenario	Scenario 1			Scenario 4			Scenario 7		
	Five randomly drawn, equally sized trials			Five equally sized trials; one trial drawn from a less health population			Five equally sized trials; three trials drawn from a less health population		
	Coverage	Bias	MAD	Coverage	Bias	MAD	Coverage	Bias	MAD
Number of QALYs									
FFE	98.1%	-0.2%	8.3%	97.4%	-2.4%	8.5%	96.0%	-5.1%	9.5%
FRE	98.8%	-0.1%	8.3%	99.5%	-2.2%	8.5%	99.3%	-5.0%	9.7%
BFE	98.3%	0.2%	8.7%	98.6%	-2.1%	8.8%	97.9%	-5.0%	9.4%
BRE	100.0%	-1.0%	9.6%	100.0%	-3.3%	10.7%	100.0%	-6.6%	13.4%
Number of LYs									
FFE	98.2%	-1.6%	13.9%	97.1%	-3.8%	14.2%	97.5%	-4.6%	15.0%
FRE	99.3%	-1.4%	14.0%	99.1%	-3.4%	14.3%	98.9%	-4.4%	15.5%
BFE	98.6%	-0.9%	14.5%	98.4%	-3.1%	14.8%	99.2%	-4.1%	15.0%
BRE	100.0%	-2.3%	16.1%	100.0%	-4.9%	17.5%	100.0%	-6.9%	20.8%
Total costs									
FFE	98.5%	0.0%	5.1%	98.2%	-1.5%	5.5%	97.1%	-2.7%	5.9%
FRE	99.3%	-0.1%	5.2%	99.5%	-1.7%	5.6%	99.4%	-2.8%	6.1%
BFE	98.5%	-0.3%	5.3%	98.5%	-1.8%	5.7%	98.4%	-3.2%	6.1%
BRE	100.0%	0.3%	6.4%	100.0%	-1.8%	6.8%	100.0%	-3.1%	8.1%

[a] FFE: Frequentist fixed effects method; FRE: Frequentist random effects method; BFE: Bayesian fixed effects method; BRE: Bayesian random effects method; MAD: Mean Absolute Deviation

Figure 4.4 shows the CE acceptability curves (CEAC) for scenario 7. The four graphs represent the four methods of meta-analysis. In each graph, we show the CEAC for ten repetitions, the median, 2.5th and 97.5th percentile over 1,000 repetitions. It is clear that even in this scenario with a lot of heterogeneity, the graphs are very similar for FFE, FRE and BFE. At a ceiling ratio of € 30,000 per QALY, which is very close to the true population ICER of € 28,020 (dashed vertical line), the median probability of New Intervention being cost effective is between 60–70%, for these three methods. At a ceiling ratio of € 21,000, the median probability for all three methods is below 20% and the 97.5th percentile is below 30%. At € 39,000, the median probability is above 95% and the 2.5th percentile is above 65%, again for all three methods. Therefore, no great difference in policy decision would arise from using these three different methods of meta-analysis.

However, using BRE, the outcome would be different. Even at a ceiling ratio of € 48,000, the 97.5th percentile is below 60%, and the median probability is below 90%. Using BRE, a policy maker would be much less certain of the cost-effectiveness of the new intervention.

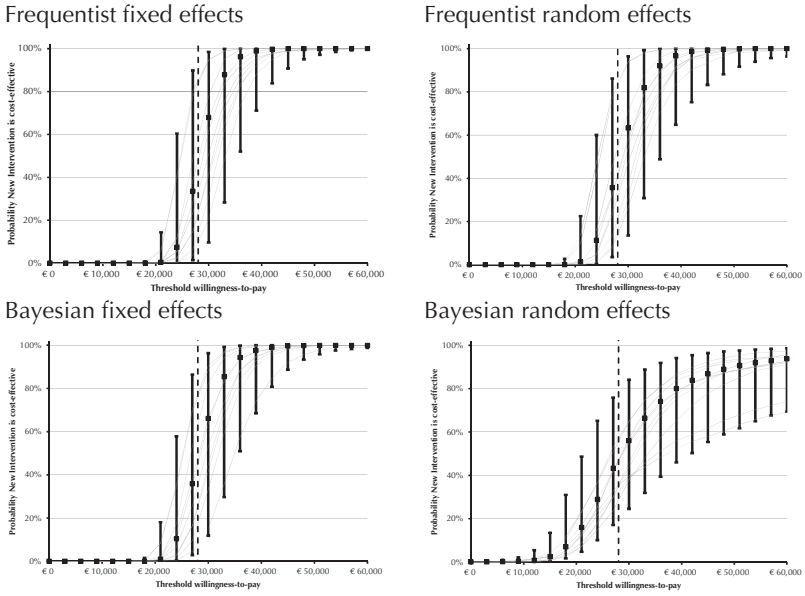

Figure 4.4: CEACs for the four models in the heterogeneous scenario 7. Graphs depicts median, 2.5th and 97.5th percentile CEACs over 1,000 repetitions, as well as the CEACs for the first 10 repetitions; horizontal line is the 'true' population ICER.[a]

[a] CEAC: Cost-effectiveness acceptability curve; ICER: incremental cost-effectiveness ratio

4.4 DISCUSSION

In this study, we compared four methods of meta-analysis. Using a simulation study we could compare the HE outcomes to a gold standard and judge their statistical performance. In order to do this, we made a few crucial assumptions. First, we assumed that the researcher wants to estimate the parameter values of the entire population, not a subpopulation. This allows us to compare the results to the outcomes for the entire population. We also assumed the researcher was unaware of the fact that any heterogeneity was caused by sampling from subgroups. A researcher might not have combined the trials at all, had they been aware of the differences, by seeing the patient characteristics or trial protocols. A researcher might also have tried to compensate using regression methods, which were not the focus of the paper, nor would they be feasible with only five to ten trials.

With almost no heterogeneity, we found that the results of the FFE, FRE and BFE methods were comparable. With heterogeneity added to the trials, we saw differences on a parameter level, but these did not translate into important differences in HE outcomes. That could be because the HE model combines all parameter estimates and their uncertainties into one estimate of QALYs and total costs. All these uncertainties together may

hide the (subtle) differences we have seen between the methods. In addition, we did not take structural uncertainty into account, which may exceed any parameter uncertainty.

Using any of these three methods would not lead to differences in policy decisions. Using BRE would, as it has a tendency to overestimate uncertainty and yield a larger probability that a new treatment is rejected or that more research is asked for where it might not be necessary. Partly, this is due to the number of trials included in the meta-analysis. Generally speaking, sophisticated methods, such as RE, require more data than simple methods, because of the increased number of parameters. This is particularly important for BRE, as it estimates between-study heterogeneity and also takes the uncertainty around this estimate into account. This can be estimated more precisely from ten trials than from five. In scenario 3, where we have the same amount of patients in ten trials instead of five, we have seen that the CI around the BRE is still larger than those of the other three methods, but the difference is much smaller. We also saw that the coverage for the BRE is much closer to 95% and that uncertainty is overestimated in fewer parameters.

We speculate that with more than ten trials the differences might be even less pronounced and the BRE method will yield almost the same results as the other three methods, although the amount of uncertainty will always exceed that of the other methods. We did not test this assumption as this situation is unlikely within the scope of the expensive drug programme. In addition, time and budget constraints did not permit the calculation time needed for a simulation of this many trials, especially in a number of different scenarios.

Based on this, we recommend not using the BRE when only few sources of evidence are available. Unfortunately, this is more the rule than the exception, especially in the expensive drugs programme which was the reason to initiate this study. With only a few differences between the other three methods, we would personally favor FRE, as it automatically reduces to FFE in the absence of heterogeneity, is easy to implement and is more easily understood by physicians and policy makers who will be using the results.

By calculating outcomes for a number of scenarios, we have covered many of the different situations that are likely to arise in meta-analysis. We have drawn a few larger trials, but more smaller trials, and trials with differences in trial sizes. We have drawn trials randomly from the same population, one trial from a subgroup of patients, several trials from different subgroups and several trials from the same subgroup. Because of this, we feel that the results of our study are generalizable to other studies that use meta-analysis to obtain pooled estimates of parameters to fill a HE model.

We have made sure that the difference between the two interventions is large. When two interventions are much closer to each other, it unlikely this will change our conclusions regarding the methods of meta-analysis. The same is true for a longer time horizon, or including discounting.

Despite our feelings that the results are generalizable to other situations, there are several limitations to our study. The first limitation is that we have assumed that all data comes

from the same set of trials. In practice, the data for transition probabilities will likely come from different sources than, for example, the RR for those transition probabilities or the utilities. The exact source of the evidence will not have an impact on the performance of the methods of meta-analysis. Therefore, we decided not to explore this extra complexity in this paper.

Another limitation is the choice of prior for the Bayesian models. The use and choice of priors is an important subject when discussing the Bayesian methodology. Any Bayesian calculation can be affected by the type of priors used. In the case of meta-analysis, a small number of studies is extra vulnerable to the type of prior.[8, 17] As we did not assume the researcher to have prior information, we also used so called vague, or flat priors. Even though they are supposed to be 'uninformative', they may influence the outcomes, especially the posterior scale parameters.[17] We tested several different specifications of the priors but did not find any differences in outcomes, likely from the relative simplicity of the models used. However, researchers using the BFE or BRE should keep these restrictions in mind and different priors may lead to different results.

Our results are not generalizable to network meta-analysis and should only be used in the case of a pair-wise comparison of two interventions. In the case that more than two comparators are available, other methods of meta-analysis are available, which make use of all the available evidence (see [18–21] and chapter 5).

We have seen that both the RE methods and the appropriate measure for heterogeneity, I^2, have a tendency to detect heterogeneity, when trials have differences in number of patients, even with a large number of total patients, randomly drawn from the same underlying population. This is a very common occurrence in meta-analysis and may lead to too conservative CIs as none of the methods can make the distinction between sampling error and heterogeneity. Trials can therefore be considered heterogeneous, not only when one or more trials are drawn from a (different) subgroup of patients, but also when all trials are randomly drawn from the same population, but with differences in trial sizes. At the same time, with heterogeneity built in, many of the parameters show no important degree of heterogeneity. From this we can see that the I^2 might be an imperfect measure for heterogeneity, at least with a relatively low number of trials.

In our simulation study, we have made sure that the reference parameters are not close to their natural limits; for example, probabilities or costs close to 0. In cases when the reference parameters are closer to these limits, we expect that the Bayesian methods will have model parameters that are closer to the true population value than the frequentist methods. First of all, frequentist methods usually need a correction term (continuity correction) if one of the trial parameters is 0, because it will not be possible to calculate the necessary standard errors otherwise. Bayesian methods do not. In addition, Bayesian methods may use a bounded likelihood function, while frequentist methods always im-

plicitly use a normal distribution. This might be a valid reason to prefer Bayesian methods over frequentist methods.

The transition probabilities and probabilities to experience an event in the New Intervention arm were calculated using the model parameter in the Usual Care arm, and the corresponding RR. Results using the risk difference were similar and therefore not shown.

In many HE models, many input parameters need to be estimated. When more than one input parameter is estimated from the same set of sources, we recommend heterogeneity is not checked for each parameter separately, but rather for the set of trials. If statistics indicate trials are homogeneous for one parameter, but heterogeneous for another, it is recommended that all parameters are calculated using the same type of model. The model type selection should be based on trial heterogeneity rather than parameter heterogeneity.

4.5 CONCLUSION

In conclusion, the FFE, FRE and BFE meta-analysis methods led to comparable HE outcomes, even in scenarios where we built in heterogeneity. The differences that we see between the methods point towards a broader CI (which is translated in a higher coverage), a higher MAD and a lower statistical power for Bayesian methods compared with frequentist methods, and for RE methods compared with FE methods. RE methods do not necessarily reduce bias when heterogeneity is added to the trials, and may even increase bias in certain situations. BRE tends to overestimate uncertainty reflected in the shape of the CEAC. Based on this study, we recommend the FRE method as the preferred method of meta-analysis.

4.6 LITERATURE

[1] NZa. Beleidsregel Dure Geneesmiddelen [Policy rule Expensive Drugs] (BR-CU-2017). 2011. http://www.nza.nl/regelgeving/beleidsregels/ziekenhuiszorg/BR-CU-2017. Accessed 15 June 2011.

[2] Cooper NJ, Sutton AJ, Abrams KR, Turner D, Wailoo A. Comprehensive decision analytical modelling in economic evaluation: a Bayesian approach. Health Econ. 2004;13(3):203–26.

[3] Ades AE, Sculpher M, Sutton A, Abrams K, Cooper N, Welton N, et al. Bayesian methods for evidence synthesis in cost-effectiveness analysis. Pharmacoeconomics. 2006;24(1):1–19.

[4] Riley RD, Simmonds MC, Look MP. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. J Clin Epidemiol. 2007;60(5):431–9.

[5] Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med. 2001;20(6):825–40.

[6] Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. 2001;10(4): 277–303.

[7] Sutton AJ, Higgins JP. Recent developments in meta-analysis. Stat Med. 2008;27(5):625–50.

[8] Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions 5.0.2, updated September 2009. Available at http://www.cochrane-handbook.org/.

[9] Oppe M, Al M, Rutten-van Molken M. Comparing methods of data synthesis: re-estimating parameters of an existing probabilistic cost-effectiveness model. Pharmacoeconomics. 2011; 29(3): 239–50.

[10] Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557–60.

[11] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177–88.

[12] Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002; 21(11):1539–58.

[13] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. London: Chapman & Hall; 1995.

[14] Carlin BP, Louis TA. Bayes and empirical Bayes methods for data analysis. London: Chapman & Hall; 1996.

[15] Jaynes E. Confidence intervals vs Bayesian intervals. In: Harper W, Hooker CA, editors. Foundations of probability theory, statistical inference, and statistical theories of science. Dordrecht: D Reidel; 1976. p. 175.

[16] O'Hagan A, Luce B. A primer on Bayesian statistics in health economics and outcomes research. Sheffield: Centre for Bayesian Statistics in Health Economics; 2003.

[17] Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Stat Med. 2005;24(15):2401–28.

[18] Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ. 2003;326(7387):472.

[19] Strassmann R, Bausch B, Spaar A, Kleijnen J, Braendli O, Puhan MA. Smoking cessation interventions in COPD: a network meta-analysis of randomised trials. Eur Respir J. 2009;34(3):634–40.

[20] Puhan MA, Bachmann LM, Kleijnen J, Ter Riet G, Kessels AG. Inhaled drugs to reduce exacerbations in patients with chronic obstructive pulmonary disease: a network meta-analysis. BMC Med. 2009;14(7):2.

[21] Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc. 2006;101:447–59.

A4 APPENDIX

Table A4.1: Monthly transition probabilities in the Reference Disease Progression and changes in probabilities due to patient characteristics or interventions.

From:/To:	Moderate	Severe	Very severe	Death
Reference disease progression (RDP)				
1 – Moderate	80%	10%	6%	4%
2 – Severe	20%	65%	10%	5%
3 – Very Severe	10%	20%	60%	10%
4 – Death	0%	0%	0%	100%
Changes due to gender. RDP = male, changes applicable to female patients.				
1 – Moderate	+1%	+1%	-1%	-1%
2 – Severe	+1%	+1%	-1%	-1%
3 – Very Severe		+1%	+1%	-2%
Changes due to age class. RDP = 18-34, changes applicable to patients aged 35-64, double these changes applicable to patients aged 65+.				
1 – Moderate	-4%	+2%	+0%	+2%
2 – Severe		-4%	+2%	+2%
3 – Very Severe			-4%	+4%
No changes due to developed/developing country.				
Changes due to BMI. RDP = low/average BMI, changes applicable to patients with high BMI, double these changes applicable to patients with double BMI.				
1 – Moderate	-1%	+1%	+0%	+0%
2 – Severe		-1%	+1%	+0%
3 – Very Severe			-1%	+1%
Changes due to smoking status. RDP = non-smokers, changes applicable smokers.				
1 – Moderate	-3%	+1%	+1%	+1%
2 – Severe		-3%	+1%	+2%
3 – Very Severe			-3%	+3%
Changes due to use of Usual Care.				
1 – Moderate	+5%	-2%	-2%	-1%
2 – Severe		+2%	-1%	-1%
3 – Very Severe			+2%	-2%
Changes due to use of New Intervention.				
1 – Moderate	+10%	-5%	-3%	-2%
2 – Severe	+3%	+3%	-4%	-2%
3 – Very Severe		+5%	-2%	-3%

Table A4.2: Probability of an event per cycle for each disease stage in the Reference Disease Progression and changes in probabilities due to patient characteristics or interventions.

Disease stage	Moderate	Severe	Very severe	Death
Reference disease progression (RDP)	5%	10%	40%	0%
No changes due to gender.				
No changes due to age class.				
No changes due to developed/developing country.				
Changes due to BMI. RDP = low/average BMI, changes applicable to patients with high BMI, double these changes applicable to patients with double BMI.	+1%	+2%	+4%	
Changes due to smoking status. RDP = non-smokers, changes applicable smokers.	+2%	+5%	+10%	
No changes due to use of Usual Care.				
Changes due to use of New Intervention.	-2%	-10%	-20%	

Table A4.3: Monthly costs per stage using a Gamma distribution in the Reference Disease Progression and changes in distributional parameters due to patient characteristics or interventions.

Disease stage	Moderate	Severe	Very severe	Death
Reference disease progression (RDP)				
Alpha	4	5	10	
Beta	50	80	100	
No changes due to gender.				
Changes due to age class. RDP = 18-34, changes applicable to patients aged 35-64, double these changes applicable to patients aged 65+.				
Alpha				
Beta	+5	+5	+5	
Changes due to developed/developing country. RDP = developed country, changes applicable to patients from developing country.				
Alpha				
Beta	-10	-10	-10	
Changes due to BMI. RDP = low/average BMI, changes applicable to patients with high BMI, double these changes applicable to patients with double BMI.				
Alpha				
Beta	+2	+2	+2	
No changes due to smoking status.				
No changes due to use of Usual Care.				
No changes due to use of New Intervention.				

Table A4.4: Quality of life weights using a Beta distribution in the Reference Disease Progression and changes in distributional parameters due to patient characteristics or interventions.

Disease stage	Moderate	Severe	Very severe	Death
Reference disease progression (RDP)				
Alpha	64	35	20	
Beta	16	15	20	
No changes due to gender.				
Changes due to age class. RDP = 18-34, changes applicable to patients aged 35-64, double these changes applicable to patients aged 65+.				
Alpha	+5	+5	+5	
Beta	+5	+5	+5	
No changes due to developed/developing country.				
Changes due to BMI. RDP = low/average BMI, changes applicable to patients with high BMI, double these changes applicable to patients with double BMI.				
Alpha				
Beta	+5	+5	+5	
No changes due to smoking status.				
No changes due to use of Usual Care.				
No changes due to use of New Intervention.				

Table A4.5: Costs due to an event using a Gamma distribution in the Reference Disease Progression. No changes due to patient characteristics or interventions.

Disease stage	Moderate	Severe	Very severe	Death
Reference disease progression (RDP)				
Alpha	10	10	10	0
Beta	200	200	200	0

Table A4.6: Quality of life decrement due to an event using a Beta distribution in the Reference Disease Progression No changes due to patient characteristics or interventions.

Disease stage	1	2	3	4
Reference disease progression (RDP)				
Alpha	6	6	6	0
Beta	4	4	4	0

Table A4.7: Summary of the result of meta-analysis on all parameters of the health-economic model for four of the seven scenarios. Means over 1,000 repetitions.

Total number of parameters for which:	Scenario 2	Scenario 3	Scenario 4	Scenario 6
Total number of parameters	33	33	33	33
Parameters influenced by added heterogeneity	0	0	24	24
Mean I^2 < 30%: heterogeneity might not be important	27	27	27	22
Mean I^2 > 50%: substantial heterogeneity	4	6	6	6
Mean coverage < 90% (underestimation of uncertainty)				
Frequentist fixed effects method (FFE)	6	6	9	17
Frequentist random effects method (FRE)	4	2	0	3
Bayesian fixed effects method (BFE)	6	6	9	17
Bayesian random effects method (BRE)	0	0	0	0
Mean coverage > 98% (overestimation of uncertainty)				
FFE	13	12	7	4
FRE	13	13	13	10
BFE	12	13	10	4
BRE	31	14	32	32
Mean bias > 2%				
FFE	0	3	12	17
FRE	0	0	13	16
BFE	0	0	12	18
BRE	3	0	13	17
Mean MAD[a] > 5%				
FFE	2	1	3	11
FRE	3	0	5	9
BFE	3	3	6	13
BRE	9	3	9	12

[a] MAD: Mean absolute deviation

Table A4.8: Health economic outcomes for two of the seven scenarios. Both intervention arms and the difference. Means and range from the 2.5th and 97.5th percentiles over 1,000 repetitions.[a]

Scenario		Scenario 2			Scenario 3	
Intervention arm	New Int	Usual	Diff	New Int	Usual	Diff
Number of QALYs						
Truth	0.540	0.485	0.054	0.540	0.485	0.054
FFE	0.542	0.488	0.054	0.543	0.489	0.053
FRE	0.541	0.488	0.054	0.542	0.489	0.053
BFE	0.541	0.486	0.054	0.540	0.487	0.054
BRE	0.540	0.487	0.053	0.540	0.486	0.054
Number of LYs						
Truth	0.786	0.740	0.046	0.786	0.740	0.046
FFE	0.789	0.745	0.044	0.790	0.747	0.044
FRE	0.788	0.744	0.044	0.790	0.746	0.044
BFE	0.787	0.742	0.045	0.787	0.743	0.044
BRE	0.786	0.743	0.043	0.787	0.742	0.045
Total costs						
Truth	€ 7,633	€ 6,116	€ 1,517	€ 7,633	€ 6,116	€ 1,517
FFE	€ 7,652	€ 6,142	€ 1,510	€ 7,667	€ 6,151	€ 1,516
FRE	€ 7,649	€ 6,141	€ 1,508	€ 7,661	€ 6,146	€ 1,515
BFE	€ 7,636	€ 6,125	€ 1,511	€ 7,629	€ 6,117	€ 1,512
BRE	€ 7,641	€ 6,129	€ 1,512	€ 7,629	€ 6,114	€ 1,515

[a] LY: Life year; QALY: Quality Adjusted LY; FFE: Frequentist fixed effects method; FRE: Frequentist random effects method; BFE: Bayesian fixed effects method; BRE: Bayesian random effects method, New Int = New Intervention, Usual = Usual Care, Diff = Difference between two intervention arms

Table A4.9: Health economic outcomes for two of the seven scenarios. Both intervention arms and the difference. Means and range from the 2.5th and 97.5th percentiles over 1,000 repetitions.[a]

Scenario		Scenario 5			Scenario 6	
Intervention arm	New Int	Usual	Diff	New Int	Usual	Diff
Number of QALYs						
Truth	0.540	0.485	0.054	0.540	0.485	0.054
FFE	0.534	0.481	0.053	0.524	0.472	0.052
FRE	0.532	0.480	0.053	0.529	0.477	0.052
BFE	0.532	0.479	0.053	0.521	0.469	0.052
BRE	0.531	0.479	0.052	0.528	0.477	0.051
Number of LYs						
Truth	0.786	0.740	0.046	0.786	0.740	0.046
FFE	0.782	0.739	0.044	0.774	0.731	0.043
FRE	0.781	0.737	0.044	0.777	0.733	0.043
BFE	0.780	0.736	0.044	0.771	0.727	0.044
BRE	0.780	0.736	0.043	0.776	0.733	0.042
Total costs						
Truth	€ 7,633	€ 6,116	€ 1,517	€ 7,633	€ 6,116	€ 1,517
FFE	€ 7,604	€ 6,103	€ 1,501	€ 7,624	€ 6,142	€ 1,482
FRE	€ 7,604	€ 6,105	€ 1,499	€ 7,597	€ 6,111	€ 1,486
BFE	€ 7,578	€ 6,081	€ 1,497	€ 7,592	€ 6,110	€ 1,481
BRE	€ 7,596	€ 6,099	€ 1,497	€ 7,622	€ 6,099	€ 1,523

[a] LY: Life year; QALY: Quality Adjusted LY; FFE: Frequentist fixed effects method; FRE: Frequentist random effects method; BFE: Bayesian fixed effects method; BRE: Bayesian random effects method, New Int = New Intervention, Usual = Usual Care, Diff = Difference between two intervention arms

Table A4.10: Health economic outcomes for two of the seven scenarios. Means of coverage, bias and mean absolute deviance (MAD) of the difference between two interventions, over 1,000 repetitions.[a]

Scenario	Scenario 2			Scenario 3		
Intervention arm	Coverage	Bias	MAD	Coverage	Bias	MAD
Number of QALYs						
FFE	97.6%	-0.9%	8.5%	97.9%	-1.5%	8.1%
FRE	98.7%	-1.1%	8.8%	98.5%	-1.3%	8.1%
BFE	98.9%	0.2%	8.6%	99.1%	-0.8%	8.5%
BRE	100.0%	-2.5%	11.4%	99.9%	-0.8%	8.8%
Number of LYs						
FFE	98.2%	-2.9%	14.5%	97.6%	-4.0%	13.8%
FRE	98.6%	-3.3%	15.0%	98.3%	-3.5%	13.8%
BFE	99.1%	-1.0%	14.6%	99.0%	-2.5%	14.2%
BRE	100.0%	-5.1%	19.0%	99.8%	-2.3%	14.8%
Total costs						
FFE	98.3%	-0.5%	5.7%	98.9%	-0.1%	5.4%
FRE	99.0%	-0.6%	5.9%	99.5%	-0.2%	5.4%
BFE	99.0%	-0.4%	5.8%	99.5%	-0.4%	5.5%
BRE	100.0%	-0.3%	7.8%	99.9%	-0.2%	5.9%

[a] LY: Life year; QALY: Quality Adjusted LY; FFE: Frequentist fixed effects method; FRE: Frequentist random effects method; BFE: Bayesian fixed effects method; BRE: Bayesian random effects method

Table A4.11: Health economic outcomes for two of the seven scenarios. Means of coverage, bias and mean absolute deviance (MAD) of the difference between two interventions, over 1,000 repetitions.[a]

Scenario	Scenario 5			Scenario 6		
Intervention arm	Coverage	Bias	MAD	Coverage	Bias	MAD
Number of QALYs						
FFE	97.6%	-3.0%	8.5%	96.7%	-4.5%	9.1%
FRE	99.2%	-2.8%	8.5%	99.7%	-3.9%	9.2%
BFE	98.5%	-2.6%	9.0%	98.0%	-3.5%	9.1%
BRE	100.0%	-3.6%	11.2%	100.0%	-5.8%	13.8%
Number of LYs						
FFE	97.6%	-4.5%	14.1%	96.8%	-5.4%	15.1%
FRE	98.9%	-4.0%	14.2%	99.3%	-5.4%	15.5%
BFE	98.2%	-3.7%	15.0%	98.7%	-3.4%	15.1%
BRE	100.0%	-5.1%	17.7%	100.0%	-7.4%	21.2%
Total costs						
FFE	98.1%	-1.1%	5.4%	97.6%	-2.3%	6.2%
FRE	99.7%	-1.2%	5.5%	99.5%	-2.1%	6.2%
BFE	98.7%	-1.4%	5.7%	98.6%	-2.4%	6.4%
BRE	100.0%	-1.3%	6.8%	100.0%	0.4%	10.5%

[a] LY: Life year; QALY: Quality Adjusted LY; FFE: Frequentist fixed effects method; FRE: Frequentist random effects method; BFE: Bayesian fixed effects method; BRE: Bayesian random effects method

Table A4.12: Statistical power for the health economic outcomes in all scenarios. Means over 1,000 repetitions.[a]

	Scen 1	Scen 2	Scen 3	Scen 4	Scen 5	Scen 6	Scen 7
Difference in number of QALY							
FFE	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
FRE	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
BFE	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
BRE	98.8%	93.3%	100.0%	96.7%	96.9%	83.9%	81.9%
Difference in number of LY							
FFE	100.0%	99.7%	100.0%	99.9%	99.9%	99.9%	99.4%
FRE	99.7%	98.7%	99.9%	99.3%	99.1%	96.9%	96.7%
BFE	99.8%	99.8%	99.8%	99.6%	99.4%	99.7%	99.8%
BRE	65.9%	37.4%	98.0%	49.7%	47.8%	17.5%	19.9%
Difference in total costs							
FFE	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
FRE	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
BFE	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
BRE	98.4%	97.8%	100.0%	99.2%	99.4%	97.1%	97.9%
Total costs per QALY							
FFE	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
FRE	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
BFE	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
BRE	99.7%	94.9%	100.0%	97.9%	97.7%	88.0%	83.7%

[a] LY: Life year; QALY: Quality Adjusted LY; FFE: Frequentist fixed effects method; FRE: Frequentist random effects method; BFE: Bayesian fixed effects method; BRE: Bayesian random effects method

Chapter 5

Mix and Match

A simulation study on the impact of mixed-treatment comparison methods on health-economic outcomes

P. Vemer, M.J. Al, M. Oppe, M.P.M.H. Rutten-van Mölken

Submitted

Acknowledgments: The authors wish to acknowledge the support of Prof. Lidia Arends, Ph.D. and Prof. Jos Kleijnen, Ph.D. for their valuable contributions to the design and execution of this study, and participants of the LoLa HESG meeting in May 2012 for their useful comments.

ABSTRACT

Background Decision-analytic cost-effectiveness (CE) models combine many parameters, often obtained after meta-analysis

Aim We compared different methods of mixed-treatment analysis (MTC), especially with respect to health-economic (HE) outcomes like (quality adjusted) life years and costs.

Methods Trials were drawn from a simulated reference population, comparing two of four fictitious interventions. The goal was to estimate the CE between two of these. The amount of heterogeneity between trials was varied in scenarios. Parameter estimates were combined using direct comparison, MTC methods proposed by Song and Puhan, and Bayesian generalized linear fixed effects (GLMFE) and random effects models (GLMRE). Parameters were entered into a Markov model. Parameters and HE outcomes were compared with the reference population using coverage, statistical power, bias and mean absolute deviation (MAD) as performance indicators. Each analytical step was repeated 1,000 times.

Results The direct comparison was outperformed by the MTC methods on all indicators, Song's method yielded low bias and MAD, but uncertainty was overestimated. Puhan's method had low bias and MAD and did not overestimate uncertainty. GLMFE generally had the lowest bias and MAD, regardless of the amount of heterogeneity, but uncertainty was overestimated. GLMRE showed large bias and MAD and overestimated uncertainty. Song's and Puhan's methods lead to the least amount of uncertainty, reflected in the shape of the CE acceptability curve. GLMFE showed slightly more uncertainty

Conclusion Combining direct and indirect evidence is superior to using only direct evidence. Puhan's method and GLMFE are preferred.

5.1 INTRODUCTION

In 2006, The Netherlands implemented conditional reimbursement of potentially innovative, but expensive hospital drugs, on the condition that further real-life evidence is collected.[1] After four years, a new reimbursement decision is made, based on all evidence available. Unfortunately, new drugs are often compared to placebo or standard care and the interventions of interest vary by country or over time. Trials incorporating all competing interventions are impractical at best, impossible at worst.[2] Much of the evidence will therefore come from indirect comparisons.

The simplest form is the indirect treatment comparison, where the relative efficacy between two interventions is obtained through a common comparator.[3] With three or more interventions, there may be several direct and indirect comparisons, analyzed simultaneously using mixed treatment comparisons (MTC).

To aid reimbursement decision making, a probabilistic decision-analytic cost-effectiveness (CE) model is often used, using parameters that are calculated from evidence combined using meta-analysis. The choice of meta-analysis method can considerably affect final CE estimates.[4] Most studies comparing meta-analysis methods focused on a single treatment effect (e.g.[5-8]). However, in modeling studies a wide range of model parameters need to be estimated.[9] In this study we aimed to compare the performance of standard methods of MTC when applied to different types of model parameters, especially with respect to their impact on health-economic (HE) outcomes. A similar comparison of direct meta-analysis methods is reported separately.[10]

5.2 METHODS

5.2.1 Simulation study

The simulation comprised several steps (figure 5.1). In **step 1: Create reference population**, we simulated a superpopulation[11] containing 50,000 patients. The disease progression was simulated four times for each patient, once for each of four fictitious interventions. The mean values of parameters and HE outcomes within the superpopulation represent the 'truth' with which parameter estimates and HE outcomes were compared, referred to as reference parameters and reference outcomes. Parameters included transition and event probabilities, maintenance and event costs, utilities and utility-decrements due to an event. HE outcomes included (quality adjusted) life years (QALY/LY), intervention and maintenance costs, number of events, incremental CE ratio (ICER) and CE acceptability curves (CEAC).

In **step 2: Trial selection**, we sampled trials comparing two treatments from the reference population. For each of the trials we calculated trial parameters. In **step 3: Meta-analysis**

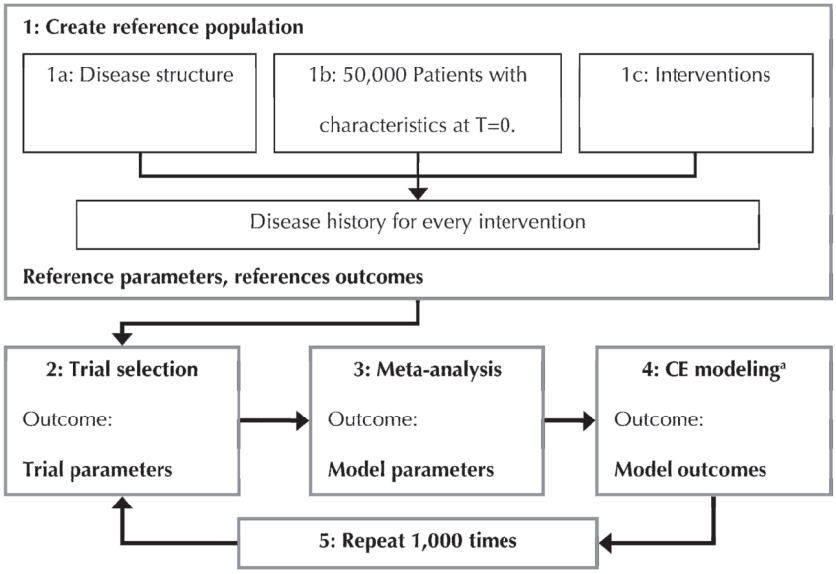

Figure 5.1: Design of the simulation study.
[a] HE: health-economic, CE: cost-effectiveness

we calculated model parameters, by pooling trial parameters using several methods of meta-analysis (paragraph 5.3.4). We used a CE model in **step 4: CE modeling**, which was filled with a set of model parameters obtained by each of the methods of meta-analysis. Probabilistic sensitivity analysis (PSA; 1,100 iterations) yielded model outcomes.

To study systematic differences between the methods of meta-analysis, we repeated steps 2 to 4 in **step 5: Repeat** in 1,000 repetitions.

5.2.2. Disease and model structure

We modeled a progressive, chronic disease with events, during which symptoms temporarily worsen, simulated using a four-stage Markov model (figure 5.2).

To simulate disease progression, we first defined the Reference Disease Progression (RDP), which can be thought of as the disease progression of an untreated, base-case patient. The RDP was modified based on individual patient characteristics and interventions, to simulate a heterogeneous population of individual patients. Table 5.1 shows characteristics of the reference population. By sampling from sub-populations, it was possible to add heterogeneity to trials in relevant scenarios

How patient characteristics and interventions influenced the RDP is stated in tables in the appendix A5.1-A5.6.

Focusing on the interventions, “No Intervention” has no effects on the RDP. “Old Intervention” decreases the probability of an event and has a positive effect on mortality, with a

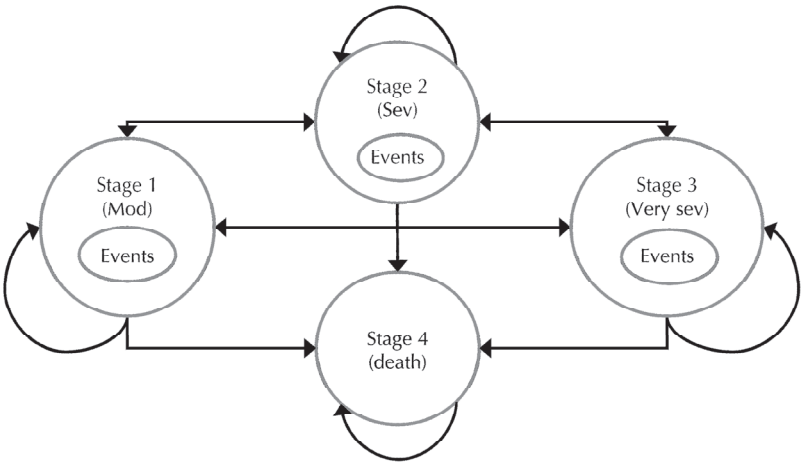

Figure 5.2: Design of the chronic disease model.

Table 5.1: Characteristics of the simulated patient population.

Size simulated cohort	50,000
Starting disease stage	5/8 in moderate, 2/8 in severe and 1/8 in very severe
Gender	50% male, 50% female
Age in years	18 – 34; 35 – 64; 65+
	Determined by a random draw from a uniform distribution from 18 to 75
Developed/developing country.	50% from developed countries, 50% developing countries
Body Mass Index (BMI)	<25 (average or low); 25-30 (high); >30 (obese),
	Determined by a random draw from a normal distribution with mean 23 and standard deviation of 4.
Smoking status	30% smokers, 70% non-smokers

one-off cost of €250 at the beginning of treatment. "Usual Care" decreases the probability of disease progression at €60 per month. "New Intervention" costs €350 per month and decreases the probability of disease progression, increases the probability of moving to a better disease stage and decreases the probability of an event. The intervention effects are dependent on the disease stage of the patient.

Changes to parameters were additive across patient characteristics and interventions. For example, for a female patient aged 35-64 who gets New Intervention, the probability to move from the severe disease stage to death was

10% (RDP) - 2% (modification for gender) + 4% (age) - 3% (intervention) = 9%.

