



Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

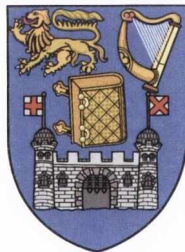
Bayesian Methods for Evidence Synthesis and Decision Making in Health Technology Assessments

A thesis submitted to the University of Dublin, Trinity College

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Statistics, University of Dublin, Trinity College



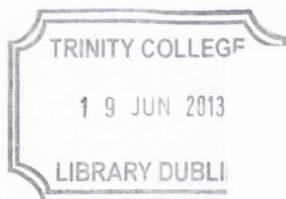
January 2013

Susanne Schmitz

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university. Except where otherwise stated, the work described herein has been carried out by the author alone.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement. The copyright belongs jointly to the University of Dublin and Susanne Schmitz.



Thes

Abstract

In this thesis statistical methodologies to improve reimbursement decisions in healthcare are developed. The allocation of healthcare resources is a very topical issue in the current economic climate. Given a limited healthcare budget it is important to distribute scarce resources fairly and efficiently. To aid decision making, new technologies are formally evaluated regarding their costs and benefits prior to reimbursement in so-called Health Technology Assessments. The work conducted in this thesis feeds into the evidence synthesis as well as the decision theoretic aspect of this process.

Bayesian mixed treatment comparison models provide a powerful methodology to estimate the relative efficacy between alternative agents, when direct evidence is not available or insufficient. To ensure efficient use of available evidence, the increased power of continuous outcome measures in such models compared to dichotomized measures is illustrated. The impact of structural uncertainty on a subsequent cost utility analysis is investigated and incorporated using value of information. To facilitate informed decision making based on all available evidence, three alternative methodologies to systematically include evidence from different trial designs are developed.

Aside from relative efficacy, the decision on reimbursement is influenced by a number of other factors. A multi criteria decision analysis approach is applied to the Irish healthcare setting to aid decision making based on multiple criteria. Based on past decisions, key drivers of reimbursement decisions and their relative importance are identified. The approach improves consistency and transparency in the decision making process.

All analyses are carried out through a Bayesian framework, which is well suited for a decision theoretic approach to health technology assessments.

Acknowledgements

First of all I would like to express my gratitude to my supervisor, Professor Cathal Walsh. Cathal has patiently guided me through these 3 years and provided all the support and help I needed. He has given me every opportunity to broaden my knowledge and to attend and present my work at conferences all over the world. He has introduced me to a lot of interesting projects, which gave my work a real application. I am grateful for his continuous excitement about my work. Cathal is an excellent supervisor, and it has always been fun to work with him!

Thank you to Roisin Adams from the NCPE, for being an excellent partner for collaboration. It has been a lot of fun and you have motivated me throughout and shown me the importance of the work we are doing. It has been a great pleasure to work with Emer Fogarty from the NCPE on my yet largest evidence network!

I would also like to thank everyone in the Statistics department in Trinity, especially my fellow postgrads, and also the people from the NCPE for making these 3 years an enjoyable experience. I have made a number of very good friends along the way!

Thank you to Emer, Javi, Roisin, Sylwia and Triona for patiently proofreading my thesis on very short notice. I appreciate your time and motivational support!

Without the constant support and encouragement of my parents throughout my life, this thesis would not have been possible. Vielen Dank, Mama, Papa und Thomas, dass ihr immer an mich geglaubt und mich unterstützt habt!!

Finally, a very special thanks goes to my dear Javi. Your love, patience and support during this time have given me the strength to keep going. Thank you for always being there for me!

Contents

Abstract	iii
Acknowledgements	v
List of Tables	xi
List of Figures	xiii
Chapter 1 Introduction	1
1.1 Evidence Synthesis	2
1.2 Multiple Criteria Decision Analysis	3
1.3 Overview of Chapters	3
1.4 Research Contributions	5
Chapter 2 Health Technology Assessment	7
2.1 Introduction	7
2.1.1 HTA in Ireland	8
2.2 Economic Evaluation	8
2.2.1 Measuring costs and effects	9
2.2.2 Decision rules	9
2.3 Uncertainty in Economic Evaluations	12
2.3.1 Sources of uncertainty	13
2.3.2 Estimating uncertainty	13
2.3.3 Presenting Uncertainty	16
2.3.4 Value of Information	18
2.4 Other components of a HTA	19
2.4.1 Combining components	20
2.5 Summary	21

Chapter 3 Statistical Methodology	23
3.1 Bayesian Inference	23
3.1.1 Prior Distribution	25
3.1.2 Likelihood Function	25
3.1.3 Posterior Distribution	25
3.1.4 Prior Elicitation	26
3.2 Monte Carlo Methods for Approximating Posterior Distributions	27
3.2.1 Monte Carlo Integration	27
3.2.2 Markov Chains	27
3.2.3 Metropolis Hastings Algorithm	28
3.2.4 Gibbs Sampler	29
3.2.5 Convergence Diagnostics	29
3.2.6 Markov Models	30
3.3 Meta-Analysis	31
3.3.1 Simple Meta Analysis	32
3.3.2 Mixed Treatment Comparisons	38
3.4 Decision Theory	44
3.4.1 Bayesian Decision Theory	47
3.4.2 Multiple Criteria Decision Analysis	49
Chapter 4 MTC: Continuous versus Binary Outcome Measures	53
4.1 Motivation	53
4.2 Background	54
4.2.1 The intervention	54
4.2.2 Aim of Analysis	56
4.2.3 Efficacy Measures	56
4.3 Methods	58
4.3.1 Data	58
4.3.2 Statistical Model	59
4.3.3 Prior Distributions	62
4.3.4 Computation / Software	63
4.4 Results	63
4.4.1 Data	63
4.4.2 Efficacy Results	70
4.4.3 Convergence	75
4.5 Sensitivity Analysis	75