Table 5.2: Reference outcomes, per patient per 12 cycles/months – Mean (Standard deviation).[a]

Variables	Usual Care	New Intervention	Difference
QALYs	0.485 (0.232)	0.540 (0.231)	0.054
LYs	0.740 (0.328)	0.786 (0.313)	0.046
Intervention costs	€ 533 (€ 236.24)	€ 3,300 (€ 1,310)	€ 2,770
Maintenance costs	€ 3,260 (€ 2,080)	€ 3,070 (€ 1,810)	- € 180
Event costs	€ 2,330 (€ 2,610)	€ 1,260 (€ 1,780)	- € 1,070
Total costs	€ 6,120 (€ 4,340)	€ 7,630 (€ 3,830)	€ 1,520
Number of cycles in:			
Moderate disease	5.171 (3.750)	6.209 (3.965)	1.038
Severe disease	2.477 (2.512)	2.313 (2.507)	-0.164
Very severe disease	1.238 (1.850)	0.911 (1.554)	-0.327
Death	3.114 (3.937)	2.567 (3.751)	-0.547
Number of events	1.160 (1.259)	0.630 (0.856)	-0.530
Proportion surviving	49.9%	58.3%	8.4%pt
ICER, total costs per QALY			€ 28,020

[a] LY: Life year; QALY: Quality adjusted LY; ICER = Incremental cost-effectiveness ratio; %pt: percentage points

The comparison of interest is that between New Intervention and Usual Care. Table 5.2 shows the reference outcomes when applying these interventions to the complete patient population.

The structure of the HE model mirrors the disease progression. We assumed that trial data was collected each month during one year. Likewise, the time horizon of the HE model was 1 year, with monthly cycles. We did not apply discounting. Simulation and modeling was performed using SAS 9.2 and WinBUGS 1.4.3.

5.2.3. Scenarios

The amount of heterogeneity in the trials sampled in step 2: Trial selection was varied in eight scenarios. Heterogeneity in the meta-analysis literature is any kind of variability between different studies.[12] All scenarios contained data from nine trials, with 500 patients in each of the two treatment arms. The comparisons made in each of the trials can be found in figure 5.3. It is clear from this graph that the nine trials provide evidence for all available contrasts. A similar structure can be found in Hasselblad[13] and Lu and Ades.[14]

The heterogeneity in all eight scenarios is described in table 5.3. In scenario 8 we used heterogeneity definitions at extreme values. This scenario included as a stress test for the methods, with very high amounts of heterogeneity between trials. In practice, trials that display this amount of heterogeneity would (should) not be combined.

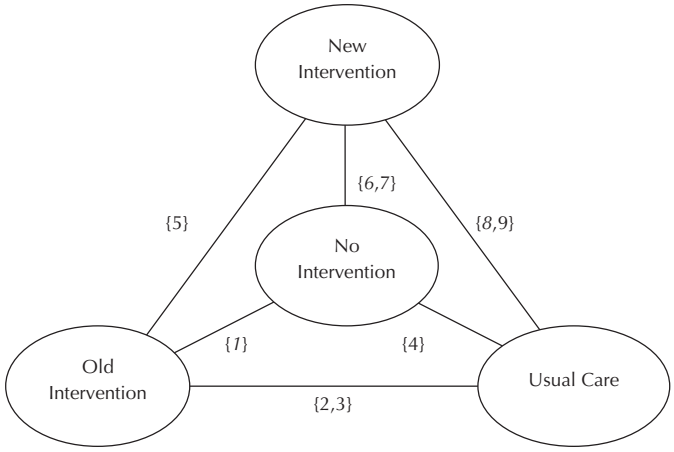

Figure 5.3: Evidence network for simulation study[a]

[a] The figures in curly brackets are the trial numbers making the corresponding comparisons, as described in the text. Trials 1, 6 and 8 are trials that may be drawn from a subpopulation in selected scenarios.

Table 5.3: Overview of different scenarios in the simulation study.

Scenario	Added heterogeneity with effect on disease progression
1	9 randomly drawn trials, with 500 patients in each of the treatment arms.
2	8 randomly drawn trials Non-random trial 1 (Old Intervention versus No Intervention), with worse average health.[a]
3	8 randomly drawn trials Non-random trial 6 (New Intervention versus No Intervention), with worse average health.[a]
4	8 randomly drawn trials Non-random trial 8 (New Intervention versus Usual Care directly), with worse average health.[a]
5	7 randomly drawn trials Non-random trial 1 (Old Intervention versus No Intervention), with worse average health.[a] Non-random trial 6 (New Intervention versus No Intervention), with lower average age.[b]
6	6 randomly drawn trials Non-random trial 1 (Old Intervention versus No Intervention), with worse average health.[a] Non-random trial 6 (New Intervention versus No Intervention), with lower average age.[b] Non-random trial 8 (New Intervention versus Usual Care directly), with higher average age.[c]
7	6 randomly drawn trials Non-random trials 1 (Old Intervention versus No Intervention), 6 (New Intervention versus No Intervention) and 8 (New Intervention versus Usual Care), with worse average health.[a]
8	6 randomly drawn trials Non-random trials 1 (Old Intervention versus No Intervention), 6 (New Intervention versus No Intervention) and 8 (New Intervention versus Usual Care), with worse average health.[a] Extreme scenario

[a] Trial contains, on average, patients with a higher age, more smokers and more obesity; patients have therefore on average a more rapid disease deterioration, higher event probability, higher maintenance costs, lower quality of life.

[b] Trial contains, on average, patients with a lower age; patients have therefore on average a slower disease deterioration.

[c] Trial contains, on average, patients with a higher age; patients have therefore on average a more rapid disease deterioration.

5.2.4. Methods of meta-analysis

In MTC, only measures of relative differences between treatments can be compared. Despite being used in many applications of MTC, the odds ratio (OR) is not commonly used in HE modeling. We have chosen to use the natural logarithm of the relative risk (ln(RR)) as relative measure of treatment benefit for the transition and event probabilities. For all non-relative variables in the model costs, quality of life weights and baseline values for the comparator , we used estimates from standard direct methods.[10,15]

As a baseline method, we combined all available direct evidence (DIRECT) on an ln(RR)-scale using the DerSimonian-Laird random effects method (DL).[15] The pooled estimate is calculated as a weighted average of individual study estimates, using the inverse of the within-study and between-study variance (heterogeneity) as weights. This is a relevant comparison, since it has been debated whether or not direct and indirect evidence can and should be combined, or even if indirect methods should be used at all. Since there is no reason not to use direct evidence when it is available, results on indirect treatment comparison methods were not reported separately in this paper.

The first MTC method is proposed by Song et al. (SONG).[17] They calculated a direct estimate using the DL method described above. Next, all possible indirect estimates are calculated. [18] The estimate of indirect association on a ln(RR)-scale between A and C, from the paired comparisons of A versus B and C versus B, is calculated as

$$\ln(RR_{AC})=\ln(RR_{AB})-\ln(RR_{CB}) \quad (1)$$

The variance of $\ln(RR)_{AC}$ can be obtained from

$$Var[\ln(RR_{AC})]=Var[\ln(RR_{AB})]+Var[\ln(RR_{CB})] \quad (2)$$

The SONG estimate of the association between A and C is calculated by performing a DL meta-analysis using all direct and indirect estimates.[15]

Puhan et al. performed a logistic regression (PUHAN).[18,19] A data set is first created based on summary tables from each included study. The number of data entries is equal to the number patients in each respective cell, with dummy variables for treatment as independent variables and the presence of an event as dependent variable. Since PUHAN uses logistic regression, ln(*OR*) is the only relative measure possible. We estimated ln(*RR*) using the ln(*OR*) estimate that come from the model, and the treatment effect of the comparator.

The most widely used method of meta-analysis is the Bayesian generalized linear model (GLM), either in a fixed effect (GLMFE) or random effect (GLMRE) variant.[20] The GLM is applicable for both direct meta-analysis and MTC. It allows the definition of many different possible link functions, depending on the nature of the data. GLMFE requires the trial data, the definition of a prior for the parameter of interest and a likelihood function

linking both. Defining r_{ik} as the number of events, out of the total number of patients in each arm n_{ik}, for arm k of trial i, we assumed that the data generation process follows a binomial likelihood:

$$r_{ik} \sim Binomial(p_{ik}, n_{ik}) \quad (3)$$

where p_{ik} represents the probability of an event in arm k of trial i. We modeled the probabilities of success p_{ik} on the logit-scale, the most commonly used link function for a binomial likelihood[20]:

$$\log it(p_{ik}) = \mu_i + \delta_{i,1k} * I(k \neq 1) \quad (4)$$

where $I(k \neq 1)$ takes the value 0 when intervention k is equal to comparator 1, and 1 otherwise. The μi are trial-specific log-odds in the comparator arm, and $\delta_{i,1_k}$ are trial-specific log-odds for the treatment group compared to control. For the GLMRE, we assumed

$$\delta_{i,1k} \sim N(d_{i,1k}, \sigma^2) \quad (5)$$

where σ_2 represents the between-trial heterogeneity. For the GLMFE, (4) reduces to

$$\log it(p_{ik}) = \mu_i + d_{i,1k} * I(k \neq 1) \quad (6)$$

which is equivalent to setting σ^2 in (5) to zero, thus assuming homogeneity of the underlying treatment effects. Using ln(*RR*) is possible, but may run into computational problems.[21] We therefore estimated ln(*RR*) using the ln(*OR*) estimate and the treatment effect of the comparator.

We used a flat beta prior *Beta*(0.5,0.5) for all baseline transitions, and a flat normal prior *N*(0,1*E*12) for all other baseline parameters. We used a flat normal prior centered on *N*(0,1*E*8) for all treatment effects of the comparator. For GLMRE we used the inverse of a squared uniform distribution *U*(0.001,10) for the between-trial heterogeneity. The minimum value of this prior was not 0, to avoid numerical problems.

Conceptually, confidence intervals in frequentist statistics and credibility intervals in Bayesian statistics have very different interpretations (e.g.[22,23]). However, for convenience and legibility, we abbreviate both as CI. For each pooled parameter estimate, we report the mean and the 95% CI. Interested readers may request code on both the simulation study and the methods of meta-analysis from the corresponding author.

5.2.5. Comparing performance

We assumed that a researcher doing a meta-analysis aims to estimate the CE of the New Intervention compared to Usual Care in the entire patient population, not a specific subgroup. Evidence on other interventions is solely used to provide extra evidence for this comparison. We further assumed that the researcher is unaware of the fact that heterogeneity, when present, was caused by sampling from subgroups. To the researcher, heterogeneity is either caused by random sampling or unobserved trial differences. These assumptions are made, because if these differences in design are known, either the trials would not be synthesized at all, or a way has to be found to control for these differences. These assumptions made it possible to judge the performance of the different methods of meta-analyses by comparing model parameters and HE outcomes with the reference values. Because the same patients were included to calculate HE outcomes for each method of meta-analysis, any difference between the methods can be attributed to the methods themselves (moderately dependent samples).[11]

Statistical performance is measured using coverage, statistical power, bias and mean absolute deviation (MAD). Coverage is the percentage of all repetitions, that the simulated CI covered the 'truth'. Since the coverage is based on 95% CIs, we would expect that, if all trials are drawn randomly, the coverage should on average be close to 95%.[5,11,24] Over-coverage, where the CI are so wide that coverage rates are above 95 per cent, suggests that the results are too conservative, thus leading to a loss of statistical power. In contrast, under-coverage, where the coverage rates are lower than 95 per cent, indicates over-confidence in the estimates. More simulations will incorrectly detect a significant result, leading to higher than expected type I errors.[11] We said a method underestimated uncertainty if the coverage was smaller than 90%; and overestimated if the coverage was higher than 98%.

Statistical power is the percentage of all repetitions where the simulated result yields a statistically significant difference between the two treatments. Bias is the difference between the point estimate in the simulated data set and the true population value, averaged over all repetitions. MAD is the average, over all repetitions, of the absolute value of the bias. The MAD indicates how far the estimated value was from the 'truth', regardless of whether it was too high or too low.

5.3 RESULTS

5.3.1. Model parameters for one set of trials

Figure 5.4 compares the methods on one example parameter for each of the scenarios, using only the first repetition. From bottom to top, we compare the different meta-analysis models for the eight scenarios. Each dot represents the point estimate for the parameter,

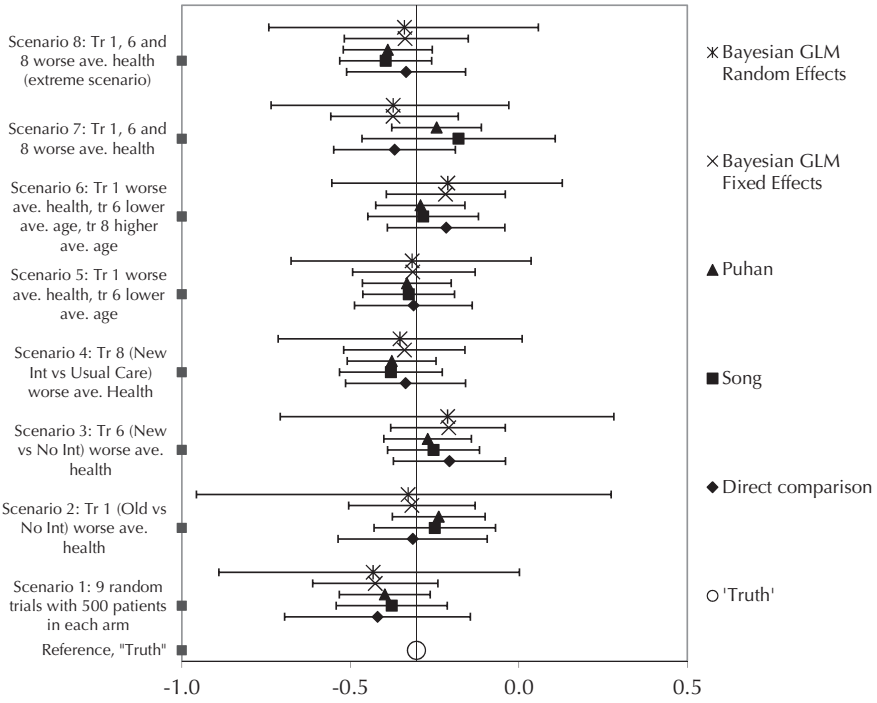

Figure 5.4: Meta-analysis on the logarithm of the risk ratio of the transition from the severe to very severe disease stage, for the New Intervention arm compared to the Usual Care arm, for one repetition.[a]

[a] All scenarios have nine trials, each with 500 patients in both treatment arms

in this case the transition probability from severe to very severe disease, and the bars the estimated CIs. The 'true' population value is displayed at the bottom. As can be seen, when all trials were drawn randomly (scenario 1), GLMRE had the broadest CI, followed by DIRECT, GLMFE and SONG. PUHAN had the smallest CI. All methods had the true parameter value in its CI and the point estimates were all very similar. In the other scenarios, each with a different amount of heterogeneity, we see a similar pattern as in scenario 1, except that in scenario 7 SONG had a relatively larger CI. The point estimate of SONG and PUHAN, and of GLMFE and GLMRE are very similar.

Based on similar patterns for other parameters (not shown), we can conclude that DIRECT and GLMRE yielded the widest CI. GLMFE had a point estimate that is generally closer to the true parameter value than DIRECT, with a smaller CI. The smallest CI was found for SONG and PUHAN. In all scenarios, for all methods, the true parameter value lay within the CI of the estimated parameters.

5.3.2. Model parameters for 1,000 repetitions

The results from the previous paragraph might be due to chance. To see if there were systematic differences, we now discuss parameter estimates averaged over 1,000 repetitions.

Table 5.4: Summary of the results of meta-analysis on parameters of the health-economic model, which require network meta-analysis. Means over 1,000 repetitions.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7	Scenario 8
Total number of parameters	12	12	12	12	12	12	12	12
Parameters influenced by added heterogeneity	0	9	9	9	9	9	9	9
Heterogeneity in the following trials	-	Trial 1(Old Int vs No Int)	Trial 6 (New Int vs Old Int)	Trial 8 (New Int vs Usual)	Trial 1 + Trial 6	Trial 1 + Trial 6 + Trial 8	Trial 1 + Trial 6 + Trial 8	Trial 1 + Trial 6 + Trial 8
Total number of parameters for which:								
Mean coverage < 90% (underestimation of uncertainty)								
Direct comparison (DIRECT)	0	0	0	0	0	0	0	4
Song's method (SONG)	0	0	0	0	0	0	0	6
Puhan's method (PUHAN)	0	0	0	0	0	0	0	7
Bayesian GLM FE method (GLMFE)	0	0	0	0	0	0	0	2
Bayesian GLM RE method (GLMRE)	0	0	0	0	0	0	0	0
Mean coverage > 98% (overestimation of uncertainty)								
DIRECT	1	2	4	0	1	2	2	1
SONG	11	11	11	4	12	10	7	3
PUHAN	3	5	6	1	5	4	3	1
GLMFE	10	10	9	6	9	5	4	3
GLMRE	12	12	12	12	12	12	12	11
Mean bias 1%-2%								
DIRECT	2	2	2	0	1	1	0	2
SONG	0	0	1	1	0	2	1	1
PUHAN	0	0	0	2	0	2	1	0
GLMFE	4	4	2	0	3	3	1	1
GLMRE	1	3	3	1	3	2	1	1

Table 5.4: Summary of the results of meta-analysis on parameters of the health-economic model, which require network meta-analysis. Means over 1,000 repetitions. (Continued)

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7	Scenario 8
Total number of parameters	12	12	12	12	12	12	12	12
Parameters influenced by added heterogeneity	0	9	9	9	9	9	9	9
Heterogeneity in the following trials	-	Trial 1(Old Int vs No Int)	Trial 6 (New Int vs Old Int)	Trial 8 (New Int vs Usual)	Trial 1 + Trial 6	Trial 1 + Trial 6 + Trial 8	Trial 1 + Trial 6 + Trial 8	Trial 1 + Trial 6 + Trial 8
Total number of parameters for which:								
Mean bias > 2%								
DIRECT	0	0	0	9	0	4	9	10
SONG	0	0	0	8	0	3	8	9
PUHAN	0	0	0	8	0	2	8	10
GLMFE	1	1	2	9	1	5	9	11
GLMRE	9	8	8	10	8	9	11	11
Mean MAD[a] 4%-7%								
DIRECT	5	7	5	4	6	4	4	3
SONG	8	9	9	8	9	7	8	3
PUHAN	8	8	8	9	8	8	9	3
GLMFE	5	5	5	4	5	5	4	2
GLMRE	4	4	4	4	4	4	4	2
Mean MAD[a] > 7%								
DIRECT	6	4	6	7	5	7	7	9
SONG	1	0	0	2	0	2	3	8
PUHAN	0	0	0	0	0	0	0	7
GLMFE	7	7	7	8	7	7	8	10
GLMRE	8	8	8	8	8	8	8	10

[a] MAD = Mean absolute deviation, minimum found is 2.6%.

Table 5.4 shows the number of parameters that correspond to several threshold values of coverage, bias and MAD. No parameter had an average coverage below 90%, which we defined as underestimation of uncertainty, except in the extreme scenario 8. In this scenario GLMFE and GLMRE had the least amount of parameters for which uncertainty is underestimated. The least overestimation of uncertainty could generally be found with DIRECT and PUHAN, regardless of the amount of heterogeneity.

GLMRE had a large number of parameters with an average bias larger than 1% or even 2%. All methods had a large number of parameters with a large bias in scenario 4, where extra heterogeneity was added to trial 8, which directly compares the Usual Care with the New Intervention. In scenarios 6 to 8, where three out of nine trials have patients drawn from a subpopulation, all methods showed bias in several parameters. The lowest amount of bias was found in SONG and PUHAN, with a similar number of parameters in each category of bias.

For all methods, the estimated parameter value was quite far from the true population value. The minimum MAD, averaged over 1,000 estimates of the same parameters (not in graphs/tables), ranged from 2.6% for PUHAN to 4.2% for GLMRE. In other words, none of the methods estimated parameters with an average MAD lower than 2.6%. The maximum MAD, averaged over 1,000 estimates of the parameters, was 27.6% for GLMRE in the extreme scenario 8. This means that one of the parameters, in this case ln(RR) of the number of events in the severe disease stage, differed from the reference value by more than 27%, averaged over 1,000 repetitions. The discrepancy will therefore be much larger for individual repetitions. SONG and PUHAN generally had the lowest number of parameters in each of the categories of MAD.

Generally, SONG, GLMFE and GLMRE overestimated uncertainty for most parameters. PUHAN overestimated uncertainty for fewer parameters. Neither of these methods underestimated uncertainty, except in the extreme scenario. The bias and MAD was generally lowest for SONG and PUHAN, followed by GLMFE.

5.3.3. Health-economic outcomes for 1,000 repetitions

In table 5.5, we show the coverage, statistical power, bias and MAD for four scenarios. Information on other scenarios can be found in the appendix A5. It shows the range in values over the four types of HE outcomes, the difference in QALYs, LYs, number of events and total costs. PUHAN had a coverage closest to the benchmark of 95%. Only in case of heterogeneity (i.e. scenarios 7 and 8) did PUHAN overestimate uncertainty. Both GLMFE and GLMRE had a coverage above 99% for all methods, for all HE outcomes. No method underestimated uncertainty.

Regardless of heterogeneity, GLMRE had the lowest statistical power. For the difference in LYs, GLMRE had a statistical power below 10% in scenario 1, where all trials were drawn randomly, and even lower in scenarios with added heterogeneity. All methods had

Table 5.5: Coverage, statistical power, absolute value of the bias and mean absolute deviation (MAD) of health-economic outcomes for four of the eight scenarios.

	Direct comparison	Song's method	Puhan's method	GLM FE method	GLM RE method
Coverage, range in values over the four health-economic outcomesa					
Scenario 1: Nine randomly drawn trials	>98%	>98%	97.0%-97.3%	>98%	100%
Scenario 4: Eight randomly drawn trials; one trial drawn from a less health population	97.1%-98.6%	>98%	96.8%-97.9%	>99%	100%
Scenario 7: Six randomly drawn trials; three trials drawn from a less healthy population	97.2%-99.1%	97.9%-99.3%	96.3%-98.2%	>99%	100%
Scenario 8: Six randomly drawn trials; three trials drawn from a less healthy population (extreme scenario)	>98%	>99%	90.0%-100%	>99%	100%
Statistical power, range in values over the four health-economic outcomesa					
Scenario 1: Nine randomly drawn trials	81.5%-100%	95.3%-100%	>99%	73.4%-100%	5.8%-95.9%
Scenario 4: Eight randomly drawn trials; one trial drawn from a less healthy population	76.3%-100%	93.5%-100%	>98%	56.8%-100%	4.1%-94.3%
Scenario 7: Six randomly drawn trials; three trials drawn from a less healthy population	79.3%-100%	91.9%-100%	>98%	60.3%-100%	3.6%-93.5%
Scenario 8: Six randomly drawn trials; three trials drawn from a less healthy population (extreme scenario)	70.0%-100%	83.7%-100%	94.1%-100%	13.0%-100%	0.5%-83.2%
Bias, range in values over the four health-economic outcomesa					
Scenario 1: Nine randomly drawn trials	0.4%-5.7%	0.2%-3.0%	0.2%-2.1%	0.3%-3.5%	0.3%-13.6%
Scenario 4: Eight randomly drawn trials; one trial drawn from a less healthy population	0.5%-11.8%	0.5%-5.5%	0.5%-5.4%	0.8%-9.3%	2.0%-6.3%
Scenario 7: Six randomly drawn trials; three trials drawn from a less healthy population	0.2%-10.1%	0.5%-9.7%	0.4%-8.1%	0.0%-7.7%	0.2%-17.8%
Scenario 8: Six randomly drawn trials; three trials drawn from a less healthy population (extreme scenario)	3.1%-11.9%	0.4%-11.9%	0.5%-9.8%	1.7%-10.5%	2.5%-17.5%

Table 5.5: Coverage, statistical power, absolute value of the bias and mean absolute deviation (MAD) of health-economic outcomes for four of the eight scenarios. (Continued)

	Direct comparison	Song's method	Puhan's method	GLM FE method	GLM RE method
MAD, range in values over the four health-economic outcomes[a]					
Scenario 1: Nine randomly drawn trials	6.0%-21.7%	5.1%-17.9%	4.9%-16.9%	6.2%-22.7%	6.9%-25.9%
Scenario 4: Eight randomly drawn trials; one trial drawn from a less healthy population	6.6%-23.5%	5.3%-18.4%	5.1%-17.4%	6.8%-25.1%	7.9%-29.9%
Scenario 7: Six randomly drawn trials; three trials drawn from a less healthy population	6.3%-22.8%	5.4%-19.2%	5.1%-18.0%	6.8%-24.1%	7.9%-28.5%
Scenario 8: Six randomly drawn trials; three trials drawn from a less healthy population (extreme scenario)	8.3%-22.5%	6.9%-19.9%	6.4%-18.0%	10.1%-27.6%	10.7%-31.0%

[a] MAD = Mean absolute deviation, the four health-economic outcomes are QALYs, LYs, number of events and total costs

a statistical power of 100% for the number of events and above 99% for total costs, in all scenarios. PUHAN generally had the lowest bias and MAD across all scenarios. GLMRE had the highest MAD for all HE outcomes in all scenarios. In the online tables A5.7 and A5.8, the results for the different HE outcomes are presented separately.

In figure 5.5 we show the CE acceptability curves (CEACs) for the heterogeneous scenario 7. The five graphs represent the methods we compared. In each graph, we show the CEAC of ten repetitions, the median and 2.5th and 97.5th percentiles over 1,000 repetitions. The vertical line indicates the true population ICER. Graphs for other scenarios can be found in the online figures A5.1 to A5.3. In this scenario where three trials are drawn from a less healthy population (scenario 7), we can see that SONG and PUHAN displayed a steeper shape than the other methods. This indicates that they were more certain of the CE of the New Intervention than the other methods. At a WTP of € 30,000 per QALY, which is close, but slightly above the true population ICER, the median likelihood that the New Intervention was cost-effective was 45%-60% for all methods. At higher WTPs, GLMFE and GLMRE were less certain than the other methods. With less heterogeneity (scenario 1 and 4), the CEACs express a higher certainty for all methods. Still, SONG and PUHAN seem to be the most certain of the CE, in these scenarios followed by GLMFE. With more heterogeneity (scenario 8) all methods displayed less certainty. SONG and PUHAN still had the most certainty around the CE. SONG had all the CEACs lying closest to each other.

Regardless of the amount of heterogeneity, SONG and PUHAN lead to the least amount of uncertainty. GLMFE model is slightly less certain. DIRECT and GLMRE have a lot of uncertainty, even at WTP values far from the true population ICER. They also display a lot of differences between the different repetitions.

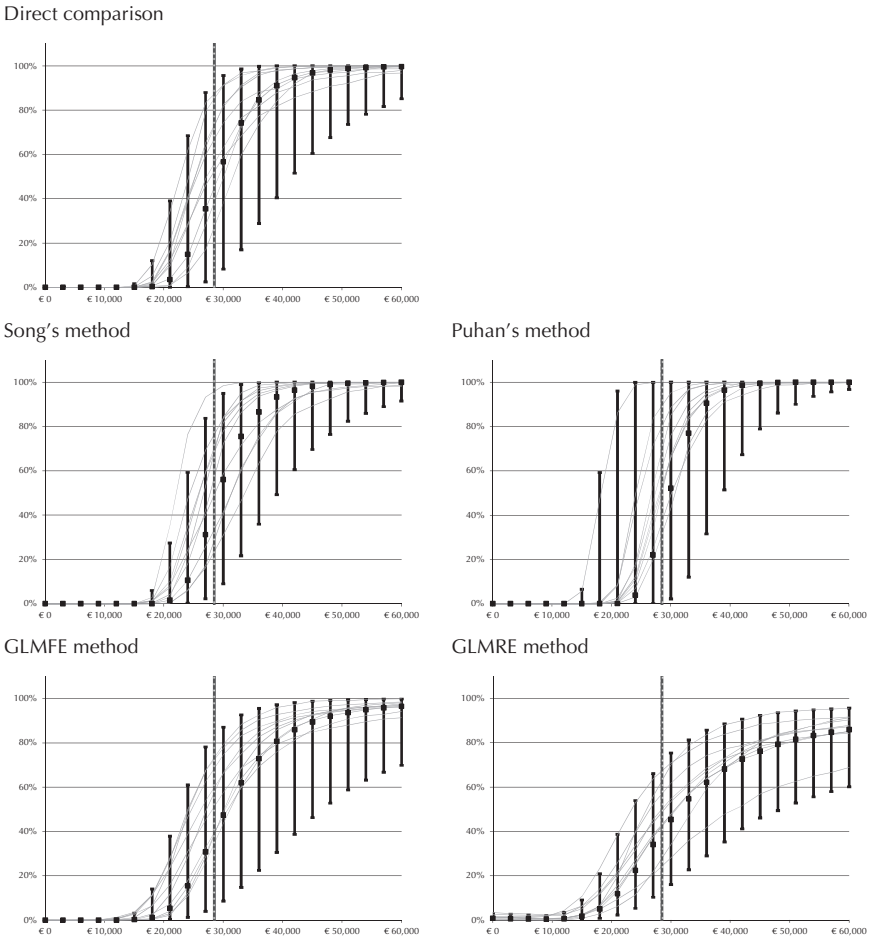

Figure 5.5: Cost-effectiveness acceptability curves (CEACs) for the five meta-analysis methods in the heterogeneous scenario 7.[a]

[a] The vertical lines depicts median, 2.5th and 97.5th percentile of the likelihood that the New Intervention is cost-effective compared with Usual care, at various threshold values of a QALY (averaged over 1,000 repetitions). The curves are the CEACs for the first 10 repetitions. The dotted vertical line is the 'true' population ICER.

5.4 DISCUSSION

In this study, we compared four methods of indirect meta-analysis in a simulation study and judged their statistical performance by creating a gold standard. On a parameter level, Puhan's method (PUHAN) showed the best performance, overestimating uncertainty for the fewest parameters with low bias and MAD. Song's method (SONG) and the Bayesian fixed effect generalized linear model (GLMFE) also had generally low bias and MAD.

On HE outcomes, PUHAN showed a coverage closest to 95%, regardless of heterogeneity. Only with high heterogeneity did PUHAN overestimate uncertainty. Both PUHAN and GLMFE performed best on bias and MAD, followed by SONG. GLMFE had a very high coverage, which we defined as overestimating uncertainty. The same is true for the Bayesian random effect generalized linear model (GLMRE), which also had the lowest statistical power and the highest MAD for all HE outcomes.

The use of these methods would lead to differences in policy decisions. Using either only the direct evidence or GLMRE would lead to more rejections of new treatments compared to the other methods or more unnecessary research. Generally speaking, sophisticated methods require more data than simple methods, because of the increased number of parameters. It is possible that the GLMRE method, which requires the largest number of parameter to be estimated, may have more desirable properties when more trials have to be combined. Unfortunately, this situation is unlikely within the scope of the expensive drug program in the Netherlands. Based on this study, we would recommend either PUHAN or GLMFE. PUHAN is easier to implement and more easily understood by physicians and policy makers who will be using the results. GLMFE is the most widely used method, but requires advanced knowledge of statistical programming.

In scenarios, we covered many likely situations. We have drawn all trials randomly, added heterogeneity on the different "legs" of the network, and changed the amount of heterogeneity. Compared to a few large trials, the effect of having more but smaller trials and trials with differences in trial sizes, on the performance of different methods is small.[10] We therefore feel our study results are generalizable to many other situations where parameters for a HE model are obtained through MTC.

However, the network is very "regular" with direct evidence for all treatment combinations. This is often not the case. New interventions are usually only compared to the latest alternative, or to placebo. Other forms of the evidence network are routinely found in MTC research. It remains open to further research whether adding irregularity to such networks will change the results of this study.

Another limitation is the choice of prior for the Bayesian models. In the case of meta-analysis, a small number of studies is extra vulnerable to the type of prior.[8,25] As we did not assume the researcher to have prior information, we used vague priors. Even though they are supposed to be "uninformative", they may influence outcomes, especially scale parameters.[25] We tested several different prior specifications but did not find any differences in outcomes.

Bayesian statistics at its heart is ideally suited for meta-analysis, since the premise of both are the same: prior available information is updated with new data.[26] However, Bayesian statistics is not ideally suited for a simulation study such as we have done. Bayesian statistics starts with the available data, which is examined in detail. This will drive all subsequent modelling decisions. Additionally, Bayesian outcomes are meaningless when

the model itself does not converge. Checking for convergence requires visual examination of plots, and careful examination of other outcome measures. However, all this is impossible in a simulation study, where many data sets are fitted one after the other.

In practice, there is still a strong preference to use direct over indirect evidence. One of the main concerns is that indirect comparisons may be subject to greater biases than direct comparisons.[17] They are essentially observational findings across trials, and may have similar biases. The Cochrane Handbook for Systematic Reviews of Interventions recommends that direct and indirect evidence is considered separately and direct comparisons should take precedence as a basis for forming conclusions.[8] In contrast, it has also been argued that it would be improper to exclude any evidence.[27] Our study seems to support this second view: the direct comparison has a smaller statistical power, leading to new interventions not being found statistically different from older interventions. The biases and MAD are also higher than the MTC methods, except for the GLMRE method.

5.5 CONCLUSION

In conclusion, when indirect evidence is available, regardless of the amount of heterogeneity present, combining all evidence is superior to using only the direct evidence. Puhan's method and GLMFE showed similar results, with GLMFE having the tendency to overestimate uncertainty, but also having lower average bias and MAD. Based on this study, where we had to combine nine trials in a network that includes evidence for all treatment combinations, we would recommend PUHAN or GLMFE as the preferred method of indirect meta-analysis.

5.6 LITERATURE

[1] NZa. Beleidsregel Dure Geneesmiddelen [Policy rule Expensive Drugs] (BR-CU-2017). 2011; Available at: http://www.nza.nl/regelgeving/beleidsregels/ziekenhuiszorg/BR-CU-2017. Accessed Jun/15, 2011.

[2] Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, et al. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1. Value Health 2011 Jun;14(4):417-428.

[3] Jansen P, Crawford B, Bergman G, Stam W. Bayesian Meta-Analysis of Multiple Treatment Comparisons: An Introduction to Mixed Treatment Comparisons. Value in Health 2008;11(5): 956-964.

[4] Oppe M, Al M, Rutten-van Molken M. Comparing methods of data synthesis: re-estimating parameters of an existing probabilistic cost-effectiveness model. Pharmacoeconomics 2011 Mar;29(3):239-250.

[5] Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med 2001;20(6):825-840.

[6] Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. Stat Med 2001;10(4):277-303.

[7] Sutton AJ, Higgins JP. Recent developments in meta-analysis. 2008;27(5):625-50.

[8] Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 5.0.2 updated September 2009. 2009.

[9] Riley RD, Simmonds MC, Look MP. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. J Clin Epidemiol 2007 May;60(5):431-9.

[10] Vemer P, Al M, Oppe M, Rutten-van Mölken M. A choice that matters? Simulation study on the impact of direct meta-analysis methods on health economic outcomes. Pharmacoeconomics 2013;31(8):719-30.

[11] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med 2006 Dec 30;25(24):4279-4292.

[12] Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 6;327(7414):557-60.

[13] Hasselblad V. Meta-analysis of multi-treatment studies. 1998;18(1):37-43.

[14] Lu G, Ades AE. Assessing evidence inconsistency in mixed treatment comparisons. J Am Stat Assoc 2006;101:447-459.

[15] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7(3):177-188.

[16] Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. 2009;338:b1147.

[17] Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997 Jun; 50(6):683-691.

[18] Puhan MA, Bachmann LM, Kleijnen J, Ter Riet G, Kessels AG. Inhaled drugs to reduce exacerbations in patients with chronic obstructive pulmonary disease: a network meta-analysis. BMC Med 2009 Jan 14;7:2.

[19] Strassmann R, Bausch B, Spaar A, Kleijnen J, Braendli O, Puhan MA. Smoking cessation interventions in COPD: a network meta-analysis of randomised trials. Eur Respir J 2009 Sep; 34(3):634-640.

[20] Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. Med Decis Making 2013 Jul;33(5):607-617.

[21] van Valkenhoef G, Ades AE. Evidence synthesis assumes additivity on the scale of measurement: response to "Rank reversal in indirect comparisons" by Norton et al. Value Health 2013 Mar-Apr;16(2):449-451.

[22] Jaynes E. Confidence Intervals vs Bayesian Intervals. In: Harper W, Hooker CA, editors. Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science Dordrecht: D. Reidel; 1976. p. 175.

[23] O'Hagan A, Luce B. A Primer on Bayesian Statistics in Health Economics and Outcomes Research. Sheffield: Centre for Bayesian Statistics in Health Economics; 2003.

[24] Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive missing-data strategies in modern missing-data procedures. Psychological Methods 2001;6:330-351.

[25] Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Stat Med 2005 Aug 15;24(15):2401-2428.

[26] Berry DA. Bayesian approaches for comparative effectiveness research. Clin Trials 2012 Feb; 9(1):37-47.

[27] Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Stat Med 2004 Oct 30;23(20):3105-24.

Table A5.1: Health-economic outcomes for three of the eight scenarios. Means of coverage, bias and mean absolute deviance (MAD) of the difference between two interventions, over 1,000 repetitions.

Scenario	Scenario 1				Scenario 2				Scenario 3			
	Nine randomly drawn trials				Eight randomly drawn trials; one trial drawn from a less healthy population				Eight randomly drawn trials; one trial drawn from a less healthy population			
	Coverage	Bias	MAD	Stat Power	Coverage	Bias	MAD	Stat Power	Coverage	Bias	MAD	Stat Power
Number of QALYs												
Direct comparison	98,9%	-3,6%	13,0%	99,6%	97,5%	-4,2%	13,1%	99,2%	97.9%	-3.7%	13.1%	99,6%
Song	98,6%	-1,9%	10,6%	99,9%	98,1%	-3,2%	10,6%	100,0%	98.4%	-5.1%	11.3%	100.0%
Puhan	97,2%	-1,3%	10,0%	100,0%	96,7%	-1,9%	10,0%	100,0%	96.4%	-4.1%	10.6%	100.0%
GLM Fixed eff	99.3%	-2.4%	13.6%	99.5%	98.9%	-2.9%	14.2%	99.2%	99.4%	-2.5%	14.2%	99.3%
GLM Rand eff	100.0%	-9.5%	16.1%	40.3%	100.0%	-9.5%	16.2%	41.1%	100.0%	-9.3%	16.2%	40.8%
Number of LYs												
Direct comparison	98,9%	-5,7%	21,7%	81,5%	97,7%	-6,6%	21,9%	80,6%	97.8%	-6.0%	21,9%	81,0%
Song	98,4%	-3,0%	17,9%	95,3%	97,5%	-4,7%	17,8%	93,5%	98.0%	-7.8%	18.8%	93.1%
Puhan	97,2%	-2,1%	16,9%	99,3%	96,7%	-2,7%	16,9%	98,9%	96.6%	-6.2%	17.6%	98.3%
GLM Fixed eff	99.6%	-3.5%	22.7%	73.4%	98.9%	-4.0%	23.5%	71.2%	99.6%	-3.5%	23.3%	71.4%
GLM Rand eff	100.0%	-13.6%	25.9%	5.8%	100.0%	-13.4%	25.9%	5.6%	100.0%	-13.4%	26.0%	6.2%
Number of Events												
Direct comparison	98,4%	-0,4%	6,0%	100,0%	98,3%	-0,3%	5,9%	100,0%	99.6%	-0.1%	7,0%	100,0%
Song	98,4%	-0,2%	5,1%	100,0%	98,7%	-0,1%	4,9%	100,0%	99.1%	-1.0%	5.1%	100.0%
Puhan	97,0%	0,2%	4,9%	100,0%	97,3%	0,3%	4,8%	100,0%	99.0%	-0.5%	4.9%	100.0%
GLM Fixed eff	98.7%	-0.3%	6.2%	100.0%	98.9%	-0.4%	6.1%	100.0%	99.6%	-0.2%	6.1%	100.0%
GLM Rand eff	100.0%	-2.7%	6.9%	90.6%	100.0%	-2.8%	6.6%	92.4%	100.0%	-2.8%	6.7%	90.9%

Table A5.1: Health-economic outcomes for three of the eight scenarios. Means of coverage, bias and mean absolute deviance (MAD) of the difference between two interventions, over 1,000 repetitions. (Continued)

Scenario	Scenario 1				Scenario 2				Scenario 3			
	Nine randomly drawn trials				Eight randomly drawn trials; one trial drawn from a less healthy population				Eight randomly drawn trials; one trial drawn from a less healthy population			
	Coverage	Bias	MAD	Stat Power	Coverage	Bias	MAD	Stat Power	Coverage	Bias	MAD	Stat Power
Total costs												
Direct comparison	99,0%	0,4%	8,0%	100,0%	99,1%	0,1%	7,8%	100,0%	99.3%	0.1%	8,2%	100,0%
Song	98,6%	0,8%	6,6%	100,0%	98,7%	0,5%	6,3%	100,0%	99.2%	0.3%	6.6%	100.0%
Puhan	97,3%	0,8%	6,3%	100,0%	97,7%	0,7%	6,1%	100,0%	98.6%	0.4%	6.2%	100.0%
GLM Fixed eff	99.7%	0.3%	8.4%	100.0%	99.3%	0.4%	8.3%	100.0%	99.6%	0.2%	8.2%	100.0%
GLM Rand eff	100.0%	0.3%	8.9%	95.9%	100.0%	0.3%	8.5%	96.1%	100.0%	0.2%	8.4%	94.8%

Table A5.2: Health-economic outcomes for two of the eight scenarios. Means of coverage, bias and mean absolute deviance (MAD) of the difference between two interventions, over 1,000 repetitions.

Scenario	Scenario 4				Scenario 5			
	Eight randomly drawn trials; one trial drawn from a less healthy population				Seven randomly drawn trials; one trial drawn from a less healthy population; one from a relatively younger population			
	Coverage	Bias	MAD	Stat Power	Coverage	Bias	MAD	Stat Power
Number of QALYs								
Direct comparison	97,1%	-8,4%	14,4%	98,7%	98,6%	-3,5%	13,1%	99,8%
Song	98,3%	-4,1%	11,2%	100,0%	99,2%	-2,2%	10,5%	100,0%
Puhan	96,8%	-4,0%	10,5%	100,0%	97,7%	-1,1%	9,9%	100,0%
GLM Fixed eff	99.2%	-7.0%	15.2%	97.9%	99.5%	-2.0%	13.8%	99.5%
GLM Rand eff	100.0%	-13.6%	18.5%	34.5%	100.0%	-9.0%	15.8%	43.6%

Table A5.2: Health-economic outcomes for two of the eight scenarios. Means of coverage, bias and mean absolute deviance (MAD) of the difference between two interventions, over 1,000 repetitions. (Continued)

Scenario	Scenario 4				Scenario 5			
	Eight randomly drawn trials; one trial drawn from a less healthy population				Seven randomly drawn trials; one trial drawn from a less healthy population; one from a relatively younger population			
	Coverage	Bias	MAD	Stat Power	Coverage	Bias	MAD	Stat Power
Number of LYs								
Direct comparison	97,1%	-11,8%	23,5%	76,3%	98,1%	-5,2%	21,9%	82,7%
Song	98,9%	-5,5%	18,4%	93,5%	98,5%	-3,5%	17,8%	95,1%
Puhan	97,1%	-5,4%	17,4%	98,6%	97,1%	-1,6%	16,7%	99,2%
GLM Fixed eff	99.3%	-9.3%	25.1%	56.8%	99.2%	-2.6%	22.9%	74.0%
GLM Rand eff	100.0%	-18.9%	29.2%	4.1%	100.0%	-12.4%	25.2%	7.7%
Number of Events								
Direct comparison	98,1%	-2,6%	6,6%	100,0%	98,9%	-0,5%	6,3%	100,0%
Song	98,1%	-0,8%	5,3%	100,0%	99,2%	-0,1%	5,1%	100,0%
Puhan	97,8%	-0,7%	5,1%	100,0%	99,0%	0,2%	5,1%	100,0%
GLM Fixed eff	99.6%	-2.2%	6.8%	100.0%	99.2%	-0.5%	6.5%	100.0%
GLM Rand eff	100.0%	-4.7%	7.9%	93.1%	100.0%	-3.0%	7.0%	93.1%
Total costs								
Direct comparison	98,6%	-0,5%	8,3%	100,0%	98,2%	0,7%	8,2%	100,0%
Song	99,0%	-0,5%	6,6%	100,0%	98,6%	0,7%	6,6%	100,0%
Puhan	97,9%	-0,5%	6,4%	100,0%	97,7%	1,0%	6,5%	100,0%
GLM Fixed eff	99.6%	-0.8%	9.0%	100.0%	99.2%	0.9%	8.6%	100.0%
GLM Rand eff	100.0%	-0.7%	9.4%	94.3%	100.0%	1.0%	8.9%	96.4%

Table A5.3: Health-economic outcomes for three of the eight scenarios. Means of coverage, bias and mean absolute deviance (MAD) of the difference between two interventions, over 1,000 repetitions.

Scenario	Scenario 6				Scenario 7				Scenario 8			
	Six randomly drawn trials; one trial drawn from a less healthy population; one from a relatively younger population; one from a relatively older population				Six randomly drawn trials; three trials drawn from a less healthy population				Six randomly drawn trials; three trials drawn from a less healthy population (extreme scenario)			
	Coverage	Bias	MAD	Stat Power	Coverage	Bias	MAD	Stat Power	Coverage	Bias	MAD	Stat Power
Number of QALYs												
Direct comparison	97,4%	-6,0%	13,9%	99,6%	97,2%	-8,0%	14,3%	99,3%	98.9%	-11.9%	15.5%	99.3%
Song	98,1%	-3,9%	11,0%	100,0%	97,9%	-7,5%	12,0%	99,9%	99.6%	-11.2%	13.7%	99.5%
Puhan	96,5%	-2,8%	10,2%	100,0%	96,3%	-6,5%	11,2%	100,0%	90.7%	-9.8%	12.3%	100.0%
GLM Fixed eff	99.1%	-4.7%	14.6%	99.1%	99.4%	-6.6%	15.0%	97.4%	100.0%	-10.5%	18.3%	73.0%
GLM Rand eff	100.0%	-11.2%	17.3%	37.4%	100.0%	-13.5%	18.7%	32.2%	100.0%	-16.3%	21.4%	10.0%
Number of LYs												
Direct comparison	97,0%	-8,8%	23,0%	79,7%	97,4%	-10,1%	22,8%	79,3%	99.1%	-11.2%	22.5%	70.0%
Song	97,9%	-5,4%	18,5%	92,9%	98,5%	-9,7%	19,2%	91,9%	99.2%	-11.9%	19.9%	83.7%
Puhan	96,7%	-3,8%	17,2%	98,7%	96,9%	-8,1%	18,0%	98,0%	99.0%	-9.3%	18.0%	94.1%
GLM Fixed eff	98.9%	-6.4%	24.3%	63.9%	99.2%	-7.7%	24.1%	60.3%	100.0%	-9.2%	27.6%	13.0%
GLM Rand eff	100.0%	-16.0%	27.5%	4.7%	100.0%	-17.8%	28.5%	3.6%	100.0%	-17.5%	31.0%	0.5%
Number of Events												

Table A5.3: Health-economic outcomes for three of the eight scenarios. Means of coverage, bias and mean absolute deviance (MAD) of the difference between two interventions, over 1,000 repetitions. *(Continued)*

Scenario	Scenario 6				Scenario 7				Scenario 8			
	Six randomly drawn trials; one trial drawn from a less healthy population; one from a relatively younger population; one from a relatively older population				Six randomly drawn trials; three trials drawn from a less healthy population				Six randomly drawn trials; three trials drawn from a less healthy population (extreme scenario)			
	Coverage	Bias	MAD	Stat Power	Coverage	Bias	MAD	Stat Power	Coverage	Bias	MAD	Stat Power
Direct comparison	98,4%	-0,8%	6,1%	100.0%	98,8%	-3,1%	6,3%	100,0%	98.3%	-10.0%	10.6%	100.0%
Song	98,1%	-0,7%	5,4%	100.0%	98,4%	-1,9%	5,4%	100,0%	99.4%	-5.8%	7.3%	100.0%
Puhan	97,7%	-0,2%	5,1%	100.0%	97,6%	-1,5%	5,1%	100,0%	90.1%	-6.0%	7.2%	100.0%
GLM Fixed eff	99.0%	-0.8%	6.3%	100.0%	99.3%	-2.8%	6.8%	100.0%	99.8%	-8.4%	10.4%	100.0%
GLM Rand eff	100.0%	-3.3%	7.1%	89.9%	100.0%	-5.3%	7.9%	88.3%	100.0%	-11.7%	13.1%	73.4%
Total costs												
Direct comparison	98,3%	-1,0%	8,1%	100.0%	99,1%	0,2%	7,8%	100,0%	99.7%	3.1%	8.3%	100.0%
Song	98,9%	-0,3%	7,0%	100.0%	99,3%	-0,5%	6,6%	100,0%	100.0%	-0.4%	6.9%	100.0%
Puhan	97,1%	-0,3%	6,6%	100.0%	98,2%	-0,4%	6,4%	100,0%	91.6%	0.5%	6.4%	100.0%
GLM Fixed eff	99.5%	-1.0%	8.7%	100.0%	99.9%	0.0%	8.7%	100.0%	100.0%	1.7%	10.1%	99.8%
GLM Rand eff	100.0%	-1.1%	9.0%	92.6%	100.0%	-0.2%	8.9%	93.5%	100.0%	2.5%	10.7%	83.2%

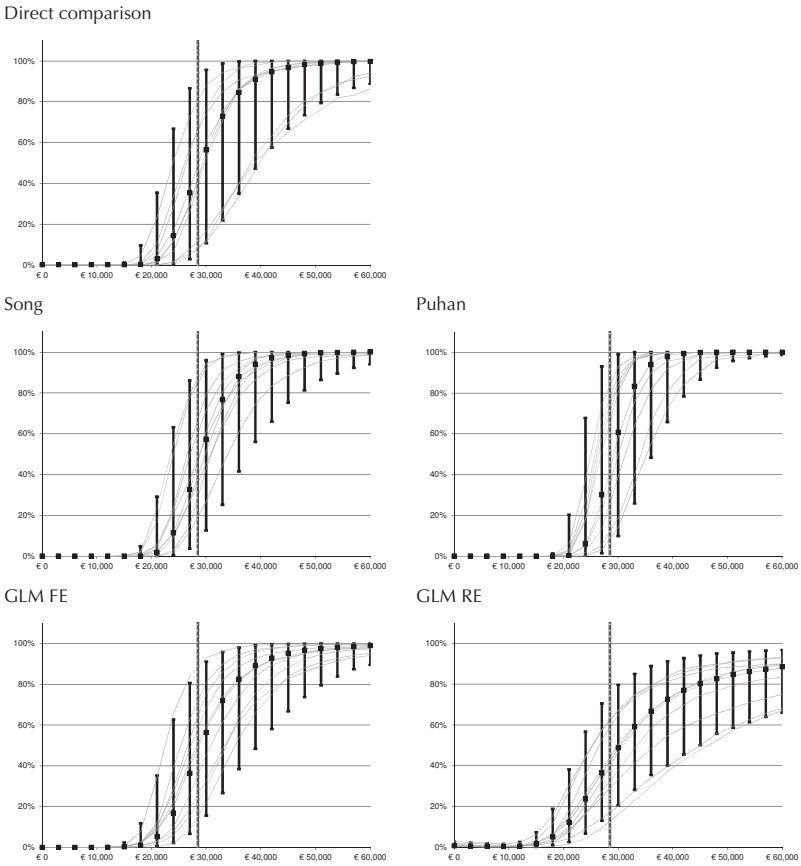

Figure A5.1: Cost-effectiveness acceptability curves (CEACs) for the five meta-analysis methods in the heterogeneous scenario 1. Graphs depicts median, 2.5th and 97.5th percentile CEACs over 1,000 repetitions, as well as the CEACs for the first 10 repetitions; vertical line is the 'true' population ICER.

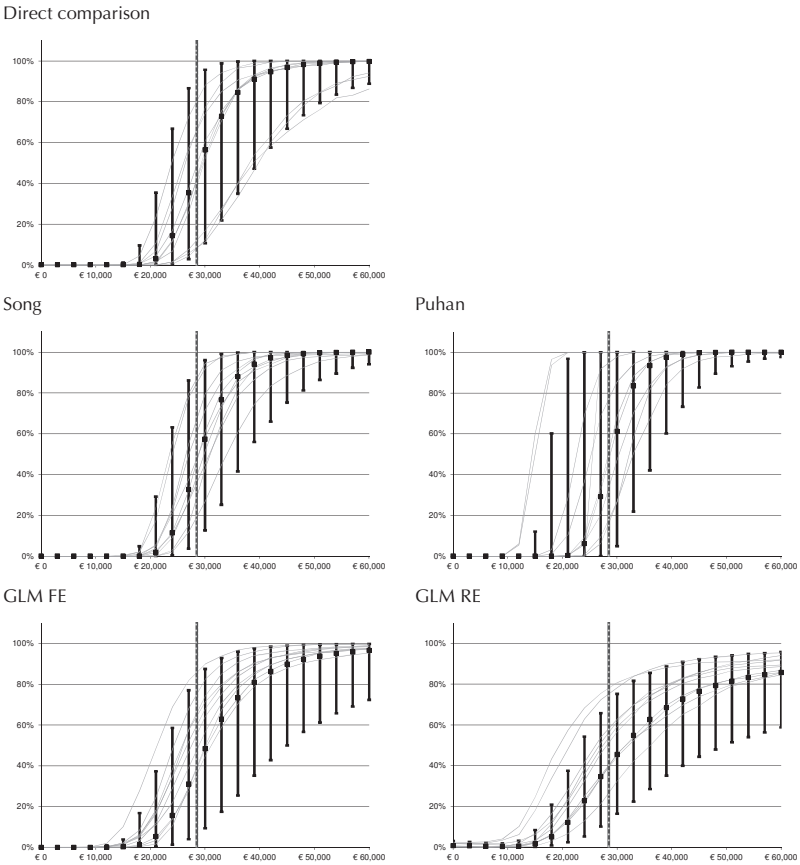

Figure A5.2: Cost-effectiveness acceptability curves (CEACs) for the five meta-analysis methods in the heterogeneous scenario 4. Graphs depicts median, 2.5th and 97.5th percentile CEACs over 1,000 repetitions, as well as the CEACs for the first 10 repetitions; vertical line is the 'true' population ICER.

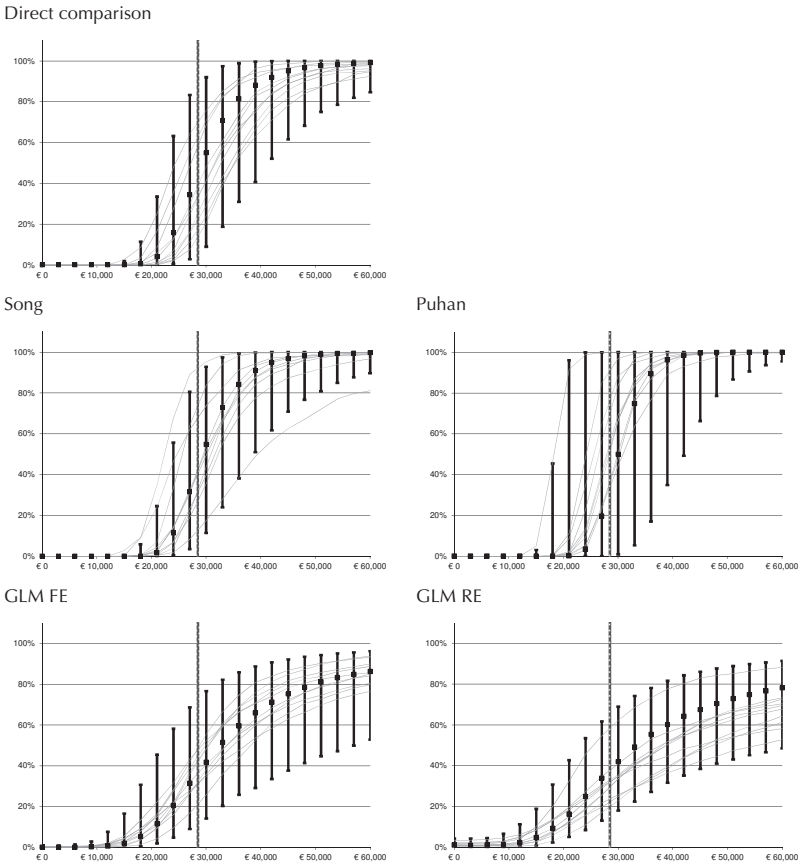

Figure A5.3: Cost-effectiveness acceptability curves (CEACs) for the five meta-analysis methods in the heterogeneous scenario 8. Graphs depicts median, 2.5th and 97.5th percentile CEACs over 1,000 repetitions, as well as the CEACs for the first 10 repetitions; vertical line is the 'true' population ICER.

Chapter 6

Crossing borders

Factors affecting differences in cost-effectiveness of smoking cessation interventions between European countries

P. Vemer, M.P.M.H. Rutten-van Mölken

Previously published in Value in Health, 2010, 13(2), 230–241
doi: 10.1111/j.1524-4733.2009.00612.x

ABSTRACT

Objectives Many different factors affect the transferability of cost-effectiveness results between countries. The objective is to quantify the impact of nine potential causes of variation in cost-effectiveness of pharmacological smoking cessation therapies (SCTs) between The Netherlands (reference case), Germany, Sweden, UK, Belgium, and France.

Methods The life-time benefits of smoking cessation were calculated using the Benefits of Smoking Cessation on Outcomes model, following a cohort of smokers making an unaided quit attempt, or using nicotine replacement therapy (NRT), bupropion, or varenicline. We investigated the impact of between-country differences in nine factors—demography, smoking prevalence, mortality, epidemiology and costs of smoking related diseases, resource use and unit costs of SCTs, utility weights and discount rates—on the incremental net monetary benefit (INMB), using a willingness-to-pay (WTP) of €20,000 per quality adjusted life year (QALY).

Results The INMB of 1000 quit attempts with NRT versus unaided, varies from €0.39 million (Germany) to €1.47 million (France). The differences between the countries were primarily due to differences in discount rates, causing the INMB to change between -65% to +62%, incidence and mortality rates (epidemiology) of smoking-related diseases (-43% to +35%) and utility weights. Impact also depended on the WTP for a QALY and time horizon: at a low WTP or a short time horizon, the resource use and unit costs of SCTs had the highest impact on INMB.

Conclusions Although all INMBs were positive, there were significant differences across countries. These were primarily related to choice of discount rate and epidemiology of diseases.

6.1 INTRODUCTION

An increasing number of regulatory agencies across the world require evidence on the cost-effectiveness of new pharmacotherapies. All these agencies need results that represent their own unique national or regional setting. Nevertheless, time and budget constraints limit the number of clinical trials and economic evaluations pharmaceutical companies can conduct in potential markets. In addition, there is increased acknowledgement of the limited external validity of country-specific cost-effectiveness data. In recognition of these difficulties, ISPOR initiated the Transferability of Economic Data Task Force. Their mission was to develop good research practices on the transferability of economic data in health technology assessment.[1]

The Task Force advocates the use of mathematical decision analytic models to assess setting-specific cost-effectiveness. These models synthesize and structure evidence from diverse sources, allow expanding the time horizon beyond that of a clinical trial, as well as adapting and transferring results from one setting to another.[2,3] For these reasons, models have been developed to assess the long-term cost-effectiveness of smoking cessation interventions.

A recent example is the BENESCO (Benefits of Smoking Cessation on Outcomes) model[4] which was developed by Heron Evidence Development Ltd, to support the launch of varenicline in various countries, e.g., The Netherlands[5], Sweden[6], Belgium [Annemans et al., unpubl. ms.], Germany[7], the UK[8], the Czech Republic[9], Korea[10], Japan[11], and Denmark.[12] Interesting differences in the cost-effectiveness of the various smoking cessation medications were observed[13,14], which may relate to various sources of variation, for example the incidence and prevalence of smoking and smoking-related diseases, characteristics of the population of smokers, differences in absolute and relative unit costs of medications and health-care services and many other factors.

This study was designed to unravel the factors driving differences in cost-effectiveness of pharmacological smoking cessation therapies (SCTs) between six European countries. The countries included were The Netherlands, Belgium, Germany, Sweden, the UK, and France, countries for which, at the start of the study, country-specific input data of the model were available.

6.2 METHODS

6.2.1 The model

The projections of the effects of smoking cessation were based on the BENESCO model[15], which is a probabilistic, updated, and improved version of the Health and Economic Consequences of Smoking model.[16] The BENESCO model simulates the consequences

of smoking and the benefits of quitting in terms of smoking-related morbidity, mortality, and associated medical costs in a population. The model is structured as a Markov model (cycle length 1 year) and follows a hypothetical cohort of current smokers making a single attempt to quit smoking at the beginning of the simulation. The cohort is followed from the time of their quit attempt until all members of the cohort have died. Individuals are classified into one of three smoking states, i.e., smoker, recent quitter (abstinent 1 to 5 years after successful quit attempt), or long-term quitter. Transition probabilities between smoking states in the first year depend on cessation rates of the interventions, while the probabilities after 1 year depend on relapse rates, which in turn depend on time since quitting. The model simulates the age, gender, and smoking status-specific incidence and mortality of four major diseases for which smoking is a well-established risk factor: chronic obstructive pulmonary disease (COPD), lung cancer, coronary heart disease (CHD), and stroke. Smoking state-specific incidence and mortality rates were calculated using relative risks.[17,18] The incidence and mortality rates for recent quitters were calculated using the relative risks of former smokers versus nonsmokers, while the rates for long term quitters were assumed to be the same as those of never smokers. Because COPD and lung cancer are chronic progressive conditions, these diseases were given hierarchical prominence over the other conditions with acute recurrent events. This means that individuals with COPD or lung cancer remain in this state until they die and cannot move to a CHD or stroke state, whereas individuals with CHD or stroke can move to the COPD or lung cancer state. As in all Markov models, states are mutually exclusive, which means that a patient cannot have two diseases at the same time. The model calculates the total number of smokers and quitters that have one of the smoking-related diseases as well as the number of deaths (due to one of the smoking-related diseases and overall) over the time horizon of the simulation. Based on these numbers, the total health-care costs associated with the different disease states and the total number of (quality adjusted) life years is calculated. The model uses three age bands: 18 to 34 years, 35 to 64 years, and 65 years and older. Subjects alive in the model at age 99 years are all assumed to die in the next cycle. It is assumed that there is no smoking-related morbidity or mortality in the 18 to 34 years age class.

6.2.2 Smoking cessation therapies

We calculate the cost-effectiveness of three frequently used pharmacological SCTs—nicotine replacement therapy (NRT), bupropion, and varenicline—and unaided cessation. NRT is the generic term for any form of smoking cessation aid which delivers a measured dose of nicotine to the person using it. Examples include the nicotine patch or nicotine gum. Bupropion is an antidepressant used to support smoking cessation.[19] Varenicline is designed to relieve symptoms of nicotine withdrawal including cigarette craving and block the reinforcing effects of continued nicotine use.[20] The 12-month continuous ab-

stinence rates were based on a meta-analysis of available randomized controlled trials, where the SCTs were always given in combination with counselling.[5] They were 5.0% for unaided cessation, 14.8% for NRT, 17.0% for bupropion, and 22.4% for varenicline. In all analyses, we assumed that 25% of smokers undertake a single quit attempt, using one of the smoking cessation interventions, or unaided. It is this cohort that is followed over lifetime.

6.2.3 Factors affecting transferability

A total of nine factors that could potentially cause differences in cost-effectiveness between countries were investigated. Each factor consists of a group of country-specific input parameters which are varied simultaneously. Table 6.1 gives the most important input parameters of each of the nine factors. The nine country-specific factors include:

F1: Demography. This includes the total number of people older than 18 years of age and the break-downs of the population by gender and age-classes.

F2: Smoking Prevalence. This refers to the percentage of smokers, nonsmokers, and former smokers in each age/gender class.

Table 6.1: Main country-specific input parameters for each factor potentially contributing to between-country variation in cost-effectiveness of smoking cessation interventions.[a]

	The Netherlands	Belgium	Germany	Sweden	United Kingdom	France
Population characteristics (age 18+, x mln)						
Population size	12.7	8.2	67.1	7.3	46.6	46.8
Number of smokers	3.54	2.25	18.61	1.51	12.70	11.53
As % of adult population	28%	27%	28%	21%	27%	25%
Cohort size: smokers making a quit attempt	0.88	0.56	4.65	0.38	3.17	2.88
F1: Demography						
Males, 18 to 34 years	14.1%	13.9%	13.6%	15.4%	14.4%	14.6%
Males, 35 to 64 years	27.3%	25.5%	26.5%	24.7%	25.4%	24.6%
Males, 65+ years	7.6%	8.9%	7.9%	9.0%	8.7%	8.6%
Females, 18 to 34 years	13.8%	13.7%	13.5%	14.9%	14.3%	14.4%
Females, 35 to 64 years	26.8%	25.3%	25.8%	24.0%	26.1%	25.3%
Females, 65+ years	10.4%	12.7%	12.6%	12.0%	11.1%	12.4%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Source	21	22	23	24	25	26

Table 6.1: Main country-specific input parameters for each factor potentially contributing to between-country variation in cost-effectiveness of smoking cessation interventions.[a] *(Continued)*

	The Netherlands	Belgium	Germany	Sweden	United Kingdom	France
Population characteristics (age 18+, x mln)						
Population size	12.7	8.2	67.1	7.3	46.6	46.8
Number of smokers	3.54	2.25	18.61	1.51	12.70	11.53
As % of adult population	28%	27%	28%	21%	27%	25%
Cohort size: smokers making a quit attempt	0.88	0.56	4.65	0.38	3.17	2.88
F2: Smoking prevalence						
Males, 18 to 34 years	32.3%	34.6%	38.7%	15.0%	32.6%	39.4%
Males, 35 to 64 years	34.1%	35.4%	36.0%	22.4%	27.7%	29.3%
Males, 65+ years	15.6%	19.2%	13.3%	15.4%	12.7%	10.2%
Females, 18 to 34 years	27.4%	26.1%	29.7%	23.0%	28.0%	31.2%
Females, 35 to 64 years	28.8%	27.8%	27.2%	26.8%	28.5%	21.8%
Females, 65+ years	12.3%	8.9%	6.4%	12.8%	26.7%	6.2%
Source	27	28	29	24	25	30,31
F3: All-cause mortality						
Males, 18 to 34 years	0.06%	0.12%	0.08%	0.07%	0.09%	0.09%
Males, 35 to 64 years	0.40%	0.54%	0.58%	0.39%	0.47%	0.55%
Males, 65+ years	5.48%	4.89%	4.14%	4.88%	4.72%	
Females, 18 to 34 years	0.03%	0.04%	0.03%	0.03%	0.04%	0.04%
Females, 35 to 64 years	0.28%	0.29%	0.29%	0.25%	0.30%	0.25%
Females, 65+ years	4.57%	4.48%	4.67%	2.85%	3.87%	3.65%
Source	21	22	23	24	32	26
F4: Epidemiology: annual incidence rate of COPD per 1,000 inhabitants						
Males, 18 to 34 years	0.17	0.16	0	0.02	0	0.03
Males, 35 to 64 years	2.08	2.31	0.1	0.74	0.15	0.17
Males, 65+ years	9.68	12.77	3.26	15.22	3.82	1.73
Females, 18 to 34 years	0.17	0.17	0	0.02	0	0.01
Females, 35 to 64 years	2.13	2.63	0.06	0.97	0.11	0.09
Females, 65+ years	6.46	12.17	2.25	9.58	1.95	2.05
Source	33	33	29	34	35,36	37–42

Table 6.1: Main country-specific input parameters for each factor potentially contributing to between-country variation in cost-effectiveness of smoking cessation interventions.[a] *(Continued)*

	The Netherlands	Belgium	Germany	Sweden	United Kingdom	France
Population characteristics (age 18+, x mln)						
Population size	12.7	8.2	67.1	7.3	46.6	46.8
Number of smokers	3.54	2.25	18.61	1.51	12.70	11.53
As % of adult population	28%	27%	28%	21%	27%	25%
Cohort size: smokers making a quit attempt	0.88	0.56	4.65	0.38	3.17	2.88
F4: Epidemiology: annual incidence rate of lung cancer per 1,000 inhabitants						
Males, 18 to 34 years	0.01	0.01	0.01	0.01	0	0.01
Males, 35 to 64 years	0.56	0.61	0.69	0.69	0.56	1
Males, 65+ years	4.24	5.15	5.38	4.95	4.84	4.25
Females, 18 to 34 years	0.01	0.01	0	0.01	0.01	0.01
Females, 35 to 64 years	0.42	0.48	0.39	0.56	0.36	0.2
Females, 65+ years	1.07	1.71	1.43	1.88	1.63	0.82
Source	43	43	29,44–47	34	25	48
F4: Epidemiology: annual incidence rate of CHD per 1,000 inhabitants, all events						
Males, 18 to 34 years	0.35	0.33	0.8	0.05	0.04	0
Males, 35 to 64 years	7.6	7.95	8.36	6.97	2.32	4.27
Males, 65+ years	24.35	26.26	33.15	33.84	25.58	33.42
Females, 18 to 34 years	0.06	0.05	0.6	0.02	0.01	0
Females, 35 to 64 years	2.37	2.54	5.29	2.32	0.53	1.21
Females, 65+ years	15.29	17.44	25.76	20.84	15.48	15.45
Source	49	49	29,50	34	25,51	52
F4: Epidemiology: annual incidence rate of CHD per 1,000 inhabitants, first event only						
Males, 18 to 34 years	0.33	0.32	0.46	0.04	0.04	0
Males, 35 to 64 years	5.69	5.95	4.86	4.39	1.6	3.8
Males, 65+ years	17.5	18.87	19.25	17.24	14.93	26.27
Females, 18 to 34 years	0.06	0.05	0.35	0.02	0.01	0
Females, 35 to 64 years	1.81	1.94	3.07	1.57	0.46	1.12
Females, 65+ years	11.42	13.04	14.96	12.11	11.27	13.42
Source	49	49	29,50	34	25,51,53	52

Table 6.1: Main country-specific input parameters for each factor potentially contributing to between-country variation in cost-effectiveness of smoking cessation interventions.[a] *(Continued)*

	The Netherlands	Belgium	Germany	Sweden	United Kingdom	France
Population characteristics (age 18+, x mln)						
Population size	12.7	8.2	67.1	7.3	46.6	46.8
Number of smokers	3.54	2.25	18.61	1.51	12.70	11.53
As % of adult population	28%	27%	28%	21%	27%	25%
Cohort size: smokers making a quit attempt	0.88	0.56	4.65	0.38	3.17	2.88
F4: Epidemiology: annual incidence rate of stroke per 1,000 inhabitants, all stroke events						
Males, 18 to 34 years	0.03	0.03	0.12	0.06	0.08	0.14
Males, 35 to 64 years	1.31	1.39	1.89	1.69	2.29	0.79
Males, 65+ years	11.69	12.61	17.8	14.67	13.25	9.07
Females, 18 to 34 years	0.1	0.1	0.03	0.05	0.07	0.13
Females, 35 to 64 years	0.88	1	1.34	0.93	1.58	0.73
Females, 65+ years	11.43	12.71	11.95	13.23	12.28	6.19
Source	54	54	29	34	25	55
F4: Epidemiology: annual incidence rate of stroke per 1,000 inhabitants, first event only						
Males, 18 to 34 years	0.03	0.03	0.06	0.05	0.06	0
Males, 35 to 64 years	1.19	1.25	0.99	1.44	1.73	0.43
Males, 65+ years	10.53	11.36	9.29	11.29	8.74	8.15
Females, 18 to 34 years	0.1	0.1	0.02	0.04	0.05	0
Females, 35 to 64 years	0.8	0.91	1.07	0.8	1.17	0.43
Females, 65+ years	10.52	11.7	9.51	10.66	8.03	5.07
Source	54	54	29	34	25	55
F5:Annual costs[b] per patient with a smoking-related disease						
COPD	1,036	1,928	2,245	2,907	1,127	2,220
Lung Cancer						
First year	13,236	13,505	33,983	10,355	5,132	17,629
After first year	13,236	13,505	33,983	5,502	5,132	17,629
CHD						
First year	4,841	4,867	1,969	4,795	1,348	5,721
After first year	2,949	796	985	1,374	1,348	5,721
Stroke						
First year	23,119	7,685	10,741	7,056	22,006	9,641
After first year	5,229	5,439	4,618	1,884	22,006	9,641
Source	56–59	57,60–63	29,64	65,66[c]	67–70	71

Table 6.1: Main country-specific input parameters for each factor potentially contributing to between-country variation in cost-effectiveness of smoking cessation interventions.[a] *(Continued)*

	The Netherlands	Belgium	Germany	Sweden	United Kingdom	France
Population characteristics (age 18+, x mln)						
Population size	12.7	8.2	67.1	7.3	46.6	46.8
Number of smokers	3.54	2.25	18.61	1.51	12.70	11.53
As % of adult population	28%	27%	28%	21%	27%	25%
Cohort size: smokers making a quit attempt	0.88	0.56	4.65	0.38	3.17	2.88
F6: Resource use: intervention costs using different resource use across countries, but equal unit costsb						
Varenicline	391.79	304.79	294.59	391.79	381.59	294.59
Bupropion	327.81	244.81	226.60	335.51	160.30	230.61
NRT	323.35	213.94	207.00	298.79	234.56	231.05
Unaided cessation	0.00	0.00	0.00	0.00	0.00	0.00
Source	20,72	73,74	75	19,76,77	78	
F7: Unit costs: intervention costs using different unit costs across countries, but equal resource useb						
Varenicline	391.79	391.78	337.28	401.90	290.62	390.60
Bupropion	327.81	277.42	292.22	350.92	285.02	327.15
NRT	323.35	311.05	317.13	365.76	213.15	387.03
Unaided cessation	0.00	0.00	0.00	0.00	0.00	0.00
Source	79,80	73,74	75	19,76,77	81	
F8: General population utility weights						
Males, 18 to 34 years	0.91	0.91	0.93	0.93	0.93	0.93
Males, 35 to 64 years	0.91	0.91	0.877	0.877	0.877	0.877
Males, 65+ years	0.82	0.82	0.8	0.8	0.8	0.8
Females, 18 to 34 years	0.92	0.92	0.91	0.91	0.91	0.91
Females, 35 to 64 years	0.89	0.89	0.853	0.853	0.853	0.853
Females, 65+ years	0.76	0.76	0.77	0.77	0.77	0.77
Source	82–90					
F8: Disease-specific utility weights						
COPD	0.69	0.76	0.76	0.76	0.76	0.76

Table 6.1: Main country-specific input parameters for each factor potentially contributing to between-country variation in cost-effectiveness of smoking cessation interventions.[a] *(Continued)*

	The Netherlands	Belgium	Germany	Sweden	United Kingdom	France
Population characteristics (age 18+, x mln)						
Population size	12.7	8.2	67.1	7.3	46.6	46.8
Number of smokers	3.54	2.25	18.61	1.51	12.70	11.53
As % of adult population	28%	27%	28%	21%	27%	25%
Cohort size: smokers making a quit attempt	0.88	0.56	4.65	0.38	3.17	2.88
Lung Cancer first year	0.61	0.61	0.61	0.61	0.61	0.61
Following years	0.5	0.5	0.5	0.5	0.5	0.5
CHD	0.71	0.76	0.76	0.76	0.76	0.76
Stroke first year	0.54	0.74	0.74	0.74	0.74	0.74
Following years	0.29	0.15	0.15	0.15	0.15	0.15
Source	83–93					
F9: Discount rates						
Costs	4.00%	3.00%	5.00%	3.00%	3.50%	3.00%
Outcomes	1.50%	1.50%	5.00%	3.00%	3.50%	0.00%
Source	94	95	96	97	98	99

[a] COPD: chronic obstructive pulmonary disease; CHD, coronary heart disease; NRT, nicotine replacement therapy

[b] In 2006 Euros, accounting for differences in purchasing power

[c] Bolin K, Dozet A, unpublished data.

F3: All-cause mortality. Mortality in the general population is expressed as the all-cause mortality rate, which is the percentage of the total number of people in each age/gender class that dies during a single year.

F4: Epidemiology of smoking-related diseases. The epidemiology of smoking-related diseases consists of three elements: the incidence rates, prevalence rates, and annual cause-specific mortality rates by age/gender class. We applied the disease definitions that were actually used in each country at the time of writing the reimbursement dossiers for varenicline. To identify COPD, all countries used ICD-10 codes J40–44, UK and Sweden also used J47. To identify lung cancer, all countries used C33–34, except Sweden that defined lung cancer as C34. CHD is identified in all countries as I20–25. Stroke in The Netherlands and Belgium is identified as I60–I69 plus G45, in Sweden as I61 and I63, in Germany as I60, I61, I63 and I64, and in the UK and France as I60–I64.

Given the causal relationship, there is a strong association between smoking prevalence and the epidemiology of smoking related diseases. To enter these two factors as independent factors in the univariate analysis, we calculated the country-specific incidence, prevalence, and mortality of smoking-related diseases among nonsmokers, i.e., the country-specific baseline risk. This was done using the country-specific epidemiology and smoking prevalence, and the relative risks for smokers, former smokers, and nonsmokers used within the model. When studying the impact of the factor "epidemiology" in the univariate analysis, the Dutch baseline-risk was replaced by the country-specific baseline risk, which was then combined with the relative risks and the Dutch smoking prevalence to estimate the incidence (or prevalence or mortality) of smoking-related diseases among smokers and ex-smokers.

F5: Costs of smoking-related diseases. The model makes a distinction between the first-year costs and subsequent-year costs for lung cancer, CHD, and stroke, diseases for which high initial costs are generally followed by lower maintenance costs. As COPD does not have (much) higher initial costs, this distinction is not relevant for COPD.

F6: Resource use and **F7: Unit costs of SCTs.** The intervention costs of SCTs are separated into two components: the amount of resource use (i.e., medication and counselling) associated with the SCTs and the unit costs of these resources. We have investigated both these factors separately.

F8: Utility weights. The BENESCO model requires two categories of utility inputs: utility weights for the general, disease-free (developed no smoking-related disease) population, which vary by age, and the disease-specific utility weights, which vary by type of smoking-related disease. The Netherlands is the only country in our sample for which country-specific utility weights for both categories were provided. Germany, Sweden, the UK, and France all have used the provided default values within the model. Belgium has used the general population utility weights from The Netherlands and the default disease-specific utility values.

F9: Discount rates. All costs and outcomes are discounted using the country-specific values that are recommended in the national guidelines for economic evaluations. In the reference case, costs are discounted at 4%, outcomes at 1.5%.

We adopted a health-care perspective and included healthcare costs that are either covered from the health-care budgets or paid for by patients. All prices and costs were inflated to 2006, using the Harmonised Indices of Consumer Prices—all items.[100] We also compensated for differences in purchasing power, using the average exchange rates on January 2, July 3, and December 31, 2006, and 2006 purchasing power parities.[100]

6.2.4 Analyses

The starting point of all between-country comparisons were the results of the BENESCO model populated with Dutch input data. Hence, The Netherlands was the reference case.

In a series of univariate analyses we replaced the group of input parameters belonging to the same factor by its country-specific estimates. We changed one factor at a time; all other factors were kept constant at the reference values. We compared the impact of each factor on the outcomes. In the subsequent multivariate analysis, we consecutively enter parameters from the highest to the lowest impact. Eventually, this results in models that are filled completely with country-specific parameters. In all analyses, the time horizon is lifetime. Sensitivity analyses were done using different time horizons and different threshold values of the willingness-to-pay (WTP) for a quality adjusted life year (QALY).

6.2.5 Outcomes

Outcomes were presented as incremental costs, QALYs gained, incremental cost-effectiveness ratios (ICERs), and incremental net monetary benefits (INMBs). The ICER is the difference in total costs between two smoking cessation interventions, divided by the difference in total QALYs. The percentage change in the INMB of the reference case caused by each factor was our primary measure of interest. The INMB was calculated as the difference in QALYs between two interventions, times societies' WTP, for a QALY (threshold value) minus the difference in costs. The INMB was calculated with a relatively low threshold value of €20,000 per QALY. For each country,we have ranked all country-specific input parameters according to the percentage of change in INMB compared with the reference case. A rank order of 1 indicates that this factor caused the INMB to change most; a rank order of 9 indicates that this factor had the least impact on the INMB. We present the outcomes according to a hierarchy of effectiveness of the interventions: NRT versus unaided cessation, bupropion versus NRT, and varenicline versus bupropion. The ranking was averaged over these three pair-wise comparisons.

6.3 RESULTS

6.3.1 Reference case

In table 6.2, the outcomes of the reference case are given. The ICER of NRT compared with unaided cessation was about €1,600 per QALY. Bupropion dominated NRT, and varenicline dominated bupropion. Using a WTP of €20,000 per QALY, the INMBs of all three comparisons were positive.

6.3.2 Univariate analysis

F1: Demography. Demography influences the age and gender distribution of the cohort of smokers that is followed over lifetime. If the cohort of smokers that attempts to quit becomes older than in the reference case, the INMB and ICER worsen, for all three pair-wise comparisons of smoking-cessation interventions. This is primarily due to a decrease

Table 6.2: Lifetime outcomes of the Benefits of Smoking Cessation on Outcomes model filled with Dutch input data, expressed per 1000 smokers making a quit attempt.[a]

	NRT versus unaided cessation	Bupropion versus NRT	Varenicline versus bupropion
Difference in total costs (€1000)[b]	126.7	-40.4	-44.7
Difference in QALYs	77.4	17.7	42.8
Incremental net monetary benefit (INMB) (€1 mln)[c]	1.42	0.39	0.90
Incremental cost-effectiveness ratio (ICER)	1636.7	Dominant	Dominant

[a] Outcomes differ from [5] because all cost inputs were updated to 2006 prices, Harmonised Indices of Consumer Prices were used, asthma exacerbations were excluded and the price of varenicline was updated. NRT: nicotine replacement therapy; QALYs: quality adjusted life years; WTP: willingness-to-pay.

[b] Intervention costs plus total costs of smoking-related diseases.

[c] WTP is €20,000.

Table 6.3: Effect of changing the reference case input values to the country-specific input values on cost-effectiveness outcomes compared with the reference case.[a]

	Effect on INMB and ICER compared with the reference case
F1: Demography	
Older cohort	Worsens
F2: Smoking prevalence	
Higher smoking prevalence among elderly	Worsens
F3: All-cause mortality	
Lower mortality	Improves
F4: Smoking-related disease epidemiology	
Higher incidence	Improves
Higher mortality	Improves
F5: Costs of smoking-related diseases	
Higher costs	Improves
F6: Resources used for SCTs	
Resource use of a more effective SCT increases more than the resource use of a less effective SCT	Worsens
F7: Unit costs of SCTs	
Unit cost of a more effective SCT increases more than the unit costs a less effective SCT	Worsens
F8: Utility weights	
Higher disease-specific utility weights	Worsens
Lower general population utility weights	Worsens
F9: Discount rates	
Higher discount rates on costs	Worsens
Higher discount rates on outcomes	Worsens

[a] INMB: incremental net monetary benefit; ICER: incremental cost-effectiveness ratio; SCTs: smoking cessation therapies.

in QALYs that is greater for the more effective intervention because the number of people who remain disease free and survive to old age is greater for this intervention. Table 6.3 summarizes these effects. Replacing the age and gender distribution in the reference case by the country-specific age and gender distribution caused the INMB of NRT versus unaided cessation to change between -2.1% in Belgium and -0.4% in Sweden. The change in INMB of bupropion versus NRT varies from -1.7% in Belgium to -0.4% in Sweden. The change in INMB of varenicline versus bupropion varies from -1.8% in Belgium to -0.4% in Sweden (figure 6.1).

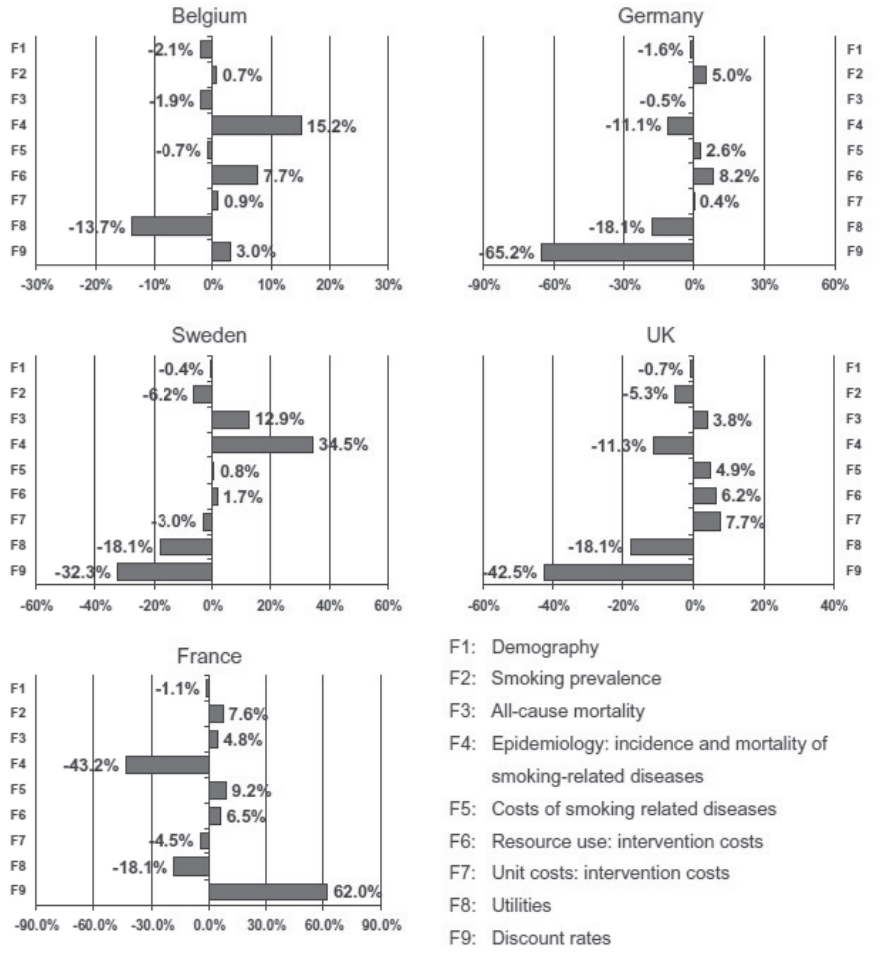

Figure 6.1: Relative change in incremental net monetary benefits of nicotine replacement therapy versus unaided cessation of the reference case, caused by applying each country-specific factor univariately.

F2: Smoking prevalence. Like demography, smoking prevalence primarily influences the age and gender distribution of the cohort of smokers attempting to quit. When smoking prevalence among the elderly gets higher, the cohort of smokers attempting to quit becomes older and the INMB and the ICER worsen. Compared with the reference case, the change in INMB for NRT versus unaided cessation that is due to a change towards country-specific smoking prevalence varies from -6.2% in Sweden to +7.6% in France. The change in INMB for bupropion versus NRT varies from -5.1% in Sweden to +6.3% in France. The change in INMB for varenicline versus bupropion varies from -5.4% in Sweden to +6.6% in France (figure 6.1).

F3: All-cause mortality. In countries where the all-cause mortality rate is lower (i.e., the life expectancy is higher) than in the reference case, the INMBs of the smoking cessation interventions are higher and the ICERs improve, primarily because of higher QALY gains. The increase in QALYs that result from a lower mortality rate is largest for the most effective treatment, because the number of people who stop smoking and remain disease free is highest for this intervention. The change in INMB for NRT versus unaided cessation because of a change in all-cause mortality varies from -1.9% in Belgium to +12.9% in Sweden. The change in INMB for bupropion versus NRT varies from -1.6% in Belgium to +10.6% in Sweden. The change in INMB for varenicline versus bupropion varies from -1.7% in Belgium to +11.3% in Sweden (figure 6.1).

F4: Epidemiology of smoking-related diseases. In countries where the incidence of all smoking-related diseases is higher than in the reference case, the INMBs and ICERs improve, because preventing more diseases results in higher QALY gains and greater cost savings. The same holds for countries where the mortality due to smoking-related diseases is higher. Change in prevalence has only a limited effect on the INMB. The change in INMB for NRT versus unaided cessation because of a change in epidemiology varies from -43.2% in France to +34.5% in Sweden. The change in INMB for bupropion versus NRT varies from -35.6% in France to +28.5% in Sweden. The change in INMB for varenicline versus bupropion varies from -37.7% in France to +30.1% in Sweden (Figure 6.1).

F5: Costs of smoking-related diseases. In countries where the health-care costs of smoking-related diseases are higher than in the reference case, the INMBs are higher and the ICERs improve, because the savings from preventing these diseases increase. This increase gets greater when the effectiveness of the smoking cessation intervention improves. The change in INMB for NRT versus unaided cessation because of a change in costs per patient with a smoking-related disease varies from -0.7% in Belgium to +9.2% in France. The change in INMB for bupropion versus NRT varies from -0.6% in Belgium to +7.6% in France. The change in INMB for varenicline versus bupropion varies from -0.6% in Belgium to +8.0% in France (figure 6.1).

F6: Resource use and **F7: Unit costs of SCTs.** When the intervention costs of a more effective SCT increase relatively more than the intervention costs of a less effective treat-

ment, the INMB will go down and the cost-effectiveness will worsen. The change in INMB for NRT versus unaided cessation because of a change in the resource use component of the intervention costs varies from +1.7% in Sweden to +8.2% in Germany. The change in INMB for bupropion versus NRT varies from -8.2% in Sweden to +20.0% in the UK. The change in INMB for varenicline versus bupropion varies from -17.5% in the UK to +0.9% in Sweden. The change in INMB for NRT versus unaided cessation because of a change in the unit cost component of the intervention costs varies from -4.5% in France to +7.7% in the UK. The change in INMB for bupropion versus NRT varies from -17.1% in the UK to +16.3% in France. The change in INMB for varenicline versus bupropion varies from -5.6% in Belgium to +6.5% in the UK (figure 6.1).

F8: Utilities. In countries where the utility weights of the smoking-related diseases are higher than in the reference case, the QALY gains from preventing these diseases are lower. The reduction in QALY gain is greatest for the intervention with the highest effectiveness. Thus, higher disease-specific utility weights lead to lower INMBs and a worsening of the cost-effectiveness. This applies to all five countries in our analysis, because the reference case represents the only country that has changed the model's default utility values. If the utility weights for the general, disease-free population of a country are lower than in the reference case, the QALY gains from preventing a smoking-related disease are lower. Again, the reduction in QALY gains is greater if the treatment is more effective because more people stay disease free and their live years are thus weighted with the lower utility weights. This causes the INMBs to go down and the ICERs to worsen. The change in INMB for NRT versus unaided cessation because of a change in utility weights varies from -18.1% in Germany, Sweden, the UK, and France, to -13.7% in Belgium. The change in INMB for bupropion versus NRT varies from -14.9% in Germany, Sweden, the UK and France, to -11.3% in Belgium. The change in INMB for varenicline versus bupropion varies from -15.8% in Germany, Sweden, the UK, and France, to -12.0% in Belgium (figure 6.1).

F9: Discount rates. In countries where the costs and outcomes are discounted more than in the reference case, the INMBs and the ICERs worsen because the cost savings and QALY gains of smoking cessation that occur far into the future are reduced. The change in INMB for NRT versus unaided cessation because of a change towards country-specific discount rates varies from -65.2% in Germany to +62.0% in Sweden. The change in INMB for bupropion versus NRT varies from -53.7% in Germany to +51.1% in Sweden. The change in INMB for varenicline versus bupropion varies from -56.9% in Germany to +54.1% in Sweden (figure 6.1).

6.3.3 Ranking of impact on INMB

The ranking of factors according to their impact on the INMB of NRT versus unaided cessation is largely similar for the comparisons bupropion versus NRT and varenicline versus bupropion. Table 6.4 shows the rank order when averaged over all three pairwise treatment comparisons. The first row shows the rank orders after averaging the impact of each factor over all countries. When substituting the reference case input univariately by country-specific input, F9: discount rates had the biggest impact on the cost-effectiveness. This is followed by F4: epidemiology and F8: utility weights. The least important factor in

Table 6.4: Univariate ranking of factors according to the percentage change in incremental net monetary benefit, averaged over the three pair-wise comparisons of smoking cessation therapies, using a threshold value €20,000 per quality adjusted life year and a lifetime horizon, unless otherwise stated.

	F9 Discount rates	F4 Epidemi-ology	F8 Utility weights	F6 Resource use	F7 Unit costs	F2 Smoking prevalence	F3 All-cause mortality	F5 Costs of smoking-related diseases	F1 Demog-raphy
Rank order averaged over all countries	1	2	3	4	5	6	7	8	9
Rank order for each country									
Belgium	5	1	2	4	3	8	7	9	6
Germany	1	3	2	5	6	4	9	7	8
Sweden	2	1	3	6	7	5	4	8	9
United Kingdom	1	5	2	3	4	6	8	7	9
France	1	2	3	8	5	6	7	4	9
Rank order averaged over all countries at different time horizons									
2 years	7	4	5	2	1	6	9	3	8
10 years	5	3	4	1	2	7	9	6	8
Lifetime	1	2	3	4	5	6	7	8	9
Rank order averaged over all countries at different threshold values									
€100	4	5	9	1	2	6	7	3	8
€500	4	5	7	1	2	6	8	3	9
€1,000	4	5	6	1	2	7	8	3	9
€5,000	1	2	5	3	4	7	8	6	9
€10,000	1	2	3	4	5	7	8	6	9
€20,000	1	2	3	4	5	6	7	8	9
€50,000	1	2	3	6	7	4	5	8	9
€100,000	1	2	3	7	6	5	4	9	8

terms of its effect on the INMB is F1: demography, i.e., the age/gender distribution of the cohort of smokers making a quit attempt.

6.3.4 Sensitivity analysis: impact of using a different time horizon

A shorter time horizon changes the importance of the various causes of variability in cost-effectiveness between countries. The importance of the three factors with the largest long-term mpact, i.e., F9: discount rates, F8: utility weights, and F3: all-cause mortality decreases. Using a time span of 2 or 10 years, the two most important factors become the two factors determining the costs of smoking cessation treatment: F6: resources used and F7: unit costs. These factors become so important because a time horizon of 2 and 10 years is insufficient to capture the full gains in QALYs and savings in costs that result from the prevention of smoking-related diseases. In other words, time has been insufficient to fully get the returns on the investments in SCT.

6.3.5 Sensitivity analysis: impact of using a different threshold value

Using different threshold values to calculate the NMB also changes the rank order of the factors. When the threshold value increases, the factors with a large influence on the QALYs, i.e. all-cause mortality, smoking prevalence, and demography, become more important. When the threshold value decreases, factors with a large influence on costs become more important. This includes resources used for the smoking cessation treatments, unit costs, and costs of smoking-related diseases. The discount rates remain important, irrespective of the threshold value. For threshold values of €5,000 or higher, discounting is the single most important factor. Using a threshold value of €1,000 or lower, the discount rate becomes the fourth most important factor. For a threshold value of €1,000 or lower, the most important factor is the resources used to deliver SCT, followed by costs of smoking-related disease.

6.3.6 Multivariate analysis

In the multivariate analysis, we enter all country-specific input parameters at the same time. Table 6.5 shows how the INMB differs between countries when fully accounting for all known between-country differences. In Belgium, the decrease in INMB due to higher disease-specific utilities is offset by an increase in the INMB because of a higher incidence of all smoking-related diseases. As a result, the INMBs increase, except for varenicline versus buproprion because the difference in unit costs between the two SCTs is greater than in The Netherlands.

In Germany, the INMBs of all three pair-wise comparisons decrease primarily because of the relatively high discount rate for costs and outcomes. Other causes are a lower incidence of COPD, higher disease-specific utility values and lower general population utility values.

Table 6.5: Incremental Net Monetary Benefit (INMB) per 1000 smokers undertaking a quit attempt, using a threshold of €20,000 per quality adjusted life year, for three pair-wise smoking cessation therapy comparisons, influenced by all nine identified factors.[a]

	INMB (x€1 mln)		
	NRT versus unaided cessation	Bupropion versus NRT	Varenicline versus bupropion
The Netherlands	1.42	0.39	0.90
Belgium	1.46	0.45	0.86
Germany	0.39	0.17	0.32
Sweden	1.29	0.22	0.82
UK	0.95	0.25	0.54
France	1.47	0.64	1.01

[a] NRT, nicotine replacement therapy.

In Sweden, the INMB is also lower than in the reference case for all pair-wise comparisons, because of higher utility weights for the smoking-related diseases and because QALYs were discounted at 3.5% instead of 1.5%. This decrease offsets the increase in INMB caused by lower all-cause mortality rates and higher incidence rates for all smoking-related diseases in most age/gender classes.

In the UK, the INMB of all three pair-wise comparisons is lower than in the reference case, primarily because of a higher discount rate (3%) for outcomes, higher utility weights for the smoking-related diseases and a lower incidence of COPD.

In France, lower smoking-related disease mortality, higher disease-specific utility values, and lower general population utility values cause the INMB of all three pair-wise comparisons to decrease. Nevertheless, the effect of no discounting (0%) on outcomes has such a large effect that the INMB is higher than in the reference case.

Figure 6.2 shows the differences between countries in terms of ICERs for NRT versus unaided cessation. Incremental costs per QALYs gained in the reference case were estimated to be €1,600, represented by the dotted line. This ICER improved for Belgium (BE), Sweden (SE), and the UK; it worsened for Germany (DE) and France (FR). Note that the ICER in Sweden and the UK improved whereas the INMB decreased. In France, the ICER worsened whereas the INMB improved. This is due to the valuation of the QALY gains with €20,000 per QALY, as a result of which a decrease in QALYs, as in the UK, has a much greater impact on the INMB than on the ICER.

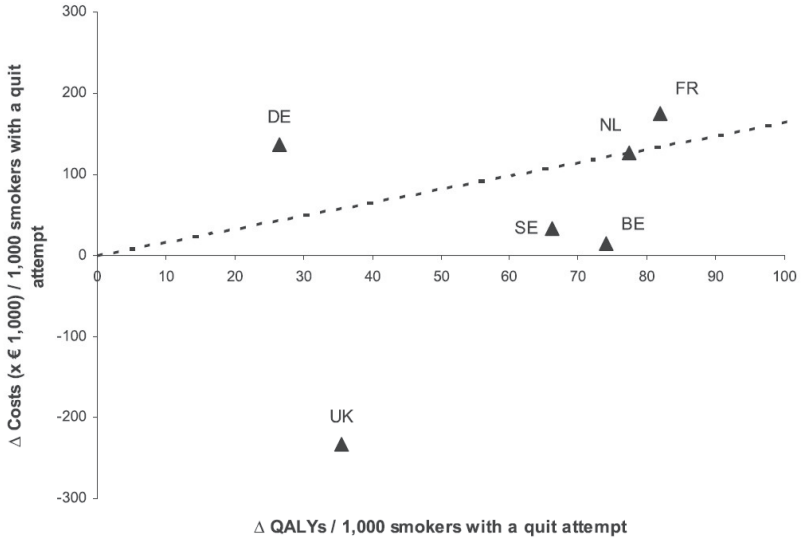

Figure 6.2: Cost effectiveness per 1,000 smokers with a quit attempt of nicotine replacement therapy versus unaided cessation influenced by all nine identified factors. Dotted line shows all points with the same ICER as NL (€1,637 / QALY). ICERs: BE = €207 / QALY; GE = €5,184 / QALY; SE = €495 / QALY; UK = -/- €6,566 / QALY; FR = €2,125 / QALY.[a]

[a] ICER: Incremental cost-effectiveness ratio

6.4 DISCUSSION

Many factors should be taken into account when transferring cost-effectiveness results across countries and settings and there are many interactions between these factors. This stresses the importance of carefully considering whether foreign results can be applied and adapted to its own setting. In this paper, we systematically investigated the impact of nine groups of country-specific model input parameters (factors) on the cross-country variability in long-term cost-effectiveness of pharmacological smoking cessation interventions. An earlier article[13] has already shown that outcomes from cost-effectiveness studies on SCTs differ considerably between countries, but causes were not unravelled. Among the factors that we have investigated, the choice of discount rate was the factor contributing the most to the between-country differences in cost-effectiveness, followed by the incidence and mortality of smoking-related diseases and the utility values used to calculate QALYs.

It is important to note that the importance of a factor in terms of its impact on the INMB depends on the WTP for a QALY. At a WTP of €20,000 per QALY, the impact of between-country differences in the cost parameters is relatively low, because the changes in the INMB are largely driven by factors affecting the QALYs. At lower values of the WTP

for a QALY, the costs of SCTs, in terms of both the unit costs and resource use, as well as the costs of smoking-related diseases, become much more important. Irrespective of the WTP for a QALY, the impact of differences between countries in demography and all-cause mortality in the INMBs is small, because the differences between the countries investigated were relatively small and do not greatly alter the cohort of smokers undertaking a quit attempt.

Despite the differences between countries, all pair-wise treatment comparisons in our study showed that the more effective smoking cessation treatments were also cost-effective, and in some case even cost-saving. INMBs were positive and ICERs were consistently below €5,300 per QALY gained. Hence, there are strong health economic arguments to support these treatments across all countries.

It is further relevant to note the differences between the changes in the ICERs and the changes in the INMBs compared with the reference case, which stresses the importance of the threshold value for a QALY in decision-making. Using the change in INMB instead of the change in ICER as a measure of the importance of a country-specific factor gives relatively greater weight to changes in QALYs. We have seen that the larger emphasis on QALYs in the INMB also affects the relative importance of a factor. For example, applying the Swedish discount rates (3% for costs and outcomes) causes both the incremental costs and QALYs gained to decrease. This leads to a change in ICER of -3% and thus a slight improvement of cost-effectiveness, whereas the INMB is a significant 32% lower.

In each country, we have used the same base-case estimates of the 12-month continuous abstinence rates.[5] This is based on the assumption that the pure biological effect of a drug can be expected to be the same, irrespective of the country. Otherwise we have used as many country-specific estimates of model input parameters as available. Some of the input data were very difficult to compare across countries. For example, smoking prevalence data may differ, because countries use different definitions and methods to determine the number of current smokers, like including only daily smokers or also including irregular smokers. In The Netherlands for example, a smoker is defined as somebody that has smoked in the 7 days before being asked[101], while in Belgium, people are asked whether they have smoked 100 cigarettes during their lives and whether they consider themselves a smoker or not.[102] In addition, the epidemiological data on smoking-related diseases are difficult to compare across countries. Different countries also used different definitions of the four diseases included in the model, especially with respect to COPD and CHD. For example, COPD was identified with ICD-10 code J40–44 in The Netherlands, but as J40-44 plus J47 in Sweden. Such a difference in definition could potentially be a source of the difference in reported epidemiology and its associated costs between countries. Furthermore, not all countries distinguish between the first-year costs of lung cancer, CHD and stroke and the costs of these diseases in later years, often because these data are not available. Such differences complicate the comparison of cost-effectiveness

between countries. Nevertheless, we have deliberately chosen to use the definitions of the smoking-related diseases and the associated cost that were actually applied at the time of writing the country-specific reimbursement dossiers for varenicline. By doing so, we highlight best the differences between countries and the influence of these differences on cost-effectiveness and cost-benefit estimates as they drive actual decision-making.

Despite its large influence on the outcomes, in only one of our six countries, The Netherlands, country-specific utility weights were available. This lack of country-specific utility data is probably due to the difficulty to collect these data and the assumption that utility values for a specific health state will probably not differ much between countries. Nevertheless, as this study shows, it is worthwhile to invest more time and resources in finding country-specific utility weights, because their impact on the INMB is large, especially at higher levels of the threshold value of a QALY.

6.5 CONCLUSION

The Transferability of Economic Data Task Force from ISPOR states on their webpage[103] that one of the most important questions to be answered with regard to transferability is "[w]hich elements of economic data vary most from setting to setting?" The results from this study suggest that it is not only important to see which factors vary, but also how much this variation in factors causes variation in cost-effectiveness. The factors that cause the most variation in cost-effectiveness do not necessarily have to be the same as the factors that vary most themselves. For example, the unit costs of the smoking cessation drugs differ considerably between countries, but the impact on the cost-effectiveness is limited when adopting a lifetime time horizon. We spent considerable time and effort on identifying data sources, adjusting input data to fit into the model and especially assessing the comparability of input parameters between countries. Based on this observation, we wholeheartedly agree with the concluding remark of the Task Force that "those developing national guidelines for economic evaluations should think carefully about the need for local data or methods, since this increases the burden on those undertaking studies in multiple jurisdictions." The results of our study underline that, when studying the cost-effectiveness of smoking cessation, there is a need for local data even for countries within a similar region of the world.

6.6 LITERATURE

[1] Drummond M, Barbieri M, Cook JR, et al. Transferability of economic evaluations across jurisdictions, in International Society for Pharmacoeconomic and Outcomes Research (ISPOR)—13th International Meeting. 2008; Toronto, Ontario, Canada.

[2] Weinstein M, O'Brien B, Hornberger J, et al. Principles of good practice for decision analytic modeling in health-care evaluation: report of the ISPOR Task Force on Good Research Practices—Modeling Studies. Value Health 2003;6:9–17.

[3] Weinstein M. Recent developments in decision-analytic modelling for economic evaluation. Pharmacoeconomics 2006;24: 1043–53.

[4] Howard P, Knight C, Boler A, Baker C. Cost-utility analysis of varenicline versus existing smoking cessation strategies using the BENESCO Simulation model: application to a population of US adult smokers. Pharmacoeconomics 2008;26:497–511.

[5] Hoogendoorn M,Welsing P, Rutten-van Mölken MPMH. Costeffectiveness of varenicline compared with bupropion, NRT and nortriptyline for smoking cessation in the Netherlands. Curr Med Res Opin 2008;24:51–61.

[6] Bolin K, Mörk A, Willers S, Lindgren B. Varenicline as compared to bupropion in smokingcessation therapy—cost–utility results for Sweden. Respir Med 2008;102:699–710.

[7] Rasch A, Greiner W. Gesundheitsökonomisches Modell der Raucherentwöhnung mit Vareniclin. Suchtmedizin in Forschung und Praxis. 2008;11:47–55.

[8] Hind D, Tappenden P, Peters J, Kenjegalieva K. Varenicline for Smoking Cessation: A Single Technology Appraisal. Sheffield, UK: ScHARR, University of Sheffield, 2007.

[9] Skoupá J, Doležal T, Hájek P, Kovárˇ P. Long term costeffectiveness and cost-utility analysis for smoking cessation in Czech Republic, in International Society for Pharmacoeconomic and Outcomes Research (ISPOR)—13th International Meeting. 2008; Toronto, Ontario, Canada.

[10] Bae J, Song HJ, Joe KH, et al. A long term cost-effectiveness analysis model for smoking cessation in Korea, in International Society for Pharmacoeconomic and Outcomes Research (ISPOR)—13th International Meeting. 2008; Toronto, Ontario, Canada.

[11] Igarashi A, Takuma H, Fukuda T, Tsutani K. Cost-utility analysis of varenicline, an oral smoking cessation drug, in Japan. Pharmacoeconomics 2009;27:247–61.

[12] Poulsen P, Dollerup J, Keiding H. The cost-effectiveness of varenicline in smoking cessation in Denmark. 2008; ISPOR 11th Annual European Congress Athens.

[13] Cornuz J, Gilbert A, Pinget C, et al. Cost-effectiveness of pharmacotherapies for nicotine dependence in primary care settings: a multinational comparison. Tob Control 2006;15:152–9.

[14] Rutten-van Mölken MPMH, Hoogendoorn M, Rasch A, Bolin K. Cost-effectiveness of varenicline for smoking cessation in five European countries, in Meeting of the Society for Research on Nicotine and Tobacco. 2007; Madrid, Spain.

[15] O'Regan C, Baker C, Marchant N. The cost-effectiveness of the novel presciption therapy varenicline in Scotland, in Meeting of the Society for Research on Nicotine and Tobacco. 2006.

[16] Orme M, Hogue SL, Kennedy LM, et al. Development of the health and economic consequences of smoking interactive model. Tob Control 2001;10:55–61.

[17] Thun M, Apicella LF, Henley SJ. Smoking vs other risk factors as the cause of smoking-attributable deaths: confounding in the courtroom. J Am Med Assoc 2000;284:706–12.

[18] Feenstra T, van Genugten ML, Hoogenveen RT, et al. The impact of aging and smoking on the future burden of chronic obstructive pulmonary disease: a model analysis in the Netherlands. Am J Respir Crit Care Med 2001;164:590–6.

[19] Hughes J, Stead L, Lancaster T. Antidepressants for smoking cessation. Cochrane Database Syst Rev 2007;(1):CD000031.

[20] Jorenby D, Hays JT, Rigotti NA, et al. Efficacy of varenicline, an alpha4beta2 nicotinic acetylcholine receptor partial agonist, vs placebo or sustained-released bupropion for smoking cessation: a randomized controlled trial. J Am Med Assoc 2006;296:56–63.

[21] Statistics Netherlands (CBS). Available from: http://www.cbs.nl [Accessed September 2, 2008].

[22] Nationaal Instituut voor Statistiek. Available from: http://statbel.fgov.be/ [Accessed November 19, 2007].

[23] Statistisches Bundesamt Deutschland. Leben in Deutschland—Mikrozensus 2005. 2006; Available from: http://www.destatis.de/presse/deutsch/pk/2006/mikrozensus2005_tabellenanhang.pdf#search=%22%22durchschnittliches%20Alter%20des%20Rauchbeginns%22%22 [Accessed August 14, 2007].

[24] Statistics Sweden (SCB). Available from: http://www.scb.se/default____2154.asp [Accessed August 31, 2007].

[25] Office for National Statistics (ONS). Available from: http://www.statistics.gov.uk/ [Accessed March 31, 2007].

[26] Beaumel C, Daguet F, Richet-Mastain L, Vatan M. La situation démographique en 2004. 2006; Available from: http://www.insee.fr/fr/ppp/ir/accueil.asp?page=SD2004/dd/sd2004_nais.htm [Accessed February 9, 2007].

[27] Zeegers T, Segaar D, Willemsen M. Roken de Harde Feiten: Volwassen 2004. Den Haag: STIVORO-voor een rookvrije toekomst, 2004.

[28] Demarest S, Drieskens S, Gisle L, et al. Health Interview Survey, Belgium. 2004; Available from: http://www.iph.fgov.be/EPIDEMIO/hisia/index.htm [Accessed November 19, 2007]. Cost-Effectiveness of Smoking Cessation 239

[29] Statistisches Bundesamt Deutschland. Gesundheitsberichterstattung des Bundes. 2006. Available from: http://www.gbe.bund.de/ [Accessed August 14, 2007].

[30] Auvray LDS, Le Fur P. L'état de Santé de la population en France en 2006. Indicateurs associé à la loi relative à la politique de santé publique, in Baromètre Santé 2005—premiers résultats.

INPES, 2006. Available from: http://www.sante.gouv.fr/drees/santepop2006/santepop2006.htm [Accessed February 9, 2007].

[31] Dumesnil S, Le Fur P, Auvray L. Santé, Soins et Protection sociale en 2000, in Série Résultats. Nancy: CREDES, 2001.

[32] GAD. UK Interim Life Tables 2002–4. 2006. Available from: http://www.gad.gov.uk/ [Accessed March 31, 2007].

[33] Poos M. COPD: Prevalentie, Incidentie, Ziekenhuisopnamen en Sterfte naar Leeftijd en Geslacht, in Volksgezondheid Toekomst Verkenning, Nationaal Kompas Volksgezondheid. Bilthoven: RIVM, 2006.

[34] Swedish National Board of Health and Welfare. Available from: http://www.socialstyrelsen.se/en/ [Accessed August 31, 2007].

[35] Lung and Asthma Information Agency. 2006. Available from: http://www.laia.ac.uk/ [Accessed December 4, 2006].

[36] Soriano J, Maier WC, Egger P, et al. Recent trends in physician diagnosed COPD in women and men in the UK. Thorax 2000; 55:789–94.

[37] Detournay B, Pribil C, Fournier M, et al. The SCOPE Study: healthcare consumption related to patients with COPD in France. Value Health 2004;7:168–74.

[38] Benard E, Detournay B, Neukirch F, et al. Prévalence de la Bronchopneumopathie Chronique Obstructive (BPCO): estimation pour la France. La Lettre du Pneumologue 2005;8:158–63.

[39] Pham Q, Gimenez M, Myre M, et al. Contribution à l'épidémiologie de la bronchite chronique chez les travailleurs du bâtiment. Bull Physio-patho Resp 1972;8:769–95.

[40] Perdrizet S, Amphoux M, Liard R. Pathologie respiratoire en médecine du travail: évaluation de trois modes de recueil des données et recherche des facteurs de risque. Rev Mal Respir 1984;1:99–103.

[41] Neukirch F, Cooreman J, Liard R, et al. Occupational exposure to bad weather, dusts and chemical fumes and respiratory abnormalities in 404 men at work (abstract). Am Rev Respir Dis 1986;133:A357.

[42] Neukirch F, Perdrizet S. La bronchite chronique. Evolution, prévention. Rev Mal Respir 1988; 5:331–46.

[43] Poos M. Longkanker naar Leeftijd en Geslacht, in Volksgezondheid Toekomst Verkenning, Nationaal Kompas Volksgezondheid. Bilthoven: RIVM, 2005.

[44] Schmidtmann I, Husmann G, Krtschil A, Seebauer G. Krebs in Rheinland-Pfalz, Inzidenz und Mortalität im Jahr 2002. Mainz: Krebsregister Rheinland-Pfaltz, 2004.

[45] Epidemiologisches Krebsregister Saarland. 2004. Available from: http://www.krebsregister.saarland.de/ [Accessed August 14, 2007].

[46] Batzler WU, Giersiepen K, Hentschel S, et al. Krebs in Deutschland 2003–2004, Häufigkeiten und Trends. Berlin: Robert Koch Institut, 2008.

[47] Robert Koch-Institut. Dachdokumentation Krebs. 2006. Available from: http://www.rki.de [Accessed August 14, 2007].

[48] Institut national de la santé et de la recherche médicale. Réseau français des registres du cancer. 2003; Available from: http://www.invs.sante.fr/publications/2003/rapport_cancer_2003/index.html [Accessed February 9, 2007].

[49] Poos M. Coronaire Hartziekten: Prevalentie, Incidentie, Ziekenhuisopnamen en Sterfte Naar Leeftijd en Geslacht, in Volksgezondheid Toekomst Verkenning, Nationaal Kompas Volksgezondheid. Bilthoven: RIVM, 2006.

[50] Kohler M, Ziese T. Telefonischer Gesundheitssurvey des Robert Koch-Instituts zu chronischen Krankheiten und ihren Bedingungen. Berlin: Robert Koch Institut, 2004.

[51] British Heart Foundation. 2006. Available from: http://www.bhf.org.uk/ [Accessed December 4, 2006].

[52] Fender M, Arveiller D, Facello A, et al. Vers une diminution de l'infarctus du myocarde dans le Bas Rhin. Ann Cardiol Angéiol 1994;43:373–79.

[53] Volmink J, Newton JN, Hicks NR, et al. Coronary event and case fatality rates in an English population: results of the Oxford myocardial infarction incidence study. The Oxford Myocardial Infarction Incidence Study Group. Heart 1998;80:40–4.

[54] Poos M. Beroerte: Prevalentie, Incidentie, Ziekenhuisopnamen en Sterfte naar Leeftijd en Geslacht, in Volksgezondheid Toekomst Verkenning, Nationaal Kompas Volksgezondheid 2004. Bilthoven: RIVM, 2004.

[55] Wolfe D, Giroud M, Kolominsky-Rabas P, et al. Variations in stroke incidence and survival in 3 areas of Europe. Stroke 2000;31:2074–9.

[56] Hoogendoorn M, Feenstra TL, Rutten-van Molken MP. [Projections of future resource use and the costs of asthma and COPD in the Netherlands] Dutch. Ned Tijdschr Geneeskd 2006;150:1243–50.

[57] Slobbe L, Kommer GJ, Smit JM, et al. Kosten van Ziekten in Nederland 2003. Bilthoven: RIVM, 2006.

[58] Suryapranata H, Ottervanger JP, Nibbering E, et al. Long term outcome and cost-effectiveness of stenting versus balloon angioplasty for acute myocardial infarction. Heart 2001;85:667–71.

[59] Struijs J, van Genugten ML, Evers SM, et al. Future costs of stroke in the Netherlands: the impact of stroke services. Int J Technol Assess Health Care 2006;22:518–24.

[60] Caekelbergh K, Annemans L, Spaepen E. Management of COPD in Belgium: A real life cost study. Value Health 2005;8:214.

[61] Singh B, Golden R. The uninsured patient. Am J Med 2006;119: 166.e1–5.

[62] Muls E, van Ganse E, Closon MC. Cost-effectiveness of pravastatin in secondary prevention of coronary heart disease: comparison between Belgium and the United States of a projected risk model. Atherosclerosis 1998;137(Suppl.):S111–16.

[63] TCT. TCT database. Available from: https://tct.fgov.be [Accessed November 19, 2007].

[64] Nowak D, Dietrich ES, Oberender P, et al. [Cost-of-illness Study for the Treatment of COPD in Germany] German. Pneumologie 2004;58:837–44.

[65] Läkemedelsstatistik AB, Medical Index Sweden. Stockholm: Läkemedelsstatistik AB, 2002.

[66] Ghatnekar O, Persson U, Glader EI, et al. Cost of stroke in Sweden: an incidence estimate. Int J Technol Assess Health Care 2004;20:375–80.

[67] Halpin D. Health economics of chronic obstructive pulmonary disease. Proc Am Thorac Soc 2006;3:227–33.

[68] Curtis L, Netten A. Unit Costs of Health and Social Care 2005. Kent, UK: PSSRU, University of Kent, 2005.

[69] McMurray J, Hart W, Rhodes G. An evaluation of the cost of heart failure to the National Health Service in the UK. Br J Med Ec 1993;6:99–110.

[70] Youman P, Wilson K, Harraf F, Kalra L. The economic burden of stroke in the United Kingdom. Pharmacoeconomics 2003; 21(Suppl. 1):43–50.

[71] Fournier M, Tonnel A-B, Housset B, et al. Impact économique de la BPCO en France: Étude SCOPE (Economic impact of COPD in France: The SCOPE study). Rev Mal Respir 2005; 22: 247–55.

[72] Gonzalez D, Rennard SI, Nides M, et al. Varenicline, an alpha4beta2 nicotinic aceylcholine receptor partial agonist, vs sustained-release bupropion and placebo for smoking cessation: a randomized controlled trial. J Am Med Assoc 2006;296:47–55.

[73] RIZIV/INAMI database. Available from: http://www.riziv.be [Accessed November 19, 2007].

[74] Geneesmiddelen Reportorium. Available from: http://www.cbip.be [Accessed November 19, 2007].

[75] Kassenärztliche Bundesvereinigung. Einheitlicher Bewertungsmaßstab. 2006; Available from: http://www.kbv.de/ebm2000plus/EBMGesamt.htm [Accessed August 14, 2007].

[76] European Medicines Agency. Summary of Product Characteristics. London: EMEA, 2006.

[77] Willers S. [Smoking cessation treatment-recommendations for Skåne, Sweden] (Swedish) in Bakgrundsmaterial till Skånelistans rekommendationer. Lund: Lund University, 2007.

[78] NHS Clinical Knowledge Summaries (formerly PRODIGY). Available from: http://cks.library.nhs.uk/home [Accessed March 31, 2007].

[79] Feenstra T, van Baal PH, Hoogenveen RT, et al. Cost-Effectiveness of Interventions to Reduce Tobacco Smoking in the Netherlands. An Application of the RIVM Chronic Disease Model. Bilthoven: RIVM, 2006.

[80] CVZ. Medicijnkosten. 2006; Available from: http://www.medicijnkosten.nl [Accessed September 2, 2008].

[81] NHS Business Services Authority. Available from: http://www.ppa.org.uk [Accessed March 31, 2007].

[82] van Baal P, Hoeymans N, Hoogenveen RT, et al. Disability weights for comorbidity and their influence on health-adjusted life expectancy. Popul Health Metr 2006;4:1.

[83] Spencer M, Briggs AH, Grossman RF, Rance L. Development of an economic model to assess the cost effectiveness of treatment interventions for chronic obstructive pulmonary disease. Pharmacoeconomics 2005;23:619–37.

[84] Mannino D, Buist AS, Petty TL, et al. Lung function and mortality in the United States: data from the First National Health and Nutrition Examination Survey follow up study. Thorax 2003;58:388–93.

[85] Trippoli S, Vaiani M, Lucioni C, Messori A. Quality of life and utility in patients with non-small cell lung cancer. Quality-of-life Study Group of the Master 2 Project in Pharmacoeconomics. Pharmacoeconomics 2001;19:855–63.

[86] Hay J, Sterling KL. Cost effectiveness of treating low HDLcholesterol in the primary prevention of coronary heart disease. Pharmacoeconomics 2005;23:133–41.

[87] Duncan P, Lai SM, Keighley J. Defining post-stroke recovery: implications for design and interpretation of drug trials. Neuropharmacology 2000;39:835–41.

[88] Tengs T, Lin TH. A meta-analysis of quality-of-life estimates for stroke. Pharmacoeconomics 2003;21:191–200.

[89] Szende A, Svensson K, Stahl E, et al. Psychometric and utilitybased measures of health status of asthmatic patients with different disease control level. Pharmacoeconomics 2004;22: 537–47.

[90] Gage B, Cardinalli AB, Owens DK. Cost-effectiveness of preference-based antithrombotic therapy for patients with nonvalvular atrial fibrillation. Stroke 1998;29:1083–91.

[91] Stouthard M, Essink-Bot ML, Bonsel GJ, et al. Disability Weights for Diseases in the Netherlands. Rotterdam: Department of Public Health, Erasmus University Rotterdam, 1997.

[92] Feenstra T, Hamberg-van Reenen HH, Hoogenveen RT, et al. Cost-effectiveness of face-to-face smoking cessation interventions: a dynamic modeling study. Value Health 2005;8:178–90.

[93] McNamara R, Lima JA, Whelton PK, et al. Echocardiographic identification of cardiovasular sources of emboli to guide clinical management of stroke: a cost-effectiveness analysis. Ann Intern Med 1997;127:775–87.

[94] Rodenburg–van Dieten HEM. Richtlijnen voor farmacoeconomisch onderzoek. Evaluatie en Actualisatie. Diemen: CVZ, 2008.

[95] Cleemput I, van Wilder Ph, Vrijens F, Huybrechts M, Ramaekers D. Richtlijnen voor farmacoeconomische evaluaties in België. Health Technology Assessment (HTA). Brussels: Federaal Kenniscentrum voor de Gezondheidszorg (KCE), 2008.

[96] Hannoveraner Konsensus Gruppe. Deutsche Empfehlungen zur gesundheitsökonomischen Evaluation—Revidierte Fassung des Hannoveraner Konsens. Gesundheitsökonomie und Qualitätssicherung 1999;4:A62–5.

[97] Edling A, Stenberg AM. General Guidelines for Economic Evaluations from the Pharmaceutical Benefits Board (LFNAR 2003:2). Solna: Pharmaceutical Benefits Board (LFN), 2003.

[98] NICE. Guide to the Methods of Technology Appraisal. London: NICE, 2004.

[99] Lévy É, De Pouvourville G. French Guidelines for the Economics Evaluation of Health Care Technologies. Paris: Collège des Économistes de la Santé, 2004.

[100] OECD.stat. Available from: http://www.oecd.org/statsportal/ [Accessed September 2, 2008].

[101] STIVORO. Available from: http://www.stivoro.nl [Accessed September 2, 2008].

[102] Bayingana K, Demarest S, Gisle L, et al. Gezondheidsenquête België. Brussel: Wetenschappelijk Instituut Volksgezondheid, 2004.

[103] The Transferability of Economic Data Task Force. Available from: http://www.ispor.org/councils/transferability.asp [Accessed September 2, 2008].

Chapter 7

Largely ignored

The impact of the threshold value for a QALY on the importance of a transferability factor

P. Vemer, M.P.M.H. Rutten-van Mölken

Previously published in European Journal of Health Economics, 2011, 12:397–404
doi: 10.1007/s10198-010-0253-3

ABSTRACT

Recently, several checklists systematically assessed factors that affect the transferability of cost-effectiveness (CE) studies between jurisdictions. The role of the threshold value for a QALY has been given little consideration in these checklists, even though the importance of a factor as a cause of between country differences in CE depends on this threshold.

In this paper, we study the impact of the willingness-to-pay (WTP) per QALY on the importance of transferability factors in the case of smoking cessation support (SCS). We investigated, for several values of the WTP, how differences between six countries affect the incremental net monetary benefit (INMB) of SCS. The investigated factors were demography, smoking prevalence, mortality, epidemiology and costs of smoking-related diseases, resource use and unit costs of SCS, utility weights and discount rates. We found that when the WTP decreased, factors that mainly affect health outcomes became less important and factors that mainly effect costs became more important. With a WTP below €1,000, the factors most responsible for between country differences in INMB were resource use and unit costs of SCS and the costs of smoking-related diseases. Utility values had little impact. At a threshold above €10,000, between country differences were primarily due to different discount rates, utility weights and epidemiology of smoking-related diseases. Costs of smoking-related diseases had little impact. At all thresholds, demography had little impact.

We concluded that, when judging the transferability of a CE study, we should consider the between country differences in WTP threshold values.

7.1 INTRODUCTION

Transferability has been defined as the degree to which the results of a cost-effectiveness study performed in one jurisdiction are representative for another jurisdiction, frequently another country. The term is also used to express the amount of adaptation that is necessary to make the results applicable to another jurisdiction.[1] A study that requires just a few simple adaptations, like a change in unit costs or discount rate, is more easily transferable than a study that requires many complicated adaptations, like a change in resource use pattern or a change in the case mix of the study population.

The question arises to what extent we can assess transferability. Recently, several checklists have been developed which systematically check the transferability of CE studies between jurisdictions [2–4]. Such checklists contain a list of issues which must be satisfied before a study is considered transferable. Examples of such issues are whether the perspective and comparators are relevant for the country of interest, whether the disease epidemiology is comparable in the country of interest and whether discount rates were specified. Some of the checklists generate a summary transferability score, assigning either equal importance to each issue[2] or weighting the importance of each issue.[3]

The advantage of such checklists is that they force the user to systematically think about transferability. The disadvantage is that the transferability scores are difficult to interpret. What does a particular transferability score really say? Moreover, in all of these checklists, there is one aspect that has been given little consideration, namely the threshold value for a QALY. Although some checklists encourage the user to discuss the generalizability of the results in the light of country-specific decision criteria, the checklists ignore the fact that the importance of a transferability issue depends on the threshold value for a QALY. Issues that are important contributors to between country differences in CE at a low threshold need not be of equal importance at a higher threshold. For example, a difference in the incidence of an infectious disease between two countries, which causes the incremental cost-effectiveness ratio (ICER) of a vaccination program to double from €12,000 in country A to €24,000 in country B, is less important when the threshold value is €50,000 per QALY but highly important when the threshold value is €20,000.

As we will discuss later in this paper, no country has an explicit threshold value for a QALY. However, this does not mean that there is no threshold. Since no country has infinite resources available for health care, decisions have to be made which interventions to reimburse and which not. At the same time, although heavily criticized, the QALY is widely regarded as a relevant outcome for decision-makers. Assuming both the existence of a country-specific threshold and the acceptance of QALYs as an outcome measures that is relevant to the decision-makers in our countries of interest, country-specific INMBs can be calculated for all relevant countries.

The aim of this paper is to assess whether the extent to which between country differences influence the country-specific INMB depends on the threshold value for a QALY. We ranked several possible causes (in this paper ''transferability factors'' or ''factors'') of between country differences in CE of smoking cessation support, according to their impact on the INMB. By doing this for different levels of the threshold value for a QALY, we studied how the importance of a transferability factor depends on this threshold value. In an earlier study, we found that at a threshold value of €20,000, discount rates and the epidemiology of smoking-related diseases were important drivers of between country differences in INMB of smoking cessation interventions.[5]

7.2 METHODS

7.2.1 The BENESCO model

The economic evaluation of smoking cessation support was based on the Benefits of Smoking Cessation on Outcomes (BENESCO) model, using a lifetime horizon and a healthcare sector's perspective.[6] The BENESCO model simulates the benefits of quitting smoking in terms of smoking-related morbidity, mortality and associated medical costs. Diseases included in the model are chronic obstructive pulmonary disorder (COPD), lung cancer, chronic heart disease (CHD) and stroke. The model is structured as a Markov model and follows a hypothetical cohort of current smokers making a single attempt to quit smoking at the beginning of the simulation. More information on the model can be found in the online appendix and several publications.[7-9]

7.2.2 Smoking cessation support

If a person decides to quit smoking, (s)he can do this unaided, or using one of the available forms of smoking cessation strategies. In this paper, we compare three forms of pharmacological smoking cessation support: nicotine replacement therapy (NRT), bupropion and varenicline. NRT is the generic term for any form of smoking cessation aid which delivers a measured dose of nicotine to the person using it. Examples include the nicotine patch or nicotine gum. Bupropion is an antidepressant used to support smoking cessation.[10] Varenicline is designed to relieve symptoms of nicotine withdrawal including cigarette craving and block the reinforcing effects of continued nicotine use.[11] The reference case to which the results of other countries in this analysis are compared is an economic evaluation performed from the Dutch health care perspective. In this study, NRT has been shown to be costeffective compared to unaided cessation, bupropion was dominant over NRT and varenicline was dominant over bupropion.[7]

7.2.3 Cost-benefit

The primary outcome of the economic evaluation was the INMB. The INMB was chosen because it explicitly incorporates the threshold value for the willingness-to-pay (WTP), or the societal value for a QALY (k). Hence, we could study the extent to which the importance of the transferability factors depends on the λ (see paragraph below). The INMB was calculated for different values of λ ranging from €0 to €50,000. Comparing two smoking cessation interventions, A and B, the INMB was calculated as:

$$[QALY(A) - QALY(B)] \times \lambda - [Costs(A) - Costs(B)].$$

A positive INMB indicates that the net benefits of intervention A are higher than the net benefits of intervention B. A negative INMB indicates that the net benefits of intervention A are lower than the net benefits of intervention B.

7.2.4 Factors affecting transferability

Within the BENESCO model, all variables except the risk ratio of getting a smoking-related illness were changed to calculate country-specific cost-effectiveness. These variables can intuitively be grouped in a total of nine transferability factors, which could potentially cause differences in the cost-benefit of smoking cessation support between countries. We investigated each of these factors. Each transferability factor consisted of a group of country-specific input parameters which were varied simultaneously. Demography included the total number of people older than 18 years of age in six age/gender classes. Smoking prevalence refers to the percentage of smokers, non-smokers and former smokers in each age/gender class. All-cause mortality was the percentage of the total number of people in each age/gender class that dies during a single year. The epidemiology of smoking-related diseases consisted of three elements: the incidence rates, prevalence rates and annual cause-specific mortality rates by age/gender class. The costs of smoking-related diseases were separated into first year costs and costs in subsequent years, except for COPD. The amount of resource use (i.e., medication and counselling) associated with the SCS and the unit costs of these resources were the two elements defining the intervention costs. The utility weights were defined both for the general population and for patients with a smoking-related disease. The discount rates for both costs and outcomes were set equal to the values recommended in 2007 in the national guidelines for economic evaluations. For each country, we adopted a health care perspective. Information on the main input factors can be found in appendix A7.

7.2.5 The importance of each transferability factor

Based on the hierarchy of effectiveness of the smoking cessation interventions, we calculated the INMB of NRT versus unaided cessation, bupropion versus NRT and varenicline

versus bupropion in The Netherlands.[7] These results were used as the reference values. We then changed the Dutch input values of each of the nine factors individually to the country-specific values for Germany, Sweden, the UK, Belgium and France.[5, 8, 12, 13] This changes the INMB. The percentage of change in the INMB was then averaged over the three treatment comparisons. Each factor is then rank ordered from highest (1) to lowest (9) percentage of change in INMB. We compare these rankings for different threshold values, ranging from €0 to €50,000 per QALY, for each country separately.

7.3 RESULTS

The INMB of NRT versus unaided cessation in The Netherlands at a WTP threshold of €20,000 per QALY was €1.42 million. The INMB of bupropion versus NRT was €0.39 million, and the INMB of varenicline versus bupropion was €0.90 million. Table 7.1 shows the ranking of factors at different threshold values for a QALY, after replacing the Dutch reference values with the German country-specific values. It is clear from this table that the threshold value has a considerable impact on the importance of each factor.

This can be seen most clearly for the cost of smoking-related diseases, resource use, unit costs and utility weights. The INMB is calculated as the difference in QALYs multiplied by λ minus the difference in costs. Hence, the QALY gains are weighted with the λ before subtracting the additional costs. When λ decreases, the transferability factors that mainly affect the health outcomes are given less weight and become less important, whereas the transferability factors which mainly affect the costs become more important. Among the latter are the costs of smoking cessation support, which are driven by the resources

Table 7.1: Univariate ranking of factors after changing to the German values, at different threshold values for the WTP for a QALY.[a]

Threshold value for WTP	Demography	All-cause mortality	Smoking prevalence	Epidemiology	Costs of smoking-related diseases	Resource use	Unit costs	Utility weights	Discount rates
€0	7	8	6	5	3	1	2	9	4
€100	7	8	6	5	4	1	2	9	3
€500	8	9	6	5	4	1	3	7	2
€1,000	8	9	7	5	3	1	4	6	2
€5,000	8	9	7	4	6	2	5	3	1
€10,000	8	9	6	3	7	4	5	2	1
€20,000	8	9	4	3	7	5	6	2	1
€50,000	7	9	4	3	8	5	6	2	1

[a] WTP: Willingness-to-pay; Ranked according to the percentage change in INMB compared to the reference country, averaged over three pair-wise comparisons of smoking cessation strategies.

used and the unit costs of these resources, and the costs of smoking-related diseases. When λ increases, the transferability factors with a large impact on the QALYs are given a higher weigh, increasing their importance as a cause of between country differences in net benefit. As a consequence, utility weights, epidemiology (mainly incidence and mortality) of smoking-related diseases and smoking prevalence become more important.

When replacing Dutch by German input values, the discount rate, which changed from 4% for costs and 1.5% for effects in The Netherlands to 5% for both costs and effects in Germany, led to the highest percentage of change in INMB at thresholds of €5,000 or higher. At lower thresholds, the relative importance of the discount rate decreases, with the resource use taking over as the most important factor. All-cause mortality and demography are relatively unimportant at all threshold values.

Similar tables result for the other countries. They are shown in appendix A7. For the other four countries, the same general overview can be seen, although there are individual differences. Tables on the outcomes of other countries can be found in appendix A7. In all countries, unit costs and/or resource use are amongst the two most important causes of between country differences in cost benefit at low thresholds. At high thresholds, utility weights, discount rates and the epidemiology of smoking-related diseases are the three most important factors for all countries, except for Belgium which had almost the same discount rates as in The Netherlands.

Figure 7.1 shows the percentage of change in the INMB of NRT versus unaided cessation, when substituting the reference case input values to the German values, at various levels of λ. The results were comparable for the other pair-wise comparisons of smoking cessation interventions (not shown, available on request). At a λ of €0, substituting Dutch resource use by German resource use led to the greatest change in INMB, a decrease of 92%. The second most important transferability factor was the costs of smoking-related diseases, decreasing the INMB with 29%. At a λ of €500, Germany-specific resource use still caused the INMB to change most, but now followed by the discount rates. At a λ of €5,000 or above, the discount rates had the highest impact on the INMB, causing a decrease in INMB of 99%, followed by the utility weights, causing a decrease in INMB of 25%.

The closer the threshold value of a QALY was to the original ICER in the reference country—in this case €1,637 per QALY—the more the INMB was affected by a change to country-specific input data. This was true for every transferability factor, since the INMB of the base case comparison is close to zero with a λ close to the ICER. For most transferability factors, the sign of the impact on the INMB at a threshold of €1,000 or lower switched from positive to negative or vice versa at a threshold of €5,000 or higher. Only the factor discount rates for Sweden and the UK and the factor epidemiology for Belgium and Sweden show a slightly different pattern. The reason is that the INMB of the base case comparison is always positive with a λ above the value of the ICER, but negative with a λ

below the ICER. A similar change in INMB due to a new country-specific variable value will therefore have a directly opposite effect.

Figure 7.2 shows, for various levels of λ, the change in INMB of NRT versus unaided cessation, when all reference case input values are simultaneously replaced by country-specific values. This change to the country-specific INMB is shown for each country separately. At lower levels of λ, the INMB changes most when adapting the Dutch input data to the UK input data. This is because the resource use and costs of NRT in the UK

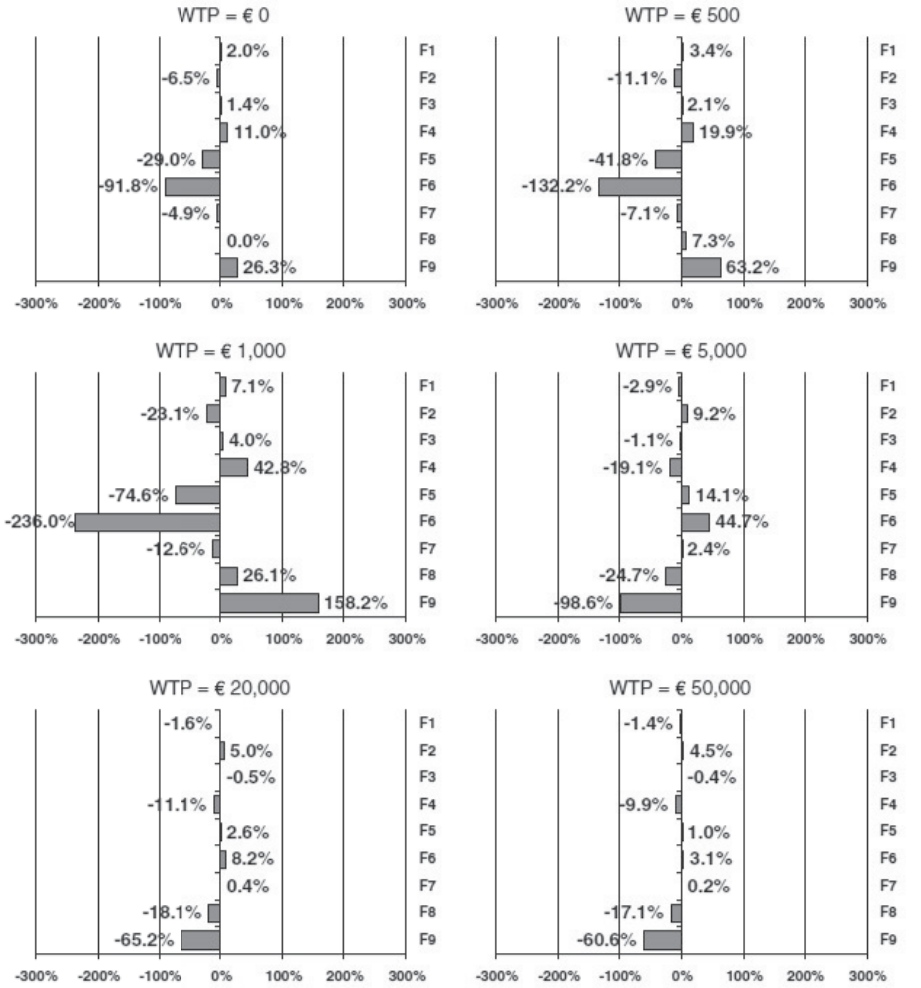

Figure 7.1: Percentage of change in the INMB of NRT versus unaided cessation, when replacing Dutch input values for each factor by the German input values, at different threshold values for the WTP for a QALY. F1: demography, F2: smoking prevalence, F3: mortality, F4: epidemiology, F5: costs of smoking-related diseases, F6: resource use, F7: unit costs, F8: utility weights, F9: discount rates.

[a] WTP: Willingness-to-pay

differ most from that in The Netherlands, i.e., NRT is considerably less expensive in the UK than in The Netherlands. At higher levels of λ, the INMB changes most when adapting the Dutch input data to the German input data. This is because of the great difference in discount rates.

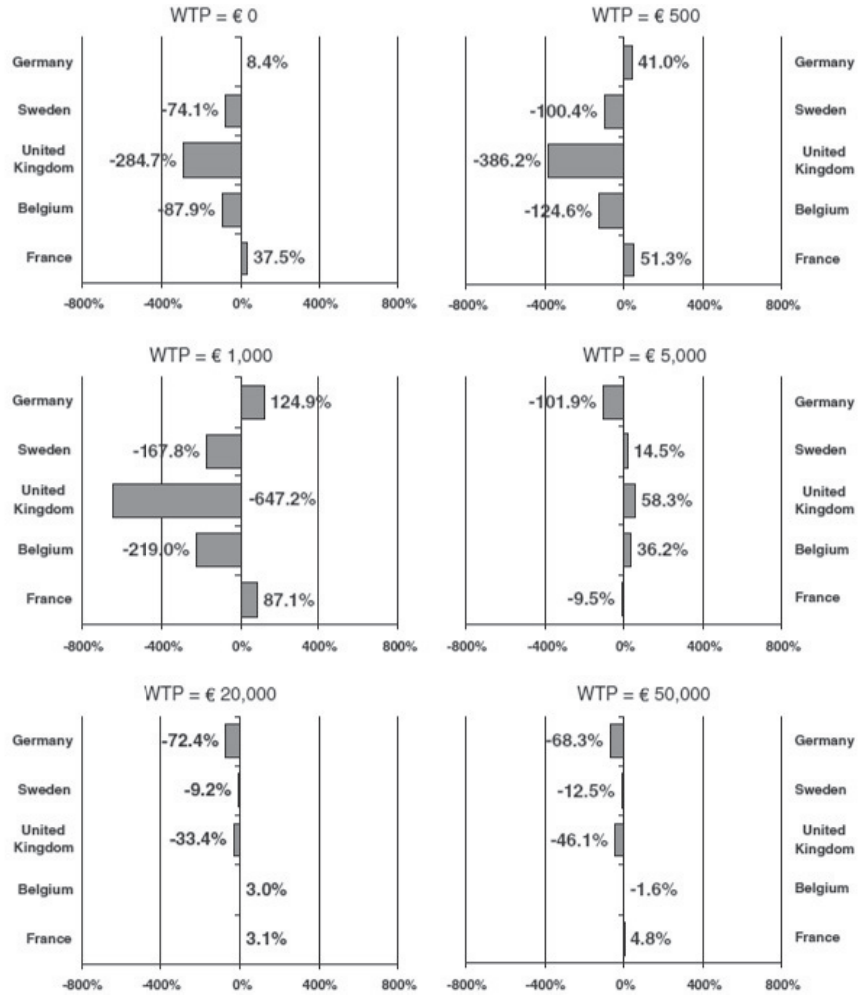

Figure 7.2: Percentage of change in the INMB of NRT versus unaided cessation, when simultaneously replacing Dutch input values for all transferability factors by the country-specific input values at different threshold values for the WTP for a QALY.

[a] WTP: Willingness-to-pay

7.4 DISCUSSION

This paper has clearly shown that, when we transfer a foreign economic evaluation to the country of interest, the factor which should most carefully be adapted is driven by the threshold value for a QALY. We feel that this aspect has been given too little attention in the transferability debate until now. When discussing the transferability of smoking cessation evaluations in countries with low threshold values, we should pay most attention to the country-adaptation of the cost drivers, i.e., the costs of the interventions and the disease that are studied. In countries with high threshold values, we should pay most attention to the country-adaptation of the factors that affect the health outcomes. These include disease epidemiology and utility values. Although the results are specific for smoking cessation interventions, and indeed for the BENESCO model, we feel this conclusion is applicable to similar interventions of a preventative nature that require initial investments which by far precede the returns on these investments in terms of improved health outcomes and savings in the costs of health care utilization. In addition, we feel that based on this paper, the threshold value for a QALY should be an integral part of the investigation of transferability for all economic evaluations.

The approach taken in this paper assumes the acceptance of the QALY as a relevant decision-making outcome and the existence of threshold values, either explicit or implicit. The reason that the role of the threshold value has been largely ignored in the transferability discussion may be related to the fact that we are far from reaching a consensus on the maximum willingness to pay for a QALY. There is not a single jurisdiction where the threshold value is really fixed. With respect to the countries in our current study, NICE (National Institute for Health and Clinical Excellence) in the UK mentions the most explicit threshold value, but they too reject the use of a single, absolute threshold, instead preferring to make decisions on a case-by-case basis. NICE is unlikely to reject a technology with a ratio in the range of £ 5,000 – £ 15,000 per QALY solely on the grounds of cost ineffectiveness but would need special reasons for accepting technologies with ratios over £ 25,000 – £ 35,000 per QALY as cost effective.[14] In The Netherlands, a threshold of €20,000 per QALY is often cited. However, this threshold was obtained from economic evaluations of preventive interventions and is certainly not used consistently.[15] Currently, there is discussion in The Netherlands on increasing the threshold value depending on the burden of the disease of interest, with a maximum threshold of up to €80,000 for very severe diseases.[16,17] In Germany, IQWiG (Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen) has prepared guidelines for economic evaluation as part of the evaluation of the value of pharmaceuticals[18], but coverage does not depend on any externally set maximum standard. In 2008, the Belgian KCE (Federal Knowledge Centre for Health Care) started to produce methodological reports in order to help standardize the methodology used for health technology assessment.[19] One such report mentions

that CE is now only rarely used as an argument for reimbursement. The budget impact is considered more important. This report was intended as a starting point for a discussion on the role of threshold values in Belgium.[20] In Sweden, no guidance has been given as to acceptable cost-effectiveness ratios, defined in terms of cost per QALY or otherwise, although cost-effectiveness is considered a crucial aspect in the Swedish reimbursement system.[21] Rather than apply a single threshold, there may be different (implicit) thresholds depending on the severity of the disease or an assessment of patient need.[22] In France, no threshold value is used when making health care decisions.[23]

In a cost-utility analysis, an intervention is found to be cost-effective if the cost per QALY falls below the WTP for a QALY. This value is used as an external threshold value against which the ICER is compared. In the cost benefit analysis, the societal value of a QALY is directly incorporated into the analysis. For an intervention to be cost-effective, the INMB (calculated as [QALY(A) – QALY(B)] x λ – [Costs(A) – Costs(B)]) needs to be positive. In the net benefit approach, it is apparent that the impact of a country-specific model parameter on the decision to adopt an intervention depends on the threshold value of the QALY. Though not immediately visible in the cost-utility analysis itself, the impact of a parameter on the probability that the ICER falls below the threshold value equally depends on the level of this threshold value. Whether or not the threshold value is explicitly incorporated into the analysis does to affect the importance of a country-specific parameter in the transferability discussion. The consequence of our findings may be that we have to adjust the available transferability check lists to encourage checking whether there might be differences in the willingness to pay between the countries, next to checking how well a foreign study represents the circumstances in the country of interest and how much effort is required to adapt particular data inputs to the country of interest. For example, NRT, bupropion and varenicline are currently not covered by the basic health care insurance in The Netherlands, although the health insurance board CVZ has recently advised the Ministry of Health to reimburse them as part of an integrated smoking cessation program.[24] This implies a current willingness to pay of €0. Consequently, when future foreign CE studies of new smoking cessation interventions (e.g., a vaccine) become available, the transferability of the cost of this intervention to The Netherlands will be driving the reimbursement discussion. In contrast, in the UK, where NRT, bupropion and varenicline are already paid by the NHS, the discussion on the transferability of foreign cost-effectiveness studies of such a drug might focus more on the representativeness of the epidemiology of smoking-related diseases for the UK and the utility values of patients having a smoking-related disease.

7.5 CONCLUSION

When judging the transferability of a CE study from one jurisdiction to another, it is relevant to consider the between country differences in threshold values per QALY. Between country, differences in cost-effectiveness are determined by between country differences in unit costs, disease epidemiology, discount rates etc., but the importance of each of these is influenced by the threshold value for a QALY. Between country differences that are important at a low threshold value might be less important at a high threshold value and vice versa.

7.6 LITERATURE

[1] Drummond, M., Barbieri, M., Cook, J., Glick, H., Lis, J., Malik, F., et al.: Transferability of economic evaluations across jurisdictions: ISPOR good research practices task force report. Value Health 12, 409–418 (2009)

[2] Boulenger, S., Nixon, J., Drummond, M.F., Ulmann, P., Rice, S., de Pouvourville, G.: Can economic evaluations be made more transferable? Eur. J. Health Econ. 6, 334–346 (2005)

[3] Antonanzas, F., Rodriguez-Ibeas, R., Juarez, C., Hutter, F., Lorente, R., Pinillos, M.: Transferability indices for health economic evaluations: methods and applications. Health Econ. 18, 629–643 (2009)

[4] Nixon, J., Rice, S., Drummond, M., Boulenger, S., Ulmann, P., de Pouvourville, G.: Guidelines for completing the EURONHEED transferability information checklists. Eur. J. Health Econ. 10, 157–165 (2009)

[5] Vemer, P., Rutten-van Mölken, M.P.M.H.: Crossing borders: factors affecting differences in cost-effectiveness of smoking cessation interventions between European countries. Value Health 13, 230–241 (2010)

[6] O'Regan C, Baker C, Marchant N. The cost-effectiveness of the novel prescription therapy varenicline in Scotland. In: Meeting of the Society for Research on Nicotine and Tobacco, 14–18 Feb, Madison (WI), USA (2006)

[7] Hoogendoorn, M., Welsing, P., Rutten-van Mo¨lken, M.P.M.H.: Cost-effectiveness of varenicline compared with bupropion, NRT and nortriptyline for smoking cessation in the Netherlands. Curr. Med. Res. Op. 24, 51–61 (2008)

[8] Bolin, K., Mörk, A., Willers, S., Lindgren, B.: Varenicline as compared to bupropion in smoking cessation therapy—Cost–utility results for Sweden. Respir. Med. 102, 699–710 (2008)

[9] Howard, P., Knight, C., Boler, A., Baker, C.: Cost-utility analysis of varenicline versus existing smoking cessation strategies using the BENESCO Simulation model: application to a population of US adult smokers. Pharmacoeconomics 26, 497–511 (2008)

[10] Hughes, J., Stead, L., Lancaster, T. Antidepressants for smoking cessation. Cochrane Database Syst. Rev. 1 Art.No. CD000031 (2007). doi:10.1002/14651858.CD000031.pub3

[11] Jorenby, D.E., Hays, J.T., Rigotti, N.A., et al.: Efficacy of varenicline, an alpha4beta2 nicotinic acetylcholine receptor partial agonist, vs. placebo or sustained-released bupropion for smoking cessation: a randomized controlled trial. J. Am. Med. Assoc. 296, 56–63 (2006)

[12] Rasch, A., Greiner, W.: Gesundheitsökonomisches Modell der Raucherentwöhnung mit Vareniclin. Suchtmedizin in Forschung und Praxis 11, 47–55 (2009)

[13] Annemans, L., Nackaerts, K., Bartsch, P., Prignot, J., Marbaix, S.: Cost effectiveness of varenicline in Belgium, compared with bupropion, nicotine replacement therapy, brief counselling and unaided smoking cessation: a BENESCO Markov cost-effectiveness analysis. Clin. Drug Investig. 29, 655–665 (2009)

[14] Rawlins, M.D., Cuyler, A.J.: National Institute for Clinical Excellence and its value judgements. Br. Med. J. 329, 224–227 (2004)

[15] Pomp, M., Brouwer, W.B.F., Rutten, F.F.H.: QALY-tijd. Nieuwe medische technologie. kosteneffectiviteit en richtlijnen. Den Haag, Centraal Planbureau (CPB) (2007)

[16] RVZ. Zinnige en duurzame zorg (Sensible and sustainable Care). Zoetermeer. Raad voor de Volksgezondheid en Zorg (Council for Public Health and Health Care) Report nr. 06/06 (2006)

[17] RVZ. Rechtvaardige en duurzame zorg (Justified and sustainable care). Den Haag. Raad voor de Volksgezondheid en Zorg (Council for Public Health and Health Care) Report nr. 07/04. ISBN 978-90-5732-183-2 (2007)

[18] Perleth, M., Gibis, B., Gohlen, B.: A short history of health technology assessment in Germany. Int. J. Technol. Assess. Health Care. 25(Suppl 1), 112–119 (2009)

[19] Cleemput, I., Van Wilder, P.: History of health technology assessment in Belgium. Int. J. Technol. Assess. Health Care. 25(Suppl 1), 82–87 (2009)

[20] KCE. Drempelwaarden voor kosteneffectiviteit in de gezondheidszorg. Brussels. Federaal Kenniscentrum voor de Gezondheidszorg (KCE). Report nr. 100A (2008)

[21] LFN. The Swedish pharmaceutical reimbursement system. Pharmaceutical Benefits Board (LFN), Solna (2007)

[22] Moïse P, Docteur E. Pharmaceutical pricing and reimbursement in Sweden. OECD health working papers no 28, OECD, Paris (2007)

[23] Rochaix L. In France, no threshold value is used when making health care decision. Report nr. 29006531, personal communication, 5 Jan (2010)

[24] Kroes ME, Mastenbroek CG. Stoppen-met-rokenprogramma: te verzekeren zorg! Diemen. CVZ (2009)

A7 APPENDIX

A7.1 Background information BENESCO model and additional results

The projections of the effects of smoking cessation were based on the BENESCO (Benefits of Smoking Cessation on Outcomes) model.[1] The BENESCO model simulates the consequences of smoking and the benefits of quitting in terms of smoking-related morbidity, mortality and associated medical costs in a population. The model is structured as a Markov model (cycle length 1 year) and follows a hypothetical cohort of current smokers making a single attempt to quit smoking at the beginning of the simulation. The cohort is followed from the time of their quit attempt until all members of the cohort have died. Individuals are classified into one of three smoking states, i.e., smoker, recent quitter (abstinent 1–5 years after successful quit attempt) or long-term quitter. Transition probabilities between smoking states in the first year depend on cessation rates of the interventions, while the probabilities after 1 year depend on relapse rates, which in turn depend on time since quitting. The model simulates the age, gender and smoking status-specific incidence and mortality of four major diseases for which smoking is a well-established risk factor: chronic obstructive pulmonary disease (COPD), lung cancer, coronary heart disease and stroke. Smoking state-specific incidence and mortality rates were calculated using relative risks.[2, 3] The incidence and mortality rates for recent quitters were calculated using the relative risks of former smokers versus never smokers, while the rates for long-term quitters were assumed to be the same as those of never smokers. Because COPD and lung cancer are chronic progressive conditions, these diseases were given hierarchical prominence over the other conditions with acute recurrent events. This means that individuals with COPD or lung cancer remain in this state until they die and cannot move to a CHD or stroke state, whereas individuals with CHD or stroke can move to the COPD or lung cancer state. A patient cannot have two diseases at the same time. The model calculates the total number of smokers and quitters that have one of the smoking-related diseases as well as the number of deaths (due to one of the smoking-related diseases and overall) over the time horizon of the simulation. Based on these numbers, the total health care costs associated with the different disease states and the total number of (quality adjusted) life years are calculated. The model uses three age bands: 18–34 years, 35–64 years and 65 years and older. Subjects alive in the model at age 99 years are all assumed to die in the next cycle. It is assumed that there is no smoking-related morbidity or mortality in the 18–34 years age class.

A7.2 LITERATURE

[1] Hoogendoorn, M., Welsing, P., Rutten-van Mölken, M.P.M.H.: Cost-effectiveness of varenicline compared with bupropion, NRT and nortriptyline for smoking cessation in the Netherlands. Curr. Med. Res. Op. 24, 51–61 (2008)

[2] Thun, M.J., Apicella, L.F., Henley, S.J.: Smoking vs. other risk factors as the cause of smoking-attributable deaths: confounding in the courtroom. J. Am. Med. Assoc. 284, 706–712 (2000)

[3] Feenstra, T.L., van Genugten, M.L., Hoogenveen, R.T., Wouters, E.F., Rutten-van Mölken, M.P.M.H.: The impact of aging and smoking on the future burden of chronic obstructive pulmonary disease: a model analysis in the Netherlands. Am. J. Respir. Crit. Care Med. 164, 590–596 (2001)

Chapter 8

The road not taken

Transferability issues in multinational trials

P. Vemer, M.P.M.H. Rutten-van Mölken

Previously published in PharmacoEconomics (2013) 31(10):863–876
DOI 10.1007/s40273-013-0084-z

Acknowledgments: We would like to acknowledge the useful comments made by three anonymous reviewers.

ABSTRACT

Background National regulatory agencies often have to use cost-effectiveness (CE) data from multinational randomized controlled trials (RCTs) for national decision making on reimbursement of new drugs. We need to make the best use of these patient-level data to obtain estimates of country-specific CE. Several methods, ranging from simple to statistically complex, have existed for years. We investigated which of these methods are used to estimate CE ratios in economic evaluations performed alongside recent, multinational RCTs that enrolled at least 500 patients.

Methods In this systematic literature review, studies were classified based on whether resource use, unit costs, health outcomes and utility value sets were obtained from all countries, a subset of countries or one country. We recorded if the study presented trial-wide and country-specific CE results and reported the statistical analyses that were used to estimate them.

Results We included 21 studies, of which the majority used measurements of health care utilization and health outcomes from all countries to estimate CE. Thirteen studies used a one-country valuation of health care utilization; six used a multi-country valuation. Despite the availability of country-specific utility value sets, none of the studies that presented quality-adjusted life-years (QALYs) used multi-country valuation. Valuation of health care utilization and health outcomes was not always consistent within a study: three studies combined a multi-country valuation of health care utilization, with a one-country valuation of health outcomes. Most studies calculated trial-wide CE estimates, while 11 studies calculated country or region-specific estimates. Thirteen studies used relatively simple methods, which do not take the possible interaction between the country and treatment effect on health care utilization and health outcomes into account. Eight studies used more advanced statistical methods. Three of them used a fixed-effects modeling approach. Five studies explicitly took the hierarchical structure of the data into account, which leads to more appropriate estimates of population average results and associated standard errors. In this way, they help improve transferability of the published results.

Conclusion Based on this systematic review, we concluded that the uptake of more advanced statistical methods has been relatively slow, while simpler naïve methods are still routinely employed.

8.1 INTRODUCTION

Once regulatory approval has been obtained, pharmaceutical companies have to file for reimbursement of a drug in many different countries, each with their own specific regulations. In many countries, decision makers are in principle prepared to consider cost-effectiveness (CE) evidence from an international origin, provided that the data are adapted to their own country and setting. Such adaptations are often done with decision analytic models that are filled with country-specific epidemiological and economic data. Sometimes, the country of interest has participated in a CE study that was linked to a multinational clinical trial. In that case, there are patient-level data on cost-effectiveness. The challenge is to make the best use of this data to obtain a CE estimate that represents the country of interest best. This is the topic of this paper.

Three simple methods are frequently used to calculate CE estimates from multinational randomized controlled trials (RCTs). The first simple method is to use only data from the country of interest. This method disregards all data collected in other countries.[1] For example, in a trial with both Dutch and UK data, only the data from Dutch patients is included. The second method is to combine measurements of health care utilization and health outcomes from all or a subset of countries included in the RCT, which are then valued with weights (unit costs, utility values) from a single country (one-country valuation, see table 8.1). In our example, we count the number of hospital bed days in both countries, which are then multiplied with the Dutch costs for each bed day. In effect, all patients in the trial are treated as if they come from the country where the valuation comes from.[2,3] The final simple method that is frequently used, combines measurements of resource utilization and health outcomes with country-specific values that only apply to that specific country (multi-country valuation). In our example, the hospital bed days used by Dutch patients are multiplied by the Dutch unit costs, and the bed days from UK patients are multiplied by the UK unit costs. These three methods do not take into account the interaction between country and treatment effect on health care utilization and health outcomes, an interaction that may be due to differences in for example epidemiology, practice patterns, payment systems and unit costs. In this sense, they might be called naïve.

More advanced statistical techniques have been developed, which do take this interaction into account. The first such method was introduced in 1998 by Willke and colleagues[4], who proposed the use of two separate patient-level regression models, modeling both the direct effect of a treatment on costs and the indirect effect of treatment on costs through a change in health outcomes. They were soon followed by other studies proposing different fixed effect models with country-level covariates, for example Koopmanschap and colleagues and more recently Clarke and colleagues.[5,6]

Table 8.1: Terminology used in classifying transferability issues of multinational trials.

Measurement	
One-country (1C)	Measurements from a single country are included
Subset	Measurements from a subset of countries are included
All countries (All)	Measurements from all countries in the RCT are included
Measurements in	
health care utilization	E.g.: number of days in a hospital, number of pills taken
health outcomes	E.g.: survival, number of exacerbations, number of events
Valuation	
One-country valuation (1C)	The weights used, come from a single country
Multi-country valuation (MC)	Each country has their own weights
Unweighted (UW)	No weights are used
Measurements in	
health care utilization	Unit costs
health outcomes	Quality of life weights

Health economic data from multinational RCTs naturally fall into the hierarchical structure of multiple micro units (patients) within macro units (countries). To take this structure into account, multilevel modeling (MLM) has been used.[7-10] MLM uses a random intercept for each macro unit to model the hierarchical structure, but can be expanded to include a random slope or country-level (fixed) covariates. Hierarchical models lower the variability of country-specific CE results by borrowing strength from the other countries, and lead to more appropriate estimates and associated standard errors.[11-14] In a recent review of available methods, Manca et al. conclude that these hierarchical models are the most appropriate tool to analyze CE alongside a multinational trial.[15]

In 2005, Barbieri and colleagues showed that one of the most frequently used methods of calculating CE was to combine country-specific measurements of health care utilization and weights from a single country.[2] Because more advanced methods have been available for several years and their use has been recommended by the ISPOR Task Force on Transferability[16], we aimed to review the experience with these methods. We examined recent multinational cost-effectiveness analyses (CEAs) that were conducted alongside large RCTs and described and summarized how these studies dealt with transferability aspects. In more detail, we aimed to answer the following questions:

- In what way have the researchers looked for evidence of heterogeneity between countries and (how) have heterogeneity issues been addressed?
- How have the researchers calculated trial-wide CE estimates?
- How have the researchers calculated country-specific CE estimates?

At all points in the texts where we say country, it could also be read as any other jurisdiction.

8.2 METHODS

8.2.1 Search strategy

A systematic electronic literature search was performed for CEAs alongside multinational RCTs, published in 2005 or after, written in English, Dutch or German. Because the more advanced statistical techniques can only be applied to large numbers of patient, the RCT in question needed to be conducted on at least 500 individual patients from at least 2 different countries. The CEA needed to be performed on individual patient-level data (IPD). We searched both PubMed and EMBase with combinations of the following key words (one from each category):

- "multinational", "international", "multiple countries", "multi-country" and/or "regional";
- "cost-effectiveness"; "cost-utility"; "costs"; "ICER" (incremental cost-effectiveness ratio) and/or "QALY" (quality adjusted life year);
- "trial" and/or "RCT".

Additional studies were sought by hand searching the reference list of original research papers and review papers on transferability that were found in the initial literature search. As we only wanted to include CEAs, we excluded studies that did not measure both health care utilization and health outcomes. We excluded all (systematic) reviews, meta-analyses, evaluations or overviews of treatments and programs, guidelines and recommendations for performing an RCT. Decision analytic models, like Markov models, were also excluded. Studies where the primary source of data was not a randomized controlled trial (i.e. registries, cohort studies, longitudinal studies or non-randomized studies) and published abstracts were excluded.

8.2.2 Data extraction

For all the papers included in the study, we first looked for the name of the RCT, disease area, interventions, number of patients and number of countries. Next, we classified the studies based on whether the measurement of health care utilization and health outcomes was based on patients in just one country, a subset of countries or all countries. See table 8.1 for an overview of the terminology used. In addition, studies were classified by the source of the valuation for the health care utilization and the health outcomes: unit costs and utility values (quality of life weights). In a one-country valuation, an analyst applies the weights from one country to measurements from all countries. In a multi-country valuation, an analyst applies weights from each individual country to the quantities from

that same country. If the health outcomes reported are other than QALYs, no value sets are used for health outcomes and the valuation is classified as unweighted. Studies may be classified differently for the valuation of health care utilization and health outcomes.

We recorded if the study described trial-wide and country-specific CE results, how these were calculated and how heterogeneity between countries was measured. Finally, we classified the studies based on the statistical methods used to analyze the data. The studies were classified as a "simple method", a "fixed effects regression model" or a "hierarchical regression model".

8.3 RESULTS

8.3.1 Literature search

The literature search was performed on Jul 12th, 2013. We identified 821 potentially eligible papers, with 318 titles from PubMed and 716 from EMBase. Based on the title of the paper, we excluded 391 titles. Reasons for exclusion are listed in figure 8.1. Upon

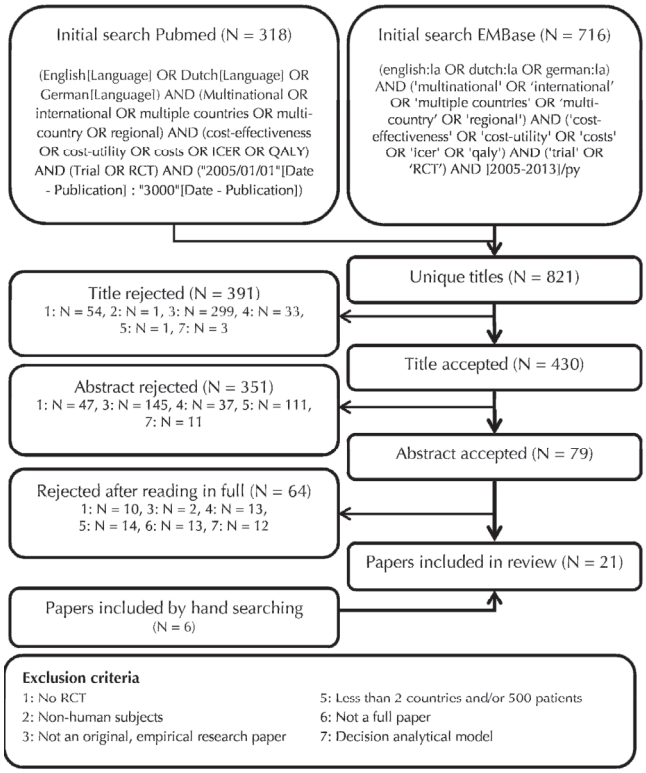

Figure 8.1: In- and exclusion of papers at various stages.

Table 8.2: Papers included in study.

First author and year	Disease[a]	RCT on which study is based		Number of countries
		Name	Interventions and number of patients[b]	
Lindgren 2005 [17]	Hypertension	ASCOT-LLA	Atorvastatin (n = 5,168) Placebo (n = 5,137)	7
Lofdahl 2005 [18]	COPD	-	Budesonide (n = 257) Formoterol (n = 255) BFC (n = 254) Placebo (n = 256)	15
Peeters 2005 [19]	Psoriasis vulgaris	-	Calipotriol / betamethasone dipropionate followed by calcipotriol alone (n = 249) Tacalcitol (n = 252)	4
Pinto 2005 [13]	Acute MI	ASSENT-3	N = 6,095 Heparin Enoxaparin Abciximab	26
Radeva 2005 [20]	Severe side effects after heart transplantation	-	Everolimus 1.5 mg/day (n = 209) Everolimus 3.0 mg/day (n = 211) Azathioprine (n = 214)	14
Reed 2005 [21]	Acute MI	VALIANT	Valsartan (n = 4,909) Captopril (n = 4,909) Both (n = 4,855)	24
Weintraub 2005a [22]	ACS	CURE	N = 12,562 Clopidogrel Placebo	28
Weintraub 2005b [23]	Acute MI	EPHESUS	N = 6,632 Eplerenone Placebo	37
Briggs 2006 [24]	Asthma	GOAL	N = 3,416 FP alone SFC	44
Tonkin 2006 [25]	ACS	LIPID	Pravastatin (n = 4,470) Placebo (n = 4,544)	2
Willan 2006 [26]	Parkinson's Dementia Disease	EXPRESS	Rivastigmine (n = 362) Placebo (n = 179)	12
Bachert 2007 [27]	ARC	- (idem Canonica)	Grazax (n = 316) Placebo (n = 318)	8
Canonica 2007 [28]	ARC	- (idem Bachert)	Grazax (n = 316) Placebo (n = 318)	8
Manca 2007 [29]	CHF	ATLAS	Low dose lisinopril (n = 1,596) High dose lisinopril (n = 1,568)	19
Rutten-van Mölken 2007 [30]	COPD	-	Roflumilast (n = 761) Placebo (n = 753)	14

Table 8.2: Papers included in study. (Continued)

		RCT on which study is based		
First author and year	Disease[a]	Name	Interventions and number of patients[b]	Number of countries
Willan 2008 [31]	?	?	"T" (n = 680) "S" (n = 676)	14
Marcoff 2009 [32]	ST-segment elevation MI	ExTRACT–TIMI 25	N = 20,506 Enoxaparin Unfractionated heparin	48
Briggs 2010 [33]	COPD	TORCH	Salmeterol (n = 1,521) FP (n = 1,534) SFC (n = 1,533) Placebo (n = 1,524).	42
Gomes 2010 [34]	Anaesthesia for carotid surgery	GALA	General Anaesthetic (n = 1,753) Local infiltration and cervical plexus nerve block (n = 1,773)	24
Lorgelly 2010 [35]	CHF	CORONA	Rosuvastatin (n = 2,514) Placebo (n = 2,497)	21
Rogkakou 2011 [36]	Persistent rhinitis	XPERT	Levocetirizine (n = 278) Placebo (n = 237)	5

[a] MI = Myocardial infarction, ARC = Allergic rhinoconjunctivitis, ACS = Acute Coronary Syndromes, CHF = Chronic heart failure, COPD = Chronic obstructive pulmonary disease.

[b] Number of patients per arm when available ("n") or total number of patients in the trial ("N"). 5fu/lv = 5-fluorouracil/leucovorin, BFC = Combination of budesonide / formoterol, FP = Fluticasone propionate, SFC = Combination of Salmeterol / Fluticasone propionate.

reading the abstract, a further 351 titles were excluded for various reasons. Hand searching resulted in six more studies. A total of 21 papers were included after reading in full (table 8.2).[13,17-36]

8.3.2 Characteristics of studies

The studies come from a variety of disease areas, with eight papers on cardiovascular diseases, four on lung diseases and three on allergies. The other studies investigated interventions in Parkinson's disease, colon cancer, anesthesia, psoriasis vulgaris and hypertension. Three studies only used the trial data as a convenient data sample for exploring a new method of analysis.[13,29,31] Because they used a particular clinical trial to demonstrate how CEAs can be performed alongside RCTs in order to obtain country-specific estimates of CE, they were included in the review. One of these papers did not reveal the original RCT nor disease area, "so as not to conflict with previous publications".[31] The minimum number of countries in an RCT was two in the LIPID trial in patients with a history of acute coronary syndromes[25]; the maximum was 48 in the ExTRACT TIMI 25 study in patient with a myocardial infarction.[32] This last study also had the highest number of patients, namely 20,506. The lowest number of patients was 501 in the trial discussed by Peeters et al. in psoriasis vulgaris.[19]

8.3.3 Classification of studies

Measurement: resource use and health outcomes

Table 8.3 classifies the studies according to the sample of countries from which data on resource use and health outcomes were obtained. Most studies based both health care utilization and health outcomes on information from all countries. None of the studies was based on data from only one country, but three based their analysis on a subset of countries.[27,29,33] Bachert et al. only used data from the five North European countries, out of eight countries included in the RCT.[27] The study by Manca and colleagues excluded two unnamed countries with a very low number of patients and/or an extremely unequal allocation of patients between the two treatment groups in the trial.[29] Briggs and colleagues based their results on 21 of the 42 participating countries for which validated translations of the EuroQol-5D (EQ-5D) instrument were available. The CE analysis was therefore limited to 70% of the trial participants.[33] Weintraub et al. used all patients for their health outcomes, but only measures utility weights in patients from English speaking countries. For other patients, the average utility by treatment arm from the English speaking countries was used to estimate utility.[23] Although theoretically possible, none of the studies based the health care utilization on a different selection of countries than the health outcomes.

Valuation: unit costs and utility value sets

Table 8.3 also reports the valuation of health care utilization (unit costs) and health outcomes (utility value sets) for each of the studies. Twelve studies used a one-country valuation for health care utilization[13,17-19,22,23,25,26,30,32,34,35], six a multi-country valuation.[20,21,24,27,28,33] In 3 studies this was unclear.[29,31,36] In two studies this was because their focus was more on the analytical methodology, than on how data was gathered and the resulting CEA.[29,31] In the study by Rogkakou et al. it was not stated if every country has their own unit cost vector or if only one unit cost vector is used.[36]

In twelve of the studies, health outcomes were unweighted. The health outcomes in these papers were, for example, the number of avoided exacerbations[18,30], major cardiovascular events[35], patient-level survival[32] and event-free-days.[34] None of the studies used a multi-country valuation to obtain utilities, while four studies used a one-country valuation. All four of these studies used the recommended tariffs for the EQ 5D for the United Kingdom (UK)[37,38], although only two studies calculated country-specific CE results for the UK. Bachert et al. used these UK tariffs for calculating country-specific results for the UK, Germany, The Netherlands, Sweden, Denmark, Norway and Finland.[27] Willan and colleagues applied the UK tariffs on data from all countries to calculate the CE results for the UK, and, in a separate calculation, applied the Canadian tariffs when calculating CE results for Canada.[26] Canonica and colleagues used the UK tariffs to calculate CE results for Spain, France, Italy and Austria.[28] The study by Briggs and colleagues used the UK

Table 8.3: Transferability issues.

First author and year	Health care utilization		Health outcomes		Trial-wide CE result?	Country-specific CE results		Other CE results	Analysis type[c]
	Measurement[a]	Valuation[b]	Measurement[a]	Valuation[b]		Within RCT	Outside RCT		
Lindgren 2005 [17]	All	1C	All	UW	No	1 country	-	-	Simple
Lofdahl 2005 [18]	All	1C	All	UW	Yes	-	-	-	Simple
Peeters 2005 [19]	All	1C	All	UW	Yes	-	-	-	Simple
Pinto 2005 [13]	All	1C	All	?	No	1 country	-	-	Hier
Radeva 2005 [20]	All	MC	All	UW	Yes	-	-	-	Simple
Reed 2005 [21]	All	MC	All	UW[d]	Yes	-	-	-	Simple
Weintraub 2005a [22]	All	1C	All	?	Yes	-	-	-	Simple
Weintraub 2005b [23]	All	1C	All	UW	No	1 country	-	-	Simple
Briggs 2006 [24]	All	MC	All	?	No	1 country	-	-	FE
Tonkin 2006 [25]	All	1C	All	UW	No	2 countries	-	-	Simple
Willan 2006 [26]	All	1C	All	1C	No	2 countries	-	-	Hier
Bachert 2007 [27]	Subset	MC	Subset	1C	No	5 countries	2 countries	-	Simple
Canonica 2007 [28]	All	MC	All	1C	No	3 countries	1 country	-	Simple
Manca 2007 [29]	Subset	?	Subset	?	Yes	17 countries	-	-	Hier
Rutten-van Molken 2007 [30]	All	1C	All	UW	Yes	-	-	-	Simple
Willan 2008 [31]	All	?	All	?	Yes	1 country	-	-	Hier
Marcoff 2009 [32]	All	1C	All	UW	Yes	-	-	-	Hier
Briggs 2010 [33]	Subset	MC	Subset	1C	Yes	-	-	4 regions	FE
Gomes 2010 [34]	All	1C	All	UW	Yes	1 country	-	-	FE
Lorgelly 2010 [35]	All	1C	All	UW	Yes	-	-	-	Simple
Rogkakou 2011 [36]	All	?	All	UW	Yes	-	-	-	Simple

[a] 1C = information from 1 country included; subset = information from some countries excluded; All = information from all countries included

[b] 1C = One-country: weights from a single country; MC = Multi-country: each country has their own weights; UW = Unweighted; ? = Unknown.

[c] Simple = Simple, naïve analysis; FE = fixed effects model; Hier = Hierarchical model.

[d] Collected quality of life information in a subset of patients, but this data was not used for CEA and is therefore classified as unweighted.

weights to calculate CE results for four groups of countries called regions, of which one, Western Europe, included the UK.[33] In five studies, the valuation of health outcomes was not apparent from the text. In three studies this was because of their methodological focus.[13,29,31] In the study by Briggs et al., the authors applied a mapping algorithm to mat the IPD on the Asthma Quality of Life Questionnaire to the EQ-5D in order to obtain a utility estimate.[24] Since no further information about this algorithm was given and no reference could be found, it is unknown whether it is transferable across countries. In the study by Weintraub et al. the source of the utility valuation was not mentioned in the text.[23]

The valuation of the health care utilization and health outcomes was not consistent in three studies. They used a one-country valuation for health outcomes, the UK EQ-5D tariffs discussed above, and a multi-country valuation for health care utilization: unit costs from each individual country.[27,28,33]

8.3.4 Trial-wide and country-specific CE estimates

Thirteen studies calculated a trial-wide CE estimate, of which three also calculated country-specific CE estimates (table 8.3).[29,31,34] Briggs and colleagues calculated both trial-wide CE results, and CE results for four regions.[33] The remaining eight studies calculated only country-specific CE estimates. Two studies extrapolated the results beyond the countries included in the original trial.[27,28] They did this by applying country-specific unit costs of countries not participating in the trial to trial-wide resource use. However, as was mentioned before, these studies calculated QALYs by using the same utility value set (UK) for all countries.[27,28] Both Lindgren and colleagues and Willan and colleagues combined all available data and applied one-country valuation for health care utilization and health outcomes.[17,26[By performing the procedure twice, once using unit costs from one country and once from the other country, they could calculate country-specific results for two countries The study by Marcoff and colleagues mentioned estimates of country-specific CE in an online appendix, but these estimates could not be found in the appendix.[32]

Differences between country-specific CE results

When the same study publishes results for more than one country, it is informative to compare these results. Five studies calculated CE results for more than one country[25-29] and one for four regions.[33]

Willan and colleagues found that the incremental costs were positive in the Canadian setting and negative in the UK setting (intervention was dominant).[26] However, the confidence intervals were very wide and both symmetrically straddle the origin. They therefore concluded that there is very little evidence of a difference in costs between the two countries.

The studies by Bachert et al. and Canonica et al. are based on the same RCT about allergic rhinoconjunctivitis. Bachert and colleagues concluded that the ICER was similar

in every country, ranging between € 13,000 and € 18,000 per QALY gained.[27] Possible reasons for the differences in CE results given by the authors were differences in discount rates, unit costs and treatment costs. Canonica and colleagues calculated ICERs ranging from € 14,000 to € 22,000. They did not list potential reasons for these differences.[28] Neither of the two studies commented on the difference in ICERs between the two studies, nor on the difference in intervention costs, which was € 1,200 for all countries in Canonica et al. versus € 1,500 for all countries in Bachert et al.[27,28]

In the study by Manca and colleagues there are large differences in outcomes between the 17 countries.[29] Incremental costs ranged from approximately -UK£ 200 to +UK£ 400 and mean incremental survival ranged from approximately -100 days to +100 days. The size of the confidence intervals also differed significantly between the countries, as did the shape of the scatter plot in the CE planes and the CE acceptability curves.

Briggs and colleagues calculated regional ICERs for two treatment comparisons in COPD.[33] In the first comparison, the region-specific ICERs ranged from US$ 21,500 to US$ 77,100, with a trial-wide estimate of US$ 43,600. The second comparison showed a range of region-specific ICERs from US$ 13,200 to US$ 46,300, with a trial-wide estimate of US$ 26,500. In both comparisons, the ICER in the USA was by far the highest estimate. The authors mentioned in their discussion that this reflects the higher unit costs in the USA, although it was also clear from the results that the incremental QALYs were also lowest in the USA, which would also have contributed.

8.3.5 Heterogeneity

Differences in CE results between countries can be partly explained by heterogeneity between the countries, which is the part of the variation, in addition to that accounted for by chance, that can be explained by local characteristics, such as average age or cost levels.[39] Only one of the studies formally tested for heterogeneity before calculating CE results: Briggs and colleagues used joint tests of significance for treatment-by-region interactions.[33,40] The only evidence of heterogeneity they found was in the costs of study medication. They made a post-hoc comparison of the combined estimate with region-specific estimates based on dividing the dataset in region-specific subsets. They could not reject homogeneity across regions, but the authors acknowledge that such an approach suffers from a lack of power, making negative test results difficult to interpret.

Six studies did not discuss the possibility of heterogeneity between countries at all.[19,20,25,28,35,36] Two studies addressed heterogeneity by performing subgroup analyses. Löfdahl and colleagues conducted a subgroup-analysis to investigate differences between European and non-European countries.[18] They assumed that healthcare delivery and utilization is somewhat more homogenous within Europe, compared with other continents. Gomes and colleagues performed a similar subgroup-analysis. They analyzed patients

from the UK and patients not from the UK, separately.[34] This did not change the CE results significantly.

Lindgren and colleagues expected no problems with averaging resource use across patients unless resource consumption differed markedly between countries. They did not detect such problems and concluded that combining countries is not likely to have biased their analyses. Reed and colleagues recognized that there are differences in practice patterns and unit costs between countries, but argued that the randomized design of the RCT would not bias the findings in either direction.[21] Weintraub and colleagues (a) stated that their costing approach, combining country-specific measurements of resource use with one-country valuation, does not fully account for possible differences in treatment practices and resource use between countries or health care systems.[22] However, they felt their method yielded unbiased results and should even reduce "unwanted variability". Weintraub and colleagues (b) stated that it is not possible to adequately account for variation in costs across countries, but using country-specific costs should have little effect on their results. Bachert and colleagues justified not checking for heterogeneity due to (assumed) "similarities in costs and healthcare systems" for North European countries.[27] Rutten-van Mölken and colleagues assumed that the relative treatment effect on outcome, the prevention of COPD exacerbations, is generalizable across countries although no test was done.[30] They acknowledged that this approach does not account for many of the differences between countries that may affect CE, since differences between countries in relative prices of resources may lead to a different mix of resource use. For example, in countries where specialist contacts are relatively expensive compared to GP contact, patients may be referred to specialists less often. Applying a single set of unit costs to multiple countries ignores the presence of such substitution effects. Briggs and colleagues used an indicator variable for the country of interest, the UK, in two of the estimated regression models. The coefficient showed that the resource use needed for the treatment of an asthma exacerbation was relatively low in the UK.[24]

Finally, there were five studies explicitly taking differences between countries into account in the analysis, as explained in the next section.[1,13,29,31,32] However, no formal test of heterogeneity was performed beforehand.

8.3.6 Statistical methods

Thirteen studies used relatively simple methods to calculate trial- and country-specific CE results; while eight studies used more advanced statistical methods. Of these, three studies used a fixed effects modeling approach and the remaining five studies explicitly took the hierarchical structure of the data into account.

Simple methods

Thirteen studies calculated a simple country-specific patient-level average of the resource use and health outcomes. The resource use was then multiplied by either the same unit costs for all countries (one-country valuation), or by country-specific unit costs (multi-country valuation). If applicable, the health outcomes were multiplied by the utility value sets, which could also be the same for all countries, or country-specific. Dividing the resulting mean country-specific costs per patient by the mean country-specific health outcomes per patient led to country-specific CE estimates. Eight studies used a one-country valuation[17,18,20,22,23,25,30,35], two a multi-country valuation[19,21] and two combined a multi-country valuation for health care utilization with a one-country valuation for health outcomes.[27,28] The method used by Rogkakou and colleagues is unknown, as the method of valuation of health care utilization is unknown.[36]

Fixed effects models

Briggs and colleagues used a series of statistical regression models for costs and health-related quality of life weights in their 2006 study, including an adjustment to the country of interest via an indicator variable.[24] In their 2010 study, Briggs and colleagues used a Weibull survival model and multivariate patient-level regression models for costs and EQ-5D preference data.[33] Explanatory variables included patient-level data and dummies for each region. Both studies estimated the models for costs and health outcomes separately. These models were then used to calculate incremental costs, incremental health outcomes and CE results. Gomes and colleagues explicitly modeled the correlation between costs and health outcomes, by estimating a system of regressions and assuming a correlation between the error terms. This is called SUR modeling, or Seemingly Unrelated Regression.[34] No specific adjustment for country was made.

Hierarchical models

Five studies took the hierarchical structure of health care data explicitly into account.[13,26,29,31,32] Despite the use of many different names –multilevel modeling with random intercepts, empirical Bayesian shrinkage estimation, bivariate hierarchical models- all these methods are variations of multilevel modelling (MLM).[7] All models were analyzed using Bayesian methods, although MLM can be estimated using Frequentist methods.[8,9,11]

The studies by Pinto and colleagues and Willan and colleagues calculated trial-wide CE results by including a random intercept for each country.[13,26] They used two normal distributions for both costs and effects. Marcoff and colleagues extended this method by also including a random slope for each country. They analyzed the incremental net health and monetary benefits.[32]

There was no correlation modeled between costs and effects in any of these three studies. In contrast, both Manca and colleagues and Willan and colleagues extended the

hierarchical model by explicitly linking costs and effects. They both used flexible specifications which treat the trial-wide mean costs and effects as fixed effects, but differ in the specification of the assumed country-specific random effects. The model of Manca et al. was based on O'Hagan et al.[29,41] and assumed a bivariate normal (BVN) distribution to model the interaction. Willan et al. extended the model by Nixon and Thompson[10], which does not necessarily assume a normal distribution, but may accommodate for example skewed distributions when necessary.[31] Costs and effects do not need to have the same distribution. In both methods, costs and effects are correlated across equations in the same way. Both models allow extension of the hierarchical structure by including explanatory variables at the country level, on top of patient level covariates.

8.4 DISCUSSION

In this paper, we looked at CEAs that were based on recent, large multinational RCTs. We categorized the studies by the methods used to calculate trial-wide and country-specific cost-effectiveness. What we found was that simple, naïve methods were still frequently used. Within these studies one-country valuation of the healthcare utilization of patients from multiple countries was the most applied method. Each of the simple methods has its own advantages and disadvantages. An analysis that only uses the data from the country of interest, provides an unbiased country-specific estimator, but ignores all the information we have about the other countries. In addition, it requires enough patients within the country of interest to counter the risk of low statistical power. This defeats the purpose of doing a multinational RCT. A one-country valuation that uses one set of unit costs multiplied with trial-wide health care utilization has the advantage of keeping the statistical power. When different sets of unit costs from different countries are repeatedly combined with trial-wide resource use, we get insight in the influence of different absolute and relative prices on CE. However, differences in treatment patterns between countries are ignored and applying the unit costs from one country to trial-wide resource use could confound 'price effects' with 'country effects'.[14] A multi-country valuation accounts for differences between countries in both unit costs (or utility value sets) and treatment patterns. When averaging over all countries, the resulting point estimate is an accurate representation of the average in the trial, but it is difficult to interpret and generalize, as it is not representative for any of the countries.

Regardless of whether the investigator is interested in trial-wide or country-specific results, not taking the interaction between country and treatment effect on health care utilization and health outcomes into account may lead to wrong conclusions. Several models have been proposed that take this interaction into account. They can be grouped into fixed effects models and hierarchical models. The key advantage is that they improve

statistical efficiency by 'borrowing' information from all available data in the estimation of the difference between treatments for an individual country.

The effect of the interaction between country and treatment on costs, due to the differences in for example epidemiology, practice patterns and prices, is widely recognized.[42-44] At the same time, most models assume that the clinical effectiveness does not differ greatly between countries. However, many factors could potentially affect the between-location variability of health outcomes, such as the availability of health care services, local treatment guidelines, and differences in quality of health care.[29] Because of this, the between-country variability in differential health outcomes can be even greater than the variability in differential costs.[13] Allowing for variation across countries in both costs and health outcomes is, in our view, a preferred strategy.

In fixed effects models, applied by 3 studies in this review, country-level covariates are used to model the country-specific differences. These covariates may just be a series of simple country-dummies, but can also measure differences more explicitly with variables that measure differences in epidemiology, medical practice and economic factors. The precise formulation of the model can differ widely, ranging in this review from simple ordinary least squares regression and survival analysis to SUR modeling. Fixed effects models are typically included to "control for" differences across countries. Because of this, Drummond et al. argue that they might not be adequate to produce country-specific results.[16]

Hierarchical modeling, applied by 5 studies included in this review, lowers the variability of the country-specific CE results, by borrowing strength from other countries. Compared to the observed difference in country means, Pinto et al. achieved a 21–59 per cent reduction in average standard error of the difference. For one of the included countries, achieving this gain in precision using only country-specific data would have required more than twice as many patients.[13] Hierarchical models also lead to more appropriate estimates of population average results and associated standard errors compared to other methods.[11,12,14] For example, Grieve et al. modeled both length of stay and total costs of stroke admissions, in an observational study across 11 countries. They compared MLM with simple OLS and showed that the OLS analysis severely overestimated the precision of centre-level associations and made incorrect inferences.[11] In particular, the OLS analyses found that centre-level variables were associated with resource use, whereas the MLM analysis showed that, once the hierarchical nature of the data was recognized, none of these variables predicted resource use.

Manca et al. have recently compared advanced statistical methods for CE estimation, alongside multinational trials.[15] They conclude that Bayesian hierarchical models, using both patient- and country-level information, are the most appropriate tool to analyze CE alongside a multinational trial. This recommendation is based on the flexibility in facilitating the inclusion of patient- and cluster-level explanatory variables, and the ability

to accommodate distributions when costs and/or effects are not normally distributed. The Bayesian methodology allows direct interpretation of the country-specific (posterior) mean estimates and it provides probability statements regarding the CE in any given country.

In MLM, the country-specific estimates are latent variables, which must be quantified rather than estimated. This is achieved using shrinkage estimation, which is a weighted sum of the country-specific observed difference (one-country analysis) and the estimates provided by the MLM trial-wide estimate. It implements the idea that although different, country-specific data might share some degree of similarity and therefore contain information that is usable for all countries. The term 'shrinkage' reflects that the country-specific estimate for all countries will be closer ('shrunken') to the combined estimate, than the observed difference is. The degree of shrinkage depends on the between- and within-country variance, with more shrinkage occurring when the within-country variability is greater relative to the between-country variability.

In addition to the statistical advantages, Grieve and colleagues propose that hierarchical models may be used to assist in the design of multinational RCTs, which often only measure costs for a subsample of centers.[11] The choice of centers to collect costs is usually based on pragmatic grounds, but the factors identified as being associated with total costs from hierarchical models could be used to choose where best to measure costs. As an example, if the level of health care spending as a percentage of GDP is associated with total or incremental costs, centers could be selected which were broadly representative of countries with high, medium or low levels of spending on health care.[11]

A limitation to MLM is that it is assumed that countries are "exchangeable". This means that there are no a priori reasons to assume that one country has higher or lower health outcomes or costs, and can be represented by the same variance across all countries. Since MLM are often used to explore these differences, making this a priori assumption may be unreasonable.[16] Adding fixed effects on country-level may solve this.[45]

Despite the availability of these fixed effects and hierarchical models and demonstrated increased precision of the estimates[11-15], these newer methods have not yet been applied on a wide scale amongst researchers, nor have they replaced the simpler, naïve methods. Studies that did use more advanced methods were often studies with a theoretical, statistical point of view focused on the (illustration of) methods.One of the reasons for not using the advanced models might be found in the relative complexity. The resulting lack of transparency may limit their use in decision making.[46] Especially the hierarchical models require an advanced knowledge of (Bayesian) statistics and programming. Moreover, if these models demonstrate a difference in effectiveness of a treatment between countries, it may be very problematic to include this in a reimbursement dossier, especially for the country in which results differ from the drug's label claim.

Another drawback of the more advanced methods is the need of a large number of patients per country. Because one of our inclusion criteria was a minimum number of 500

patients, all studies in this review could have applied the more advanced methods. This is illustrated by the study of Willan et al.[26], which had a total of 541 patients and applied hierarchical models. The smallest study in the sample, the one by Tonkin[25], applied a simple method, but still had more patients in each arm than Willan in the placebo arm (179).[26] On the other hand, one of the largest studies, the study by Reed et al. with more than 14,500 patients in three arms, was analyzed by a simple method.[21] It is unclear why this dataset was not analyzed by a more advanced method and whether it would have changed the results. In any case it would have increase the precision of the estimates.

Unfamiliarity with the methods may be a reason why more advanced methods are not used. Although the ISPOR Good Research Practices Task Force on transferability recommends the advanced methods[16], guidance on the analysis of IPD from multinational trials, and the calculation of country-specific CE estimates, in national guidelines is scarce.[47] The Canadian pharmacoeconomic (PE) guidelines mention the advanced methods discussed in this paper as a possible option to analyze IPD from multinational trials, but do not explicitly recommend them.[48] The other PE guidelines and recommendations found on the ISPOR website do not suggest way to calculate country-specific CE estimates from multinational RCTs.[47] Multilevel models may not be necessary when only a few countries are included in a trial. Drummond et al. suggest 4 or 5 as a lower bound, suggesting that all RCTs in this review but one[25] could have been be analyzed by more advanced methods.[16]

Heterogeneity between countries can be tested formally, which was done in only one study, based on the method proposed by Cook and colleagues.[33,40] Fixed effects models typically contain several interaction terms to investigate whether the treatments differ between countries. If these interaction terms are statistically significant, there is heterogeneity between the countries. On the other hand, if they are not statistically significant, and there is sufficient statistical power, one may conclude that the effect of the treatment is not different between countries. In this case, the treatment effect may best be estimated for all countries together.[40] Unfortunately, low power within a single country is a common problem in multinational trials. One way of handling this problem is using a higher level of significance, or combining several countries with similar characteristics.[53]

Another formal test, based on the likelihood ratio, was proposed by Gail and Simon.[54] They distinguish between qualitative interaction, which occurs when the treatment effect is positive for the patients in some countries and negative for those in other countries, and quantitative interaction, which occurs when the magnitude but not the direction of treatment effects varies. Other examples of formal tests for qualitative interactions include the range test proposed by Piantadosi and Gail, and a simple test based on simultaneous confidence intervals proposed by Pan and Wolfe.[55,56]

When policy makers are judging the results of a study in one country for applicability in their own country, one of the first things they compare is the characteristics of the study population. This information makes it easier to interpret CE results and to determine to

what extent the results are valid for the target population in their country. However, only ten of the 21 papers produced a table showing demographic characteristics of patients in the original trial.[20,22,23,25,27,28,30,32,33,35] Even if this information is included in the clinical paper on the same study, it is relevant for policy makers to repeat this information in the cost-effectiveness paper.[57] Other important information for policy makers is the valuation of health care utilization. Only eight papers specified the unit costs used in the study.[17,18,22,26-28,30,34] Some studies did not list total health care utilization and/or health outcomes by treatment arm, which is considered to be basic information for a CEA.[18,21,26-28,31]

Calculating country-specific CE results for many countries may lead to challenges in presentation. This is apparent in the study by Manca and colleagues, which shows results for 17 countries.[29] They improved the readability of the results by choosing to show 95% confidence ellipses and CE acceptability curves, for a selection of countries. They also refrained from presenting point estimates. However, if point estimates and results for all countries are of primary importance, these may not be viable solutions and other ways of improving readability need to be explored.

8.5 CONCLUSION

Several advanced statistical techniques are available to calculate country-specific CE results from multinational trials. These methods take the interaction between country and treatment effect on health and health care utilization into account. Hierarchical models also lower variability of the country-specific CE results and lead to more appropriate estimates of population average results and associated standard errors. However, they have not been used on a wide scale yet, while simpler, naïve methods are still routinely employed. This should change in future.

8.6 LITERATURE

[1] Willan AR, Pinto EM, O'Brien BJ, Kaul P, Goeree R, Lynd L, et al. Country specific cost comparisons from multinational clinical trials using empirical Bayesian shrinkage estimation: the Canadian ASSENT-3 economic analysis. Health economics 2005 Apr;14(4):327-38.

[2] Barbieri MA, Drummond MF, Willke R, Chancellor JVM, Jolain B, Towse A. Variability of Cost-Effectiveness Estimates for Pharmaceuticals in Western Europe: Lessons for Inferring Generalizability. Value in Health 2005;8(1):10-23.

[3] Rutten-van Mölken MPMH, Vemer P. Internationale vertaalbaarheid van kosten-effectiviteit [International transferability of cost-effectiveness]. Van Kosten tot Effecten Een handleiding voor evaluatiestudies in de gezondheidszorg [From Costs to Effects. A manual for evaluation studies in health care]. 2nd ed. Maarssen: ELSEVIER gezondheidszorg; 2010. p. 270.

[4] Willke RJ, Glick HA, Polsky D, Schulman K. Estimating country-specific cost-effectiveness from multinational clinical trials. Health economics 1998 Sep;7(6):481-93.

[5] Koopmanschap MA, Touw KC, Rutten FFH. Analysis of costs and cost-effectiveness in multinational trials. Health Policy 2001 Nov;58(2):175-86.

[6] Clarke PM, Glasziou P, Patel A, Chalmers J, Woodward M, Harrap SB, et al. Event rates, hospital utilization, and costs associated with major complications of diabetes: a multicountry comparative analysis. PLoS Med 2010 Feb 23;7(2):e1000236.

[7] Goldstein H. Multilevel Statistical Models. 2nd ed. London: Edward Arnold; 1995.

[8] Rice N, Jones A. Multilevel models and health economics. Health Econ 1997 Nov-Dec;6(6):561-575.

[9] Carey K. A multilevel modelling approach to analysis of patient costs under managed care. Health Econ 2000 Jul;9(5):435-446.

[10] Nixon RM, Thompson SG. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. Health Econ 2005 Dec;14(12):1217-29.

[11] Grieve R, Nixon R, Thompson SG, Normand C. Using multilevel models for assessing the variability of multinational resource use and cost data. Health Econ 2005 Feb;14(2):185-96.

[12] Manca A, Rice N, Sculpher MJ, Briggs AH. Assessing generalisability by location in trial-based cost-effectiveness analysis: the use of multilevel models. Health economics 2005 May;14(5):471-85.

[13] Pinto EM, Willan AR, O'Brien BJ. Cost-effectiveness analysis for multinational clinical trials. Stat Med 2005 Jul 15;24(13):1965-1982.

[14] Gauthier A, Manca A, Anton S. Bayesian modelling of healthcare resource use in multinational randomized clinical trials. Pharmacoeconomics 2009;27(12):1017-1029.

[15] Manca A, Sculpher MJ, Goeree R. The analysis of multinational cost-effectiveness data for reimbursement decisions: a critical appraisal of recent methodological developments. Pharmacoeconomics 2010;28(12):1079-1096.

[16] Drummond M, Barbieri M, Cook J, Glick H, Lis J, Malik F, et al. Transferability of Economic Evaluations Across Jurisdictions: ISPOR Good Research Practices Task Force Report. Value in Health 2009;12(4):409-18.

[17] Lindgren P, Buxton M, Kahan T, Poulter NR, Dahlof B, Sever PS, et al. Cost-effectiveness of atorvastatin for the prevention of coronary and stroke events: an economic analysis of the Anglo-Scandinavian Cardiac Outcomes Trial--lipid-lowering arm (ASCOT-LLA). Eur J Cardiovasc Prev Rehabil 2005 Feb;12(1):29-36.

[18] Löfdahl CG, Ericsson A, Svensson K, Andreasson E. Cost effectiveness of budesonide/formoterol in a single inhaler for COPD compared with each monocomponent used alone. Pharmacoeconomics 2005;23(4):365-75.

[19] Peeters P, Ortonne JP, Sitbon R, Guignard E. Cost-effectiveness of once-daily treatment with calcipotriol/betamethasone dipropionate followed by calcipotriol alone compared with tacalcitol in the treatment of Psoriasis vulgaris. Dermatology 2005;211(2):139-145.

[20] Radeva JI, Reed SD, Kalo Z, Kauf TL, Cantu E,3rd, Cretin N, et al. Economic evaluation of everolimus vs. azathioprine at one year after de novo heart transplantation. Clin Transplant 2005 Feb;19(1):122-129.

[21] Reed SD, Radeva JI, Weinfurt KP, McMurray JJ, Pfeffer MA, Velazquez EJ, et al. Resource use, costs, and quality of life among patients in the multinational Valsartan in Acute Myocardial Infarction Trial (VALIANT). Am Heart J 2005 Aug;150(2):323-329.

[22] Weintraub WS, Mahoney EM, Lamy A, Culler S, Yuan Y, Caro J, et al. Long-term cost-effectiveness of clopidogrel given for up to one year in patients with acute coronary syndromes without ST-segment elevation. J Am Coll Cardiol 2005 Mar 15;45(6):838-845.

[23] Weintraub WS, Zhang Z, Mahoney EM, Kolm P, Spertus JA, Caro J, et al. Cost-effectiveness of eplerenone compared with placebo in patients with myocardial infarction complicated by left ventricular dysfunction and heart failure. Circulation 2005 Mar 8;111(9):1106-1113.

[24] Briggs AH, Bousquet J, Wallace MV, Busse WW, Clark TJ, Pedersen SE, et al. Cost-effectiveness of asthma control: an economic appraisal of the GOAL study. Allergy 2006 May;61(5):531-536.

[25] Tonkin AM, Eckermann S, White H, Friedlander D, Glasziou P, Magnus P, et al. Cost-effectiveness of cholesterol-lowering therapy with pravastatin in patients with previous acute coronary syndromes aged 65 to 74 years compared with younger patients: results from the LIPID study. Am Heart J 2006 Jun;151(6):1305-1312.

[26] Willan AR, Goeree R, Pullenayegum EM, McBurney C, Blackhouse G. Economic evaluation of rivastigmine in patients with Parkinson's disease dementia. Pharmacoeconomics 2006;24(1):93-106.

[27] Bachert C, Vestenbaek U, Christensen J, Griffiths UK, Poulsen PB. Cost-effectiveness of grass allergen tablet (GRAZAX) for the prevention of seasonal grass pollen induced rhinoconjunctivitis - a Northern European perspective. Clin Exp Allergy 2007 May;37(5):772-779.

[28] Canonica GW, Poulsen PB, Vestenbaek U. Cost-effectiveness of GRAZAX for prevention of grass pollen induced rhinoconjunctivitis in Southern Europe. Respir Med 2007 Sep;101(9): 1885-1894.

[29] Manca A, Lambert PC, Sculpher M, Rice N. Cost-effectiveness analysis using data from multinational trials: the use of bivariate hierarchical modeling. Med Decis Making 2007 Jul-Aug; 27(4):471-90.

[30] Rutten-van Mölken MPMH, van Nooten FE, Lindemann M, Caeser M, Calverley PM. A 1-year prospective cost-effectiveness analysis of roflumilast for the treatment of patients with severe chronic obstructive pulmonary disease. Pharmacoeconomics 2007;25(8):695-711.

[31] Willan AR, Kowgier ME. Cost-effectiveness analysis of a multinational RCT with a binary measure of effectiveness and an interacting covariate. Health Econ 2008 Jul;17(7):777-91.

[32] Marcoff L, Zhang Z, Zhang W, Ewen E, Jurkovitz C, Leguet P, et al. Cost effectiveness of enoxaparin in acute ST-segment elevation myocardial infarction: the ExTRACT-TIMI 25 (Enoxaparin and Thrombolysis Reperfusion for Acute Myocardial Infarction Treatment-Thrombolysis In Myocardial Infarction 25) study. J Am Coll Cardiol 2009 Sep 29;54(14):1271-1279.

[33] Briggs AH, Glick HA, Lozano-Ortega G, Spencer M, Calverley PM, Jones PW, et al. Is treatment with ICS and LABA cost-effective for COPD? Multinational economic analysis of the TORCH study. Eur Respir J 2010 Mar;35(3):532-539.

[34] Gomes M, Soares MO, Dumville JC, Lewis SC, Torgerson DJ, Bodenham AR, et al. Cost-effectiveness analysis of general anaesthesia versus local anaesthesia for carotid surgery (GALA Trial). Br J Surg 2010 Aug;97(8):1218-1225.

[35] Lorgelly PK, Briggs AH, Wedel H, Dunselman P, Hjalmarson A, Kjekshus J, et al. An economic evaluation of rosuvastatin treatment in systolic heart failure: evidence from the CORONA trial. Eur J Heart Fail 2010 Jan;12(1):66-74.

[36] Rogkakou A, Villa E, Garelli V, Canonica GW. Persistent allergic rhinitis and the XPERT study. World Allergy Organ J 2011 /;4(SUPPL. 3):S32; S36.

[37] Dolan P. Modeling valuations for the EuroQol health states. Med Care 1997;35:1095-108.

[38] Drummond M, Sulpher M, Torrance G, O'Brien B, Stoddart G. Methods for the Economic Evaluation of Health care Programmes. 3rd ed. Oxford: Oxford University Press; 2005.

[39] Briggs A, Scuplher M, Claxton K. Decision Modelling for Health Economic Evaluation. Oxford: Oxford University Press; 2006.

[40] Cook JR, Drummond MF, Glick H, Heyse JF. Assessing the appropriateness of combining economic data from multinational clinical trials. Stat Med 2003 Jun 30;22(12):1955-76.

[41] O'Hagan A, Stevens JW, Montmartin J. Bayesian cost-effectiveness analysis from clinical trial data. Stat Med 2001 Mar 15;20(5):733-753.

[42] Drummond MF, Bloom BS, Carrin G, Hillman AL, Hutchings AC, Knill-Jones RP, et al. Issues in the cross-national assessment of health technology. Int J Technol Assess Health Care 1992; 8(4):671-82.

[43] Drummond MF, Pang F. Transferability of economic evaluation results. In: Drummond MF, McGuire A, editors. Economic evaluation in health care; merging theory with practice. 1st ed. Oxford: Oxford University Press; 2001.

[44] Pang F. Design, analysis and presentation of multinational economic studies: the need for guidance. Pharmacoeconomics 2002;20(2):75-90.

[45] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. New York: Chapman & Hall/CRC; 2004.

[46] Reed SD, Anstrom KJ, Bakhai A, Briggs AH, Califf RM, Cohen DJ, et al. Conducting economic evaluations alongside multinational clinical trials: toward a research consensus. Am Heart J 2005 Mar;149(3):434-43.

[47] ISPOR. Pharmacoeconomic Guidelines from around the world. 2013; Available at: http://www.ispor.org/PEguidelines/index.asp. Accessed March, 2013.

[48] Canadian Agency for Drugs and Technologies in Health. Guidelines for the economic evaluation of health technologies: Canada. 3rd ed. Ottowa: Canadian Coordinating Office for Health Technology Assessment (CCOHTA); 2006.

[49] Badia X, Roset M, Herdman M, Kind P. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. Med Decis Making 2001 Jan-Feb; 21(1):7-16.

[50] Greiner W, Claes C, Busschbach J, Graf von der Schulenburg J. Validating the EQ-5D with time trade off for the German population. European Journal of Health Economics 2004;[Epub ahead of print].

[51] Lamers LM, McDonnell J, Stalmeier PF, Krabbe PF, Busschbach JJ. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ 2006 Oct;15(10):1121-32.

[52] Knies S, Evers SM, Candel MJ, Severens JL, Ament AJ. Utilities of the EQ-5D: transferable or not? Pharmacoeconomics 2009;27(9):767-779.

[53] Glick H, Doshi J, Sonnad S, Polsky D. Economic Evaluation in Clinical Trials. New York: Oxford University Press; 2007.

[54] Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics 1985 Jun;41(2):361-72.

[55] Piantadosi S, Gail MH. A comparison of the power of two tests for qualitative interactions. Stat Med 1993 Jul 15;12(13):1239-1248.

[56] Pan G, Wolfe DA. Test for qualitative interaction of clinical significance. Statistics in medicine 1997 Jul 30;16(14):1645-52.

[57] Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) Statement. Pharmacoeconomics 2013 Mar 26.

Chapter 9

Discussion

Further issues in HTA

9.1 INTRODUCTION

After introducing the (simple) concept of health technology assessment (HTA) in chapter 1, chapter 2 showed a practical application of health economic (HE) modelling, by calculating the cost-effectiveness (CE) of the reimbursement of aids in smoking-cessation. The following chapters explored several types of differences, and how these differences may be handled. Chapter 3 discussed differences between patients, which can be attributed to patient characteristics, called patient heterogeneity. Chapters 4 and 5 discussed different data sources, and how they can be combined using meta-analysis. Chapters 6, 7 and 8 explored issues that occur in decision making, when dealing with several different countries. These issues fall within the field of health-economics called transferability, which tries to tackle these differences and provide useful information for the decision maker.

9.2 PRACTICAL APPLICATION: SMOKING CESSATION

Chapter 2 explored the long term societal effects of reimbursement of smoking cessation support (SCS). We found that reimbursement led to more successful quitters and a gain in life years and QALYs. Accounting for overhead, administration costs and the costs of SCS, these health gains could be obtained at relatively low cost, even when including costs in life years gained. Hence, reimbursement of SCS seems to be cost-effective in the long term from a health care perspective.

The discussion on the reimbursement of smoking cessation has known many stages in The Netherlands. At the moment the study was performed, during the summer of 2008, SCS was only partly reimbursed and pharmacological SCS was not reimbursed at all. Tobacco control policy in The Netherlands at that time aimed to reduce smoking prevalence to 20% in 2010.[1] Smoking prevalence was declining, but additional efforts were required to reach the goal. In 2007 28% of the Dutch population still smoked.[2] The health insurance board (CVZ) had advised the Dutch Ministry of Health to reimburse an integrated smoking cessation programme, consisting of a combination of behavioural counselling and pharmacotherapy.[3]

A randomized controlled trial (RCT) was performed to investigate the effects of such a reimbursement policy in the Dutch region of Friesland in May 2002.[4,5] This study was funded by the Dutch Ministry of Health and the then Minister of Health, Mrs. Borst, expressed a willingness to start reimbursement, once effectiveness was shown. The trial found that the number of participants using SCS was higher in the intervention group than in the control group.[4] It also showed that the intervention would be cost-effective, in the short term. Unfortunately, when the results were presented, the new Minister of Health,

Mr. Hoogervorst, had to reduce costs in the basic insured package. He decided not to reimburse SCS in 2003.

A pilot study investigated the feasibility of large-scale implementation of reimbursement in The Netherlands in 2008. This so-called Agis Study, performed in the Dutch province of Utrecht, looked in particular to the availability and accessibility of care, and attainability of the program.[6] At the end of the test period, a third of respondents said they had stopped smoking. Based on insurance declarations, the estimated costs of nation-wide SCS reimbursement are between €14.0 en €22.7 million. Expanding on this study, Over et al. concluded that reimbursement of SCS produced overall health gains, but did not reduce health disparities between different socioeconomic groups.[7]

The goal to reduce smoking prevalence to 20% in 2010 was not met. In 2011, SCS programs were reimbursed leading to a much higher use of SCS than in the years before and an unprecedented drop in smokers from 27.2% to 24.7%. As a recent population study concluded, full health insurance coverage for smoking cessation treatment in The Netherlands was accompanied by a significant increase in the number of (dispensed) prescriptions of stop-smoking medication and a decrease in smoking prevalence.[8] In 2012, this was changed again, with a stop of the reimbursement of nicotine replacement therapies (NRTs) and pharmacotherapy. Behavioural counselling was still reimbursed. This immediately led to a drop in the use of effective SCS, a drop in the number of SCS prescriptions of 21.6 per 1,000 smokers, a rise in smoking prevalence by 1.2% and pressure on political actors by GPs and medical specialists to reinstate reimbursement.[8-11] As of 2013, NRTs and pharmacotherapies are now reimbursed for a maximum of one attempt per calendar year, provided they are accompanied by behavioural counselling.

Public health, personal health and health-economic arguments all point in the direction that SCS should be reimbursed. This includes the results in chapters 2 and 6. In chapter 2, the long-term incremental CE ratio of the reimbursement for NRT, bupropion and behavioural counselling, compared to no reimbursement, is below €16,000 per QALY gained. In chapter 6, varenicline to aid in smoking cessation dominates buproprion in The Netherlands. However, these arguments do not seem to be enough to keep reimbursement of SCS in place.

The choice whether SCS are reimbursed in The Netherlands seems to be more a political debate on "life style" medications. This also includes contraceptives, erectile dysfunction medication and cholesterol lowering medication, for which reimbursement also changes from year to year. This discussion centers on the amount of responsibility that can be laid at the feet of Dutch citizens for their own life style and its consequences. Another issue working against the health(-economic) arguments, seems to be the influence of the tobacco industry lobby in both The Netherlands[12-14] and wider Europe.[15,16]

A separate methodological issue that is addressed in chapter 2 is the inclusion of costs in life-years gained. In the extra life years gained by successful quitters, additional costs

are generated for diseases unrelated to smoking, such as dementia or hip fractures in old age.[17] This makes CE estimates both more conservative and more accurate. In most CE-studies published these costs are not taken into account (for example chapter 3), even though the inclusion of these costs is now facilitated by an online tool called Practical Application to Include Disease costs (PAID).[18]

9.3 DIFFERENT PATIENTS

Chapter 3 discussed the difference between patient heterogeneity and parameter uncertainty. We compared four ways of dealing with heterogeneity and showed that they led to widely different results in CE. Three of the methods can all be viable options, depending on the decision makers' information need. When little or no heterogeneity is expected, or when it is not expected to influence the CE results, disregarding heterogeneity may be correct. Subgroup analyses may inform policy decisions on each subgroup, as long as they are well defined and the characteristics of the cohort that define a subgroup truly represent the patients within that subgroup. Despite the necessary calculation time, the Double Loop PSA is a viable alternative which leads to better results and better policy decisions, when accounting for heterogeneity in a Markov model.

The final method draws from all available distributions at the same time: probability distributions that reflect parameter uncertainty and frequency distributions of patient characteristics. The expected outcome of this analysis reflects parameter uncertainty and patient heterogeneity in a heterogeneous population, but ignores the fundamental difference between the two. We have shown that this Single Loop PSA should not be used in CE research. It disregards the fundamental differences between heterogeneity and sampling uncertainty and overestimates uncertainty as a result.

9.4 DIFFERENT DATA SOURCES

In chapters 4 and 5, we discussed the combining of different sources of information using meta-analysis. In chapter 4 we compared four methods of direct meta-analysis and found that three of these methods lead to comparable HE outcomes, while the Bayesian random effects methods tends to overestimate uncertainty. Based on this study, we recommended using the frequentist random effects method proposed by DerSimonian and Laird as the preferred method of meta-analysis.[19] It automatically reduces to a fixed effects model in the absence of heterogeneity. Compared to the Bayesian methods it is easier to implement and more easily understood by physicians and policy makers who will be using the results. In chapter 5 we compared four methods of network meta-analysis. The method proposed

by Puhan and the Bayesian fixed effects generalized fixed effects model are preferred.[20-22] The method proposed by Song has slightly less preferable characteristics, while the Bayesian random effects generalized model overestimated uncertainty and had shown large biases and absolute deviations.

One of the major issues in meta-analysis is how different sources of evidence can be combined. This is a particular issue with a program like 'expensive drugs', where only little evidence is available at the start of conditional reimbursement, and usually the only additional evidence available after the conditional reimbursement period is evidence from use in real world observational studies. Since the decision to continue reimbursement can have a major impact, both in a financial and medical sense, it should be based on all available evidence.

For many researchers, the randomized controlled trial (RCT) is the gold standard, because of its rigorous evaluation design. This would indicate that CE results should be based on RCTs alone. However, the strict protocols in RCTs do not generally reflect clinical practice. Because of this, other researchers may prefer observational evidence alone. However, due the lack of a rigorous design, there is a possibility that the efficacy outcomes cannot be directly interpreted, and might be biased. In either situation, valuable information may be left unused.

If it is decided that both sources of evidence can be combined, a weight needs to be defined to combine all sources of evidence. Either one of the sources may get a relatively higher weight than other evidence. In the hierarchy of evidence framework, an RCT produces stronger evidence than an observational study.[23] If the weighing of the evidence in a meta-analysis would be based on this hierarchy, the strength of evidence from observational studies would be graded relatively low compared to RCTs. However, the results of well-designed observational studies do not necessarily contain bias in the magnitude of the effects of treatment.[24] This would argue for the same weights for observational studies and RCTs, provided the observational studies are well-designed. In some instances, observational studies may provide better evidence, for example in the case of rare events. Moreover, the result of an RCT may not be applicable at all, for example if the patients are highly selected or motivated relative to the population of interest.[25] In these cases, one might argue for giving observational studies a higher weight than RCTs.

Even if the results of the real life outcomes study would be treated the same as the results of an RCT, a meta-analysis of the evidence is still likely to be driven by trial results, since the uncertainty around the trial-based estimates of the treatment effect is likely to be smaller due to the generally homogeneous nature of included patients. Moreover multiple RCTs are often available, compared to one, perhaps relatively small, real life outcomes study. In order to take other issues than study design into account, the GRADE Working Group proposed a systematic and explicit method of making judgments on the quality of evidence under consideration.[25] Next to study design, the other three key elements they

named in reviewing available evidence are study quality, consistency and directness. On each of these elements, each source of evidence is graded High, Moderate, Low or Very Low. Using this grading system, different sources may be combined qualitatively, ranking the trade-offs between health benefits and harms before considering costs. This methodology can provide insight in the way different types of evidence can be weighted.

Once an appropriate weight has been determined, this weight will have to be incorporated into the meta-analysis. As most of the frequentist methods are based on the Inverse-Variance (IV) approach, where studies with a low standard error get a larger weight, it is logical to apply the weights directly to the standard error of the weights. This also works for the Bayesian methods, which use standard errors as a proxy for the strength of evidence, except when the link function is discrete (binomial, multinomial, Dirichlet). In this case, the weights can be applied directly to both the number of patients in the trial and the number of "successes" (transitions, events, etc).

The question remains how big a certain weight should be. Since every choice of weight would be arbitrary, it is recommended to perform the same analysis several times, for different types of weights and comparing the results.[26] In addition, it is recommended not using higher weights for sources of evidence with high quality, but using lower weights for sources of evidence with low quality instead.[26] Artificially giving a trial a higher weight would imply a larger number of patients than are really available. In essence, this means "inventing patients" and artificially "adding certainty" about the newly synthesized parameter. On the other hand, giving a trial a lower weight as is recommended would imply fewer patients than were really there, thus reflecting the extra uncertainty due to the quality of evidence.

This issue of combining different sources of evidence is still being discussed. In open debates at HTA conferences, for example during the ISPOR 16th Annual European Meeting (2-6 Nov 2013, Dublin), it is now being suggested that these two types of data should not be combined at all. The reasoning is that both forms of evidence provide information for two different policy questions: the RCT on whether the new intervention is at all better than what it is compared with (efficacy), and the real life data whether this better efficacy translates to better results in the real world. Combining these two forms of information is deemed unlikely to answer either question.

9.5 DIFFERENT STATISTICAL PARADIGMS

In chapters 4 and 5, a simulation study is described, which compares several methods of meta-analysis. In these chapters, the two existing statistical paradigms, the frequentist and the Bayesian approaches, meet. These two paradigms differ in the way they approach inference. If one wants to estimate the value of an unknown parameter, a likelihood func-

tion can be defined based on the data. This likelihood function is used in both approaches, albeit in different ways. The frequentist approach is to maximise the likelihood function over all possible parameter values to obtain the maximum likelihood estimate (mle). Asymptotically, as the information available increases (usually sample size) the distribution of the mle tends towards a normal distribution. Population means, standard errors and confidence intervals can then be estimated. The Bayesian approach is to define a prior distribution, which is a summary of the knowledge the statistician has before starting the analysis. This prior knowledge can be non-existent ("any value is equally likely"), somewhat informative ("it is centered around 0"; "the probability must lie between 0 and 1") or very informative ("previous research indicates a point estimate of 4.55, with a 95% credibility interval between 4.50 and 4.62"). These (subjective) prior probabilities are then combined with the (observed) frequency probabilities, which are summarized in the likelihood function. This forms a posterior distribution, with a population mean and credibility interval.

Normally, statisticians are proponents of one or the other paradigm, which directs the kind of analysis they perform. As Crowder puts it: "Some of the more vocal proponents of the different approaches to inference have been shouting at each other for years from their respective hilltops."[27] Therefore, a study comparing frequentist and Bayesian methods is not found very often, even if both methods have their own merits and drawbacks. For example, the estimated outcomes of a frequentist exercise are always (implicitly) normally distributed, while Bayesian outcomes may have every possible shape. The frequentist approach says the parameter we want to estimate is unknown, but fixed. The Bayesian approach treats the unknown parameter as a random variable, even if we know it's a fixed number, for example the distance Groningen-Rotterdam (although this distance seems to be shorter than the distance Rotterdam-Groningen).

The biggest criticism of the frequentist approach is often about the p-value. It is interpreted as the probability that the hypothesis is correct, given the data. However, the p-value derives from the likelihood of the data given the hypothesis, and can therefore strictly speaking not be interpreted as such. A 95% Bayesian credible interval is that region in which we believe the parameter to lie with probability 95%. This is how many practitioners actually interpret a frequentist confidence interval, but the "95%" refers to the long-term frequency with which 95% intervals of multiple trials contain the true value.[28] Within the Bayesian framework one can also calculate the probability that the outcome has a particular range of values, which cannot be done in the classical framework.[28]

The biggest drawback of the Bayesian approach is the introduction of subjective knowledge, via the prior distribution. The frequentist therefore says this method is inappropriate for objective scientific decision making and should only be used for individual decisions. The Bayesian in return says that there is no such thing as objectivity, since inference is

always done via an interpreter, whose background influences the inference. In addition, given enough data, the data can speak for themselves.

Bayesian statistics at its heart is ideally suited for meta-analysis, since the premise of both are the same: you have prior information available and it is updated with new data.[29] However, during the performance of the underlying study, Bayesian statistics was found not to be ideally suited for a simulation study such as we have done. When performing Bayesian statistics, the available data has to be the starting point. Ideally, the data is examined in detail, which will drive the modelling decisions around all aspects: priors, link function, initial values, etc. Additionally, the outcomes of a Bayesian model are meaningless when the model itself does not converge. Checking for convergence requires the visual examination of plots, and careful examination of other outcome measures. However, all these aspects are impossible to do in a simulation study, where many data sets are fitted one after the other.

9.6 DIFFERENT COUNTRIES

In chapters 6, 7 and 8 we discussed several aspects of transferability. Chapter 6 showed how differences in parameters between countries, also influenced the health economic outcomes. In our case study, these differences were primarily related to the epidemiology of diseases and the choice of discount rate. The least important factor was demography, i.e. the age/gender distribution of the cohort of smokers making a quit attempt. Unfortunately, this happens to be the easiest available set of parameters for new countries, whereas less easily available parameters were not specified country by country. Only one out of the six included countries had completely country-specific data.

In the case of quality of life, this was partly due to the lack of availability of country-specific utility values. In chapter 6, the country-specific results for the Belgian model used the Dutch value set.[30] Four out of the six countries referred back to the recommended tariffs for the EQ-5D for the UK.[31,32] Three papers in chapter 8 also used this value set, even when the study is not trying to calculate country-specific CE results in the UK. The UK tariffs are often used as it was the first validated value-set available. This may have been done to improve comparability between outcomes, but can also be due to unfamiliarity of researchers with the availability of other value sets for other countries. Apart from a Dutch value set which has been available since 2006, validated value sets for other countries have also been published, for example for Spain in 2001[33] and Germany in 2004[34], and many other countries. (See http://www.euroqol.org/ for other value sets). Differences between country-specific value sets of the EQ-5D are considerable, and some of the variation is due to cultural dissimilarities between countries. Using value sets from one country for another without any form of adjustment is therefore not advisable.[35]

In order to make good policy decisions, decision makers need to have all information available. In chapter 8, we have seen that a large percentage of papers is missing crucial information. Less than half of the included studies produced a table showing the demographic characteristics of patients in the original, underlying trial. Furthermore, less than a third of the papers specified the unit costs used in the study and almost a quarter of the papers did not list total health care utilization and/or health outcomes by treatment arm. All of this is considered basic information for a CEA. It is clear that a stricter adherence to for example the CHEERS guidelines, is necessary.[36]

9.7 DIFFERENT THRESHOLD VALUES

Several different outcome measures that are used in HTA are presented within the thesis. Chapters 2 to 5 perform a CE analysis (CEA), where the HE outcomes are shown as an incremental CE ratio (ICER). The ICER is the ratio of the difference in costs between two treatment options, and the difference in health outcomes. As explained in chapter 1, the ICER is compared to an (implicit or explicit) threshold. Policy makers can deem the intervention to be cost-effective compared to the comparator when the ICER is below this threshold.

In chapters 6 and 7, the outcomes are presented as the incremental net monetary benefits (INMB). The ICER was not used, due to several well-known problems with ratio statistics[37], combined with the fact that we wanted to calculate percentile changes in CE outcomes. The INMB is the difference in health outcomes, valued in monetary terms, minus the monetary costs. If the INMB is positive, the new intervention has more value to society than costs, and can thus be considered cost-effective compared to the comparator. If the INMB is negative, the new intervention will cost more than the societal benefits, and the intervention cannot be considered cost-effective compared to the comparator. In order to value the health outcomes in monetary terms, a "price" is needed for each unit of health. In chapter 6, the price, or "Willingness-to-pay" (WTP), is set at €20,000 per QALY gained. In that chapter, this was called "relatively low".

The price which can be deemed acceptable to pay for a unit of health, whether it is the threshold value to compare the ICER with, or the WTP to calculate the INMB, should be set by public discussion. Within CE studies, as was done in for example chapters 2 and 3, the outcomes can be presented for different threshold values, which allows the reader to make their own conclusion of CE. Another presentation tool is the CE acceptability curve, which shows the uncertainty around the outcomes for a whole range of threshold values. In order to avoid interfering directly with the public debate of the "price for health", the INMB may not be ideally suited to present results of CE studies, as it assumes a threshold value. However, academic research can be used to inform this public debate, by study-

ing what threshold value is deemed acceptable by the general public, or what implicit thresholds have been applied in past decisions.[38-43]

In 1998, a Dutch publication mentioned a threshold value of fl 40,000,- (€18,000), which was used for the first time in a clinical guideline.[44] In 1999, a threshold ICER of €20,000 per QALY gained has been proposed by the CVZ.[41] More recently, the Dutch Council for Public Health and Health Care (RVZ) suggested a maximum value of €80,000 per QALY for illnesses associated with a considerable burden.[45,46] This value corresponded to the implicit £ 50,000 per QALY gained from NICE at that time, and it reflected the application of the World Health Organization (WHO) threshold ICER to The Netherlands.[41] The WHO Commission on Macroeconomics and Health suggested that health technologies costing less than three times the gross domestic product (GDP) per capita for each disability adjusted life year (DALY) averted, represents good value.[47]

However, CVZ is reluctant to explicitly name a maximum threshold value for a number of reasons.[46] Firstly, there will always be other arguments that may influence a decision. Secondly, naming an explicit threshold value may elicit strategic pricing behaviour from pharmaceutical companies. Finally, it was found to be extremely difficult to elicit "societal" preferences from the general population. Other regulatory agencies around the world seem to have the same concerns. The UK regulatory organization NICE has publically discussed an "acceptable" threshold value. As a guideline rule, NICE accepts as cost-effective those interventions with an ICER of less than £ 20,000 per QALY gained and that there should be increasingly strong reasons for accepting as cost effective interventions with an ICER above a threshold of £ 30,000 per QALY gained.[48]

9.8 FOURTH HURDLE OR FOURTH FLOOR?

In considering how a new (pharmaceutical) product comes to market, CE is often described as the "fourth hurdle" in drug development.[49-51] In this terminology, the demonstration of quality, safety and efficacy are the first three hurdles. They can be considered hurdles, as they must all be overcome to secure a successful market registration. But being on a market does not guarantee successful commercialization in that market. This can be illustrated by for example the 13-valent pneumococcal conjugate vaccine (PCV13, Prevenar13, Pfizer). This vaccine is available for use in the Dutch infant population, but the 10-valent competitor PCV10 (Synflorix, GSK) won a tender for inclusion in the national vaccination program.[52] As such, PCV10 is free for parents, but PCV13 is not. CE was a crucial criterion in this policy decision. Another illustration is roflumilast, which was discussed in chapter 2. European market access has been gained in 2010, but it is still not reimbursed in The Netherlands at the end of 2013[53], despite pressure from patient groups and medical experts.[54] This is largely due to a lack of direct evidence of effectiveness

compared with inhaled corticosteroids.[54] The available indirect evidence, used in chapter 2, was not accepted due to a possible difference in patient population between the two combined trials. Once a product gains market access, the critical fourth hurdle to successful commercialization is thus to gain product reimbursement.

However, CE can also be judged in a different way: as the top floor of a metaphorical building, representing the evidence about a certain intervention. The ground floor of this building (or first floor, if we follow the US numbering) is evidence on the quality of medicines, encompassing for example such aspects as manufacturing, packaging, stability and impurities.[55] The next floor would be made from evidence that the intervention can be considered to be safe for use, followed by evidence that the intervention is effective in the indication for which it is to be used.[55] Having reached this floor, the researchers may gain market access and move up to the top floor, which offers a view towards successful commercialization.

Evidently, the stronger each of the lower floors are, the more solid the building. However, researchers in HTA usually only work on the top floor, and can often not personally check the source of evidence for every parameter. Instead, as in chapters 5 and 6, most if not all of the parameters come from sources in the published literature. These publications are read for any anomalies and differences in case definitions, but are otherwise taken at face value. In some cases, researchers in HTA can't even work with published or properly peer reviewed material. For example, the efficacy of roflumilast in the group of severe COPD patients used in chapter 2 were obtained from a subgroup analysis of patients in two RCTs. The efficacy for this "LABA subgroup" does not have an official publication, and thus a peer-reviewed evidence base. The references used in this chapter lead to the trial publication, where outcomes from the subgroups are not mentioned.[56] The numbers used in the chapter come from an internal publication, which cannot be cited from. Others have also brought to light problems with the evidence on lower floors, pointing at possible missing or shaky evidence on safety, efficacy and effectiveness of drugs (e.g.[57-59])

These are potentially severe issues for the lower floors of the evidence building, making building on top of them a potentially hazardous business. However, in addition, the fourth floor has its own methodological problems. In chapters 4 and 5, we discussed methods of meta-analysis, which bring together evidence from several sources in order to inform a CE model. However, many CE studies are based on evidence from a single source, either because only one source of evidence existed at that time, or because of pragmatic choices made by the modeller. This leads to less confidence in the outcomes of the CE study. Chapter 7 was a review of CE studies, all based on a single trial. Additionally, since this one source of data is often a RCT, these data are collected on ideal patients, often younger, with less co-morbidities and more homogeneous than "real-life" patients.

Another important current issue with CE research is structural uncertainty, which represents a lack of knowledge of the underlying true system. Just as parameter uncertainty

which arises from uncertainty around the population average, heterogeneity which arises from measurable differences in patient characteristics, and statistical uncertainty which is representative of unknowns that differ each time we run the same experiment, structural uncertainty is an integral part of HTA.[60] For example, the choice to model the difference between two interventions as a risk difference or a risk ratio is often a pragmatic one, not based on empirical knowledge of the 'true' difference between interventions, even if it is possible that such a 'truth' can be found. Unlike the other three important forms of uncertainty in HTA, structural uncertainty cannot be measured easily, and is often ignored in practice.

In conclusion, there are many possibilities of structural problems in the lower floors, in addition to many methodological issues that need to be solved on our own floor. Because of this, the fourth floor can be a rather shaky place to be. But with good research practice, and a continuous eye open for the potential pit falls, it is a worthwhile one.

9.9 LITERATURE

[1] VWS. Kiezen voor gezond leven 2007-2010 [Dutch]. 2006(Kamerstuk 2006-2007, 22894, nr. 110).

[2] STIVORO. Website of STIVORO for a smoke free future. 2008; Available at: www.stivoro.nl. Accessed 11/29, 2008.

[3] Kroes ME, Mastenbroek CG. Stoppen-met-rokenprogramma: te verzekeren zorg! 2009(29006531).

[4] Kaper J, Wagena EJ, Willemsen MC, van Schayck CP. Reimbursement for smoking cessation treatment may double the abstinence rate: results of a randomized trial. Addiction 2005 Jul; 100(7):1012-20.

[5] Kaper J. Smoking cessation treatment and its reimbursement, the costs and effects. Maastricht: Datawyse | Universitaire Pers Maastricht; 2006.

[6] Soethout J, van den Berg B. Proefimplementatie 'Stoppen met roken' [Test implementation 'Smoking cessation']. 2009;1663.

[7] Over EA, Feenstra TL, Hoogenveen RT, Droomers M, Uiters E, van Gelder BM. Tobacco Control Policies Specified According to Socioeconomic Status: Health Disparities and Cost-effectiveness. Nicotine Tob Res 2014 Jan 4.

[8] Verbiest ME, Chavannes NH, Crone MR, Nielen MM, Segaar D, Korevaar JC, et al. An increase in primary care prescriptions of stop-smoking medication as a result of health insurance coverage in the Netherlands: population based study. Addiction 2013 Dec;108(12):2183-2192.

[9] STIVORO. Kerncijfers roken in Nederland 2012. Een overzicht van recente Nederlandse basisgegevens over rookgedrag. 2013.

[10] STIVORO. Factsheet 'Reimbursement'. 2013; Available at: http://stivoro.nl/tabaksontmoediging/stoppen-met-roken/inrichtingvandezorg/vergoeding/. Accessed 12/3, 2013.

[11] De Kanter W, Dekker P. NederlandStopt.Nu [The Netherlands Quits Now]. 2014; Available at: http://www.nederlandstopt.nu. Accessed 01/25, 2014.

[12] Bouma J. Het rookgordijn [The Smoke Screen]. ; 2003.

[13] Van Woerden I, Braam S. Den Haag en de Tabakslobby [The Hague and the Tabbaco Lobby]. Vrij Nederland 2013.

[14] Willemsen MC. Unieke inkijk in de Nederlandse tabakslobby [Unique view in the Dutch Tobacco Lobby]. 2013; Available at: http://www.maastrichtuniversity.nl/web/Main1/SiteWide/SiteWide11/UniekeInkijkInDeNederlandseTabakslobby.htm. Accessed 01/28, 2014.

[15] Boseley S. EU urged to press ahead with tobacco crackdown amid lobbying scandal. Guardian 2012;2014.

[16] Doward J. Tobacco giant Philip Morris 'spent millions in bid to delay EU legislation'. Guardian 2013.

[17] van Baal PH, Feenstra TL, Hoogenveen RT, de Wit GA, Brouwer WB. Unrelated medical care in life years gained and the cost utility of primary prevention: in search of a 'perfect' cost-utility ratio. Health Econ 2007 Apr;16(4):421-33.

[18] van Baal PHM. Practical Application to Include Disease Costs (PAID). 2012; Available at: http://oldwww.bmg.eur.nl/personal/vanbaal/paid.htm. Accessed 12/16, 2013.

[19] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7(3):177-188.

[20] Puhan MA, Bachmann LM, Kleijnen J, Ter Riet G, Kessels AG. Inhaled drugs to reduce exacerbations in patients with chronic obstructive pulmonary disease: a network meta-analysis. BMC Med 2009 Jan 14;7:2.

[21] Strassmann R, Bausch B, Spaar A, Kleijnen J, Braendli O, Puhan MA. Smoking cessation interventions in COPD: a network meta-analysis of randomised trials. Eur Respir J 2009 Sep; 34(3):634-640.

[22] Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. Med Decis Making 2013 Jul;33(5):607-617.

[23] Akobeng AK. Understanding randomised controlled trials. Arch Dis Child 2005 Aug;90(8): 840-844.

[24] Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000 Jun 22;342(25):1887-92.

[25] Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. BMJ 2004 Jun 19;328(7454):1490.

[26] Vemer P, Al MJ, Oppe M, Rutten-van Mölken MPMH. Updating parameters of decision-analytic costeffectiveness models: a systematic comparison of methods. 2011;152002002.

[27] Crowder MJ. Multivariate Survival Analysis and Competing Risks. New York, NY: Taylor & Francis Group, LLC; 2012.

[28] Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 5.0.2 updated September 2009. 2009.

[29] Berry DA. Bayesian approaches for comparative effectiveness research. Clin Trials 2012 Feb; 9(1):37-47.

[30] Lamers LM, McDonnell J, Stalmeier PF, Krabbe PF, Busschbach JJ. The Dutch tariff: results and arguments for an effective design for national EQ-5D valuation studies. Health Econ 2006 Oct;15(10):1121-32.

[31] Dolan P. Modeling valuations for the EuroQol health states. Med Care 1997;35:1095-108.

[32] Drummond M, Sulpher M, Torrance G, O'Brien B, Stoddart G. Methods for the Economic Evaluation of Health care Programmes. 3rd ed. Oxford: Oxford University Press; 2005.

[33] Badia X, Roset M, Herdman M, Kind P. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. Med Decis Making 2001 Jan-Feb; 21(1):7-16.

[34] Greiner W, Claes C, Busschbach J, Graf von der Schulenburg J. Validating the EQ-5D with time trade off for the German population. European Journal of Health Economics 2004;[Epub ahead of print].

[35] Knies S, Evers SM, Candel MJ, Severens JL, Ament AJ. Utilities of the EQ-5D: transferable or not? Pharmacoeconomics 2009;27(9):767-779.

[36] Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) Statement. Pharmacoeconomics 2013 Mar 26.

[37] Sendi P, Gafni A, Birch S. Opportunity costs and uncertainty in the economic evaluation of health care interventions. Health Econ 2002 Jan;11(1):23-31.

[38] Hirth RA, Chernew ME, Miller E, Fendrick AM, Weissert WG. Willingness to pay for a quality-adjusted life year: in search of a standard. Med Decis Making 2000 Jul-Sep;20(3):332-342.

[39] Culyer A, McCabe C, Briggs A, Claxton K, Buxton M, Akehurst R, et al. Searching for a threshold, not setting one: the role of the National Institute for Health and Clinical Excellence. J Health Serv Res Policy 2007 Jan;12(1):56-58.

[40] Mason H, Jones-Lee M, Donaldson C. Modelling the monetary value of a QALY: a new approach based on UK data. Health Econ 2009 Aug;18(8):933-950.

[41] Simoens S. How to assess the value of medicines? Front Pharmacol 2010;1:115.

[42] Baker R, Bateman I, Donaldson C, Jones-Lee M, Lancsar E, Loomes G, et al. Weighting and valuing quality-adjusted life-years using stated preference methods: preliminary results from the Social Value of a QALY Project. Health Technol Assess 2010 May;14(27):1-162.

[43] Claxton K, Martin S, Soares M, Rice N, Spackman E, Hinde S, et al. Methods for the Estimation of the NICE Cost Effectiveness Threshold. 2013;Paper 81.

[44] Casparie A, van Hout BA, Simoons ML. Guidelines and costs [Dutch]. Ned Tijdschr Geneeskd 1998;142(38):2075-7.

[45] RVZ. Zinnige en duurzame zorg (Sensible and sustainable Care). 2006;06/06.

[46] Van Busschbach JJ, Delwel GO. Het pakketprincipe kosteneffectiviteit - Achtergrondstudie ten behoeve van de 'appraisal' fase in pakketbeheer. 2010;29079523.

[47] WHO Commission on Macroeconomics Health. Macroeconomics and Health: Investing in Health for Economic Development. 2001.

[48] NICE. Methods for the development of NICE public health guidance (third edition) - (6) Incorporating health economics. 2012; Available at: http://publications.nice.org.uk/methods-for-the-development-of-nice-public-health-guidance-third-edition-pmg4/incorporating-health-economics. Accessed 12/23, 2013.

[49] Honig PK. Comparative effectiveness: the fourth hurdle in drug development and a role for clinical pharmacology. Clin Pharmacol Ther 2011 Feb;89(2):151-156.

[50] Rawlins MD. Crossing the fourth hurdle. Br J Clin Pharmacol 2012 Jun;73(6):855-860.

[51] Rogowski WH. An economic theory of the fourth hurdle. Health Econ 2013 May;22(5):600-610.

[52] Health Council of the Netherlands. Vaccinatie tegen pneumokokken [in Dutch "Vaccination of infants against pneumococcal infections"]. 2010;2010/02(2005/13).

[53] CVZ. Farmacotherapeutisch Kompas [Pharmacotherapeutic Compass]. 2013; Available at: http://www.fk.cvz.nl/. Accessed 09/12, 2013.

[54] CVZ. Pakketscan COPD. 2013;211.

[55] European Medicine Agency. Scientific guidelines. 2014; Available at: http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000043.jsp&mid=WC0b01ac05800240cb. Accessed 01/28, 2014.

[56] Calverley PM, Rabe KF, Goehring UM, Kristiansen S, Fabbri LM, Martinez FJ, et al. Roflumilast in symptomatic chronic obstructive pulmonary disease: two randomised clinical trials. Lancet 2009 Aug 29;374(9691):685-694.

[57] Ioannidis JP. Why most published research findings are false. PLoS Med 2005 Aug;2(8):e124.

[58] Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess 2010 Feb; 14(8):iii, ix-xi, 1-193.

[59] Goldacre B. Bad Pharma. How Drug Companies Mislead Doctors and Harm Patients. Hammersmith: Harper Collins UK; 2012.

[60] Groot Koerkamp B, Weinstein MC, Stijnen T, Heijenbrok-Kal MH, Hunink MG. Uncertainty and patient heterogeneity in medical decision models. Med Decis Making 2010 Mar-Apr; 30(2):194-205.

Chapter 10

Afterword

10.1 SUMMARY

At its heart, health technology assessment (HTA) is very simple. It compares two or more alternative courses of action, often pharmaceutical interventions, in terms of both their costs and health outcomes. Better health outcomes usually come at extra costs, often in the way of a higher price for the intervention. HTA makes this exchange between costs and effects explicit. However, HTA is facing many methodological challenges, calling for more complexity in the analyses.

In chapter 2, we gave an example of a health-economic (HE) decision model as is commonly used to tackle the analysis complexity. The model was used to show the long term HE effects of the reimbursement of smoking cessation treatments. The study showed that reimbursement of smoking cessation support via the obligatory health care insurance in The Netherlands would result in fewer smokers and more quality-adjusted life years (QALYs). It is a cost-effective way to contribute to a reduction in the percentage of smokers.

One of the complexities discussed in this thesis was the heterogeneous nature of patients. In chapter 3, we showed that there are several ways of dealing with heterogeneity and that the outcomes, and thus the policy decision, may change when heterogeneity is handled differently. Three of these methods discussed can be useful in cost-effectiveness (CE) research, each in different circumstances. When little or no heterogeneity is expected, or when it is not expected to influence the CE results, disregarding heterogeneity may be correct. Subgroup analyses may inform policy decisions on each subgroup, as long as they are well defined and the characteristics of the cohort that define a subgroup truly represent the patients within that subgroup. Despite the necessary calculation time, the Double Loop Probabilistic Sensitivity Analysis (PSA) is a viable alternative, which leads to better results and better policy decisions, when accounting for heterogeneity in a Markov model. The Single Loop PSA can only be used to calculate the point estimate of the expected outcome. It disregards the fundamental differences between heterogeneity and sampling uncertainty, and overestimates overall uncertainty as a result.

The second complexity discussed in this thesis was the difference between data sources that have to be combined. In chapters 4 and 5, we compared several methods of meta-analysis. Using a simulation study we could compare the HE outcomes to a golden standard, and each other. In chapter 4, which compared methods of direct meta-analysis, frequentist fixed effects (FFE), frequentist random effects (FRE) and Bayesian fixed effects (BFE) led to comparable HE outcomes, even in scenarios where we built in heterogeneity. Bayesian random effects (BRE) tends to overestimate uncertainty reflected in the shape of the CE acceptability curve.

In chapter 5 we compared several methods of indirect meta-analysis. Puhan's method and the Generalized Linear Model Fixed Effects (GLMFE) showed similar results, with GLMFE having the tendency to overestimate uncertainty, but also having lower average

bias and mean absolute deviation (MAD). Generalized Linear Model Random Effects (GLMRE) showed large bias and MAD, and overestimated uncertainty even more. Based on this study, where we had to combine nine trials in a network that includes evidence for all treatment combinations, we would recommend Puhan's method or GLMFE as the preferred method of indirect meta-analysis.

The final complexity discussed in this thesis, were differences between countries. Many factors should be taken into account when transferring cost-effectiveness results across countries and settings and there are many interactions between these factors. This stresses the importance of carefully considering whether foreign results can be applied and adapted to a different setting. We've shown in chapter 6 that it is not only important to see which factors vary, but also how much this variation in factors causes variation in CE. The factors that cause the most variation in cost-effectiveness do not necessarily have to be the same as the factors that vary most themselves. Chapter 7 showed that the importance of each of the factors is also influenced by the local threshold value for a QALY. When studying the CE of smoking cessation, there is a need for local data even for countries within a similar region of the world.

Chapter 8 discussed CE analyses that were based on recent, large multinational randomized controlled trials. Several advanced statistical techniques are available to calculate country-specific CE results from multinational trials. These methods take the interaction between country and treatment effect on health and health care utilization into account. Hierarchical models also lower variability of the country-specific CE results and lead to more appropriate population estimates. However, they have not been used on a wide scale yet, while simpler, naïve methods are still routinely employed.

10.2 NEDERLANDSTALIGE SAMENVATTING

Het principe van evaluatieonderzoek in de gezondheidszorg (*health technology assessment*, HTA) is simpel. Twee of meer alternatieve behandelingen, vaak farmaceutische interventies, worden vergeleken in termen van kosten en effecten. Betere gezondheidsuitkomsten gaan meestal gepaard met hogere kosten, vaak door een hogere prijs voor de interventie. HTA maakt deze vergelijking expliciet. Echter, diverse methodologische problemen binnen de HTA vragen om steeds complexere analyses.

In hoofdstuk 2 werd een voorbeeld van een gezondheidseconomisch (GE) beslismodel gegeven, zoals deze gebruikt worden voor dergelijke complexe analyses. Het model werd gebruikt om de lange termijn kosteneffectiviteit (KE) te berekenen voor het vergoeden van stoppen-met-roken behandelingen. De studie toonde aan dat een dergelijke vergoeding vanuit de verplichte ziektekostenverzekering in Nederland zal leiden tot minder rokers en meer voor kwaliteit van leven gewogen levensjaren (*quality-adjusted life years*, QALYs). Het is een kosteneffectieve manier om het percentage rokers in Nederland naar beneden te brengen.

Eén van de complexiteiten die in dit proefschrift zijn besproken, is de heterogeniteit in groepen patiënten. In hoofdstuk 3 lieten we zien dat er diverse manieren zijn om met heterogeniteit om te gaan en dat de KE uitkomsten, en dus de beleidsbeslissing, anders kunnen zijn als een andere methode wordt gekozen. Drie van de genoemde methodes kunnen van nut zijn in HTA, elk in andere omstandigheden. Met weinig of geen heterogeniteit, of wanneer het wordt verwacht dat dit de KE uitkomsten niet zal beïnvloeden, kan heterogeniteit worden genegeerd. Subgroepen kunnen beleidsbeslissingen voor elke subgroep apart ondersteunen, zolang deze subgroepen goed zijn afgekaderd en de karakteristieken van het cohort binnen elke subgroep ook daadwerkelijk alle patiënten binnen de subgroep vertegenwoordigen. Ondanks de benodigde rekentijd, is de *Double Loop* probabilistische gevoeligheidsanalyse (*Probabilistic Sensitivity Analysis*, PSA) een goed alternatief dat leidt tot betere resultaten en beleidsbeslissingen. De *Single Loop* PSA, waarin heterogeniteit en parameteronzekerheid tegelijk worden geanalyseerd, negeert het fundamentele verschil tusen beide en zal daardoor de onzekerheid overschatten.

De tweede complexiteit die in dit proefschrift werd beschreven, is het verschil tussen databronnen die samengevoegd moeten worden. In hoofdstukken 4 en 5 werden diverse methoden van datasynthese vergeleken. Met een simulatiestudie was het mogelijk om de GE uitkomsten te vergelijken met een Gouden Standaard en met elkaar. In hoofdstuk 4, waarin vier methoden van directe datasynthese werden vergeleken, bleken de *frequentist fixed effects* (FFE), *frequentist random effects* (FRE) en *Bayesian fixed effects* (BFE) tot vergelijkbare GE resultaten te leiden, zelfs in scenarios waarin heterogeniteit was ingebouwd. *Bayesian random effects* (BRE) neigde naar een overschatting van de onzekerheid.

In hoofdstuk 5 werden diverse methodes van indirecte datasynthese vergeleken. Puhan's methode en *Generalized Linear Model Fixed Effects* (GLMFE) leidde tot vergelijkbare resultaten, waarin GLMFE de neiging vertoonde om de onzekerheid te overschatten, maar ook een lagere systematische fout en absolute afwijking had. *Generalized Linear Model Random Effects* (GLMRE) liet een grote bias en absolute afwijking zien, en overschatte de onzekerheid nog meer. Gebaseerd op deze studie, met negen studies in een netwerk waarin bewijs is voor alle mogelijke interventiecombinaties, zouden wij Puhan's methode of GLMFE prefereren voor het doen van indirecte datasynthese.

De laatste complexiteit in dit proefschrift besproken, is het verschil tussen landen. Een grote hoeveelheid factoren moet in acht worden genomen wanneer KE resultaten worden vertaald tussen landen, en er diverse interacties tussen deze factoren zijn. Dit benadrukt het belang om goed na te gaan of resultaten uit een andere land kunnen worden toe- en aangepast in de eigen omgeving. In hoofdstuk 6 hebben we laten zien dat het niet alleen van belang is om goed te kijken welke factoren veranderen, maar ook hoeveel de variatie in deze factoren, de variatie in KE veroorzaakt. Factoren die de meeste variatie in KE veroorzaken tussen landen, zijn niet noodzakelijk factoren die zelf veel variëren tussen landen. Hoofdstuk 7 toonde aan dat het belang van elke factor ook kan veranderen als de maatschappelijke bereidheid om te betalen voor een QALY verandert. Wanneer de KE van stoppen-met-roken middelen wordt onderzocht, is er een duidelijke noodzaak voor lokale data, zelfs voor landen in vergelijkbare regio's.

Hoofdstuk 8 besprak KE studies die zijn gebaseerd op recente, grote multinationale studies. Diverse geavanceerde statistische technieken zijn beschikbaar om landspecifieke KE resultaten te berekenen. Deze methodes houden rekening met de interactie tussen het land en het behandeleffect. Hiërarchische modellen verlagen ook de variabiliteit in landspecifieke resultaten en leiden tot betere schattingen. Echter, deze methodes worden niet op grote schaal gebruikt, terwijl simpelere, naïeve methodes nog regelmatig worden toegepast.

10.3 ACKNOWLEDGMENTS

And here starts the most read part of any thesis: the acknowledgments. They are always incomplete, as every success is a sum of the interactions with many people, but I do try to be as complete as I possibly can. The first solid step on the road towards this thesis was inspired by Dave Sugano, a delightful man who I'm indebted to for the reference he gave me at iMTA. He first told me about the exciting field of health economics "which is up and coming" and has "the best European institute right next door to where you live."

To come to iMTA, I left the good people at APE behind me. They have taught me many small and big things, not in the least on the political side of organizations and the practical side of economics. I need to thank the partners and all the other colleagues for the wonderful time I have had on one of the most beautiful spots in The Hague.

During the time at iMTA, I have been involved with several organizations, which have shaped me the way I am. The people of Toastmasters of The Hague, Rotterdam Toastmasters and Toastmasters District 59 (Continental Europe) are very dear to me. Coming from all over the world, they have inspired me endlessly and helped me hone my communication and leadership skills. In the municipality of Barendrecht, I was involved with D66. My fellow democrats have rekindled my love of the common good, and my political ambitions. I have continued my work for the community in Tynaarlo, and I hope that you will be as proud of my accomplishments, as I am of what you have accomplished in the past few years. My fellow members of the *Culturele Raad* (Cultural Council) Barendrecht have taught me a lot about culture policies and have given me a new appreciation of local cultural initiatives.

For my newest endeavor, I have to thank Dr. Talitha Feenstra, Dr. Paul Krabbe and Prof. dr. Maarten Postma for giving me the opportunity to continue my work in a new environment, and my new colleagues for making me feel at home in Groningen.

Two men deserve special mention, as they have kept me sane during the past decade or longer. Without the (almost) daily correspondence between The Hague, Brussels and Rotterdam, my life would have been much more boring. Thank you, Jelte Theisens and Jan Willem Gerritsen, for the many wonderful moments that we have shared. Cold wind in Brugge, rain in Nice, more rain in Cannes, even more rain in Maastricht and of course "trains on a bridge" in Nijmegen. I hope to share many more moments, glasses and e-mails with you.

During the time I have worked on this thesis, it never ceased to amaze me how incredibly nice health economists are. They might well be the best that the dismal science has to offer. It started with the wonderful colleagues at iMTA and the iBMG: you all made the time I've spend with you special. I have warm recollections of many moments, of which daily morning coffee before the day properly starts, the Roparun in 2012 and the conferences stand out. Conferences and courses are also a nice way of meeting new people,

who you will meet again and again for many years afterwards. (Yes, you, I know you are looking for your name - admit it.)

I particularly enjoyed giving classes during my time in Rotterdam and I have to thank Dr. Erwin Birnie, Dr. Ken Redekop and Dr. Jacco Keja for allowing me time in their various courses. As much as I taught the students, I think I have learned more myself.

I would also like to thank several sponsors for their financial support for the research shown in this study: the Dutch Ministry of Health for the study on which chapter 2 is based, Takeda for the study on which chapter 3 is based, ZonMW for the study on which chapters 4 and 5 are based and Pfizer for the funding of the study discussed in chapters 6 and 7. I would also like to thank the management of iMTA for allowing me to use part of the investment budget for my investigation into the field of transferability, which allowed me to write chapters 6 and 8.

Many thanks go to Prof.dr. C.A. Uyl-de Groot, Prof.dr. F. Rutten and Prof.dr. M.J. Postma, who graciously accepted the invitation to take part in the judging committee for this thesis.

Prof.dr. Maureen Rutten-van Mölken did an exemplary job in guiding me through this thesis. There have been many ways in which my mind, and my projects can get side-tracked and she always managed to get me back on track. Maureen, I thank you for the opportunity to write this thesis under your guidance, and for all the things I have learned. You and I sometimes have polar opposite ways of doing things, and I think discussing these differences and learning from the way you approach things, have vastly improved my work.

After having spent the first month at iMTA in a separate room by myself, a new colleague was hired at iMTA. When I discussed this with the other colleagues, they were shocked (SHOCKED!) to discover that I had already run her name through the internet search engines and that I knew a lot about Saskia Schawo already. Fortunately, there was much more to discover about you, Saskia, and we soon found out we had similar interests. You brightened my days when you were around. I feel incredibly fortunate that we shared an office and I look back on that time with great fondness.

Most of the "junior" staff at a university (read: not holding a PhD-title) come straight from their Masters defense. At iMTA, only a few were a bit older, with a whole career behind us. One of these was Lucas Goossens, and this meant we had something in common. Lucas, I love the way you dive into completely new subject matters, and I treasure the memory of our discussions, be it about statistics, politics or TV series.

All of the changes in my working life since 2002, have been thoroughly discussed with Ingeborg Been. She started out as a colleague, quickly became someone to share diner with in the many restaurants in The Hague, became a close friend and is someone I consider to be a coach. Ingeborg, your opinions, your ideas, the way you look at things have always meant a great deal to me.

Lucas and Ingeborg, I am honoured to have both you as my paranymph.

Where I have lived around the whole country, my sister Erlinde and her family have always been a steady place to come home to, close to Nijmegen. We may not see each other very often, but all time spend together is valuable to me.

Dear Hans and Wilma. I always hear people say that “working in health care runs in the family”. I initially escaped this by becoming an econometrician, but as with all clichés, it was stronger than I was. Not only did I marry a doctor, but I was also drawn to the field after having worked for years in social welfare. An upbringing filled with medical terms at the dining room table has prepared me for many parts of the work I do today. It isn’t the only thing I have learned that I have taken with me on this road. Having been brought up in a loving family has shaped me to the man I am today. Thank you for all that you have given me, all that you have taught me, and every way that you inspired me.

Finally, I have to thank the three brightest rays of sunshine in my life: Eveline, Casper and Anne. You bring me joy daily. Watching every step that the children take makes me feel proud. Although I do still have nightmares of a certain red, lighted button under my desk… Anne, thank you for putting up with my evenings of work, and for allowing me to continue to do so, even now that this thesis is done and I am pursuing other things. I love you and I hope to spend many more years together; learning, growing, enjoying.

10.4CURRICULUM VITAE

Pepijn Vemer was born in Enter, on July 10th, 1975. Having spent his childhood in several cities in The Netherlands, he started his econometrics study at the University of Groningen in 1994. During this time he was an active member of the Stichting Eloquentia Groningen (Elocution Society), as a debater, board member and the organizer of the biweekly debating events. He also organized the Dutch debating championships twice. In addition, he was a member of the local political party *Student en Stad*. Besides contributing to the political work of the party, he was the editor in chief of the (irregular) newsletter.

He graduated in 2000, after an internship at IMS Health, Plymouth Meeting, USA. At IMS Health he investigated time series of drug sales, before and after going off patent. His first job was as at ANOVA Health Care Insurance (now part of Agis Health Care Insurance) in Amersfoort. There he was responsible for quarterly reports, for the setting up of a quality feed-back system for physical therapists and for premium calculations for new products. In 2002 he was hired as a consultant at APE in The Hague, where he did research and advice in several fields in public and private sectors. Having started in the field of health care insurance and disability, he later branched out into social welfare. His particular interest was in the nation-wide implementation of the model to distribute the National Welfare Budget over local municipalities objectively, and in the local consequences of this model.

During this time he was heavily involved with local chapters of Toastmasters, an international organization which teaches its members communication and leadership skills. He has acted as President of Toastmasters of the Hague (TMOTH) and as Founding President of Rotterdam Toastmasters. During his time in Toastmasters he earned the Advanced Communicator Bronze and Advanced Leadership Bronze.

In 2008 he started work at the Institute for Medical Technology Assessment (iMTA), part of the Institute for Policy and Management (iBMG) at the Erasmus University Rotterdam. Here he primarily did contract research in the area of cost-effectiveness of new medical technologies, with a main interest in transferability between jurisdictions. Several large projects were performed by him, together with several colleagues both within and outside the university. The largest of these projects was a ZonMW project dealing with the methodology of meta-analysis. Research finding have been disseminated at conferences and in academic journals. He also taught various classes at different levels, ranging from first year Bachelor students to the NIHES Summer Program. See 10.5 for a complete overview of the work done during this period.

From 2013 onwards, he has been in a Postdoc-position working for the department of Epidemiology (Unit HTA) at the University Medical Center Groningen, and Pharmacoepidemiology and Pharmacoeconomics (PE2) at the University of Groningen.

10.5PHD PORTFOLIO

The following 'deliverables' have been produced, based on work done at iBMG/iMTA, between February 2008 and January 2013.

10.5.1 Publications

An up to date list of publications by the author can be found on:
http://www.ncbi.nlm.nih.gov/pubmed?term=Vemer%20P[Author] and
http://www.researchgate.net/profile/Pepijn_Vemer

"Health Related Quality-of-Life and productivity-losses in patients with depression and anxiety disorders" Bouwmans C, Vemer P, Van Straten A, Tan SS, Rutten F, Hakkaart L. J Occup Environ Med. 2014 Apr;56(4):420-4. doi: 10.1097/JOM.0000000000000112.

"Let's get back to work. Survival analysis on the return-to-work after depression " Vemer P, Bouwmans C, Vlasveld M, Van der Feltz C, Hakkaart L. Neuropsychiatr Dis Treat. 2013;9:1637-45

"The road not taken. Transferability issues in multinational trials" Vemer P, Rutten-van Mölken MPMH. Pharmacoeconomics. 2013 Oct;31(10):863-876

"A Choice That Matters? Simulation Study on the Impact of Direct Meta-Analysis Methods on Health Economic Outcomes " Vemer P, Al MJ, Oppe M, Rutten-van Mölken MPMH. Pharmacoeconomics. 2013 Aug;31(8):719-30.

"A systematic review of hospital-at-home care. Cost savings are overestimated" Goossens LM, Vemer P, Rutten-van Mölken MPMH. Chapter in "Underestimated uncertainties. Hospital-at-home for COPD exacerbations and methodological issues in the economic evaluation of healthcare" Goossens LM, iBMG Rotterdam 2012 ISBN 978-94-6169-338-9

"Largely Ignored: The impact of the threshold value for a QALY on the importance of a transferability factor." Vemer P, Rutten-van Mölken MP. Eur J Health Ec 2011;12(5):397-404

"If you try to stop smoking, should we pay for it? The cost-utility of reimbursing smoking cessation support in the Netherlands" Vemer P, Rutten-van Mölken MPMH, Kaper J, Hoogenveen RT, Van Schayck CP, Feenstra TL. Addiction, 2010;105(6):1088–1097

"Crossing Borders: Factors Affecting Differences in Cost-Effectiveness of Smoking Cessation Interventions between European Countries." Vemer P, Rutten-van Mölken MP. Value Health. 2010;13(2):230–241

"Internationale vertaalbaarheid van kosten-effectiviteit" Rutten-van Mölken MPMH, Vemer P. Chapter in "Van Kosten tot Effecten Een handleiding voor evaluatiestudies in de gezondheidszorg " Rutten-van Mölken,M.P.M.H. (Ed.), 2nd ed, ELSEVIER gezondheidszorg, Maarssen 2010 ISBN 978 90 352 3187 0

10.5.2 Projects

International comparison of health-economic outcomes for varenicline
Reimbursement of smoking cessation support
Cost-effectiveness of new pneumococcal vaccines
Go-ahead review
Zon/MW project on Meta Analysis
Transferability: Paper Threshold value for a QALY
Tiotropium Workorder
Nycomed
Transferability: Invited paper Pharmacoeconomics
Major Depressive Disorder (Lundbeck)
Functional Family Therapy
Model validation for Celgene
Zon/MW project on Chronic Lymphocytic Leukemia (CLL)

10.5.3 Education

Workgroups "Inleiding Methoden en Technieken van Onderzoek (M&T1)", BA1, 2008-2012.
Evaluator for Health Econometrics, 2008
Workgroup "Transferability", 2009-2012
Workgroup "Excelvaardigheden", 2009, 2010, 2012
Erasmus Summer Program, course "Health Economics", 2009-2012.
Workgroup "Debatteren" (Lijnvaardigheden), 2010
Workgroup "Markov Modeling", 2010-2011
Workgroup "Uncertainty", 2011
Classes BA-M&T1 Inleiding Methoden en Technieken van Onderzoek (M&T1), "Steekproeven" (2011-2012) and "Steekproeffouten" (2012).
Supervisor master graduation Georgios Gkountouros.
Supervisor Bachelor thesis Suzanne Oomen.
Co-evaluator Master Thesis Michelle Lieuw-On
Co-evaluator Master Thesis Daisy Duell
Course coordinator "Life Sciences Pricing and Management", 2011
Preparation for new elective course Pharmaceutical Pricing and Market Access, 2012

10.5.4 Courses

Evidence Synthesis for Decision Modeling, Venice June 29 - July 3, 2009
Basiscursus didaktiek, Rotterdam May-June, 2010
Training "Nieuwe B1 programma"

10.5.5 Scientific meetings and presentations

NV-TAG, Bilthoven, Mar 14th, 2008

ISPOR 11th annual European congress, Athens, Nov 9th-12th, 2008

Poster presentation "Crossing Borders: Factors affecting differences in cost-effectiveness of smoking cessation interventions between European countries"

NV-TAG, Utrecht, Nov 14th, 2008

poster presentation "Crossing Borders: Factors affecting differences in cost-effectiveness of smoking cessation interventions between European countries"

LoLa HESG, Berg en Terblijt, May 28th and 29th, 2009

Discussant, "The Longitudinal Relationship Between Health Status And Costs Of Hospital Use", Bram Wouterse, Universiteit Tilburg / Tranzo.

Own paper discussed by Saskia Knies, Maastricht University, Faculty Health Med & Life Science, School Public Health & Primary Care: "Crossing Borders: Factors affecting differences in cost-effectiveness of smoking cessation interventions between European countries"

ISPOR 12th annual European congress, Paris, Oct 24th-27th, 2009

Podium presentation "Seven, Ten or Thirteen? The cost-utility of infant vaccination with a 7-, 10- and 13-valent Pneumococcal Conjugate Vaccine in the Netherlands"

Poster presentation "Largely Ignored: The impact of the threshold value for a QALY on the importance of a transferability factor"

ZonMw/NVTAG, HTA-methodologie geneesmiddelen, Utrecht, Jan 29th, 2010.

Oral presentation "Tussenresultaten Meta Analyse"

LoLa HESG, Egmond aan Zee, May 27th and 28th, 2010

Gespreksleider.

ISPOR 13th annual European congress, Prague, Nov 6th-9th, 2010

Poster presentation "Country adaptation of a health economic model. The case for roflumilast in The Netherlands"

ZonMw/NVTAG, HTA-methodologie geneesmiddelen.

Presentatie tussenresultaten Meta Analysis Nov 23th, 2010

Bijeenkomst HTA methodologie, Utrecht, Sep 12th, 2011

Oral presentation "Updating parameters of decision-analytic cost effectiveness models: a systematic comparison of methods."

ISPOR 14th annual European congress, Madrid, Nov 5th-8th, 2011

Oral presentation "A choice that matters: Comparing methods of data synthesis in cost-effectiveness modelling"

LoLa HESG, Landgoed Ehzerwold, May 24th, 2012

Discussant, "How should we deal with patient heterogeneity in economic evaluation: a systematic review of pharmacoeconomic guidelines", Bram Ramaekers et al., Universiteit Maastricht.

Own paper discussed by Willem Woertman Nijmegen University: "Comparing methods of indirect meta-analysis in health economic models"

ISPOR 15th annual European congress, Berlin, Nov 3th-7th, 2012

Poster presentation "Comparing methods of mixed treatment comparisons in health economic models"

Poster presentation "A systematic review of hospital-at-home care: cost savings are overestimated"

10.5.6 Other

Organizer lecture Prof Richard Gill, "Probiotics", Mar 24th, 2008

Course leader Public speaking for BMG employees, 2008, 2009.

Pier review for Eur J Health Econ (Jun 2008, May 2010, May 2012, Dec 2012), Vaccine (July 2010), British Journal of Medicine and Medical Research (Apr 2012)

Parents' Day FBMG, Feb 13th, 2009

Meeting with BI in Frankfurt, Nov 19th, 2009

Course leader Public speaking for BMG employees, two sessions from Jan 11th, 2010.

Worked for two months in San Francisco, 2012

Roparun, 2012

Astellas Advisory Board Meeting (Dec 7th, 2012)

10.5.7 Awards

Best New Investigator Podium Presentation for Podium presentation "Seven, Ten or Thirteen? The cost-utility of infant vaccination with a 7-, 10- and 13-valent Pneumococcal Conjugate Vaccine in the Netherlands" at the ISPOR 12th annual European congress, Paris, Oct 24th-27th, 2009

ERASMUS UNIVERSITEIT ROTTERDAM

iMTA