4.6	Limitations	80
4.7	Conclusions	80
Chapter 5 Impact of Structural Uncertainty in Economic Evaluations		83
5.1	Motivation	83
5.1.1	Application in Rheumatoid Arthritis	84
5.1.2	Aim of Analysis	84
5.2	Methods	84
5.2.1	Structural Choices in MTC	84
5.2.2	Economic Model	85
5.2.3	Prior distribution	87
5.3	Results	87
5.3.1	Efficacy Results	87
5.3.2	CUA Results	88
5.3.3	EVPI for structural uncertainty	90
5.3.4	Model Choice	93
5.4	Limitations	95
5.5	Conclusions	95
Chapter 6 MTC: Including observational evidence		97
6.1	Motivation	97
6.1.1	Aim of Analysis	98
6.1.2	Potential Bias in Observational Trials	98
6.2	Methods	99
6.2.1	Application in Rheumatoid Arthritis	99
6.2.2	Application in Multiple Sclerosis	99
6.3	Statistical Model	101
6.3.1	Naive Pooling	101
6.3.2	As Prior Information	102
6.3.3	Hierarchical Modelling	103
6.4	Results: Application in Rheumatoid Arthritis	105
6.4.1	Data	105
6.4.2	Naive Pooling	106
6.4.3	As Prior Information	109
6.4.4	Bayesian Hierarchical Model	110
6.4.5	Sensitivity Analysis	113

6.5	Results: Application in Multiple Sclerosis	115
6.5.1	Data	115
6.5.2	Computation / Convergence	117
6.5.3	RCT only	119
6.5.4	Naive Pooling	121
6.5.5	As Prior Information	123
6.5.6	Hierarchical Model	126
6.5.7	Sensitivity Analysis	127
6.6	Conclusions / Limitations	129
Chapter 7 MCDA		133
7.1	Motivation	133
7.1.1	Aim of Analysis	134
7.2	Methods	134
7.2.1	Data Collection	134
7.2.2	Statistical Analysis	135
7.3	Results	135
7.3.1	Data	135
7.3.2	Logistic Regression	141
7.4	Conclusions / Limitations	143
Chapter 8 Conclusions and Future Work		147
8.1	Conclusions	147
8.2	Further Work	148
8.2.1	Simulation Study	148
8.2.2	Inclusion of Registry Data in MTC	148
8.2.3	Normative Approach to MCDA	149
8.3	Final Remarks	150
Appendix A Code and Data for Chapter 4		151
Appendix B Code and Data for Chapter 6		157

List of Tables

3.1	Decision Table	46
4.1	Measuring improvement in RA	57
4.2	Data extracted from RCTs	64
4.3	Baseline Characteristics	66
4.4	Input Data for discretised HAQ score and continuous ACR	67
4.5	Mean improvement estimated from patient level data.	69
4.6	Results from HAQ analysis	71
4.7	Results from ACR analysis	72
4.8	Results from Sensitivity Analysis (1)	78
4.9	Results from Sensitivity Analysis (2)	79
5.1	Results of four alternative MTC models	88
5.2	Beta distributions for HAQ multipliers	89
5.3	Incremental cost and QALY estimates	89
6.1	RA case study: Observational trial data:	106
6.2	RA case study: Naive Pooling Results	107
6.3	RA case study: Node split results	109
6.4	RA case study: OBS data as prior Results	110
6.5	RA case study: Hierarchical Model Results	112
6.6	RA case study: Sensitivity Analysis	114
6.7	MS case study: Data extracted from RCTs	116
6.8	MS case study: Data from observational trials	119
6.9	MS case study: Results based on RCT only and Naive Pooling	121
6.10	MS case study: Node Split Results	123
6.11	MS case study: Results based on OBS data only	124
6.12	MS case study: Results based on OBS data as prior information	125
6.13	MS case study: 3-level model Results	126

6.14 MS case study: Sensitivity Analysis 129

7.1 MCDA: Details on criteria selected for MCDA 136

7.2 MCDA: Data extracted from assessment summaries 138

7.3 MCDA: Forward Stepwise Regression 141

7.4 MCDA: Regression coefficients 142

List of Figures

1.1	Schematic illustration of Decision Making	2
2.1	Cost-effectiveness plane	10
2.2	Non linear utility curve	12
2.3	Sources of uncertainty in an economic evaluation.	14
2.4	Tornado plot	15
2.5	PSA results on a cost-effectiveness plane.	16
2.6	CEAC Example	17
3.1	Time series plot of 2 chains of a converged MCMC algorithm.	30
3.2	Autocorrelation Plot.	31
3.3	Three stage markov model	31
3.4	Directed acyclic graph for meta analysis	33
3.5	Funnel plot	38
3.6	Indirect comparison: Simplest case	39
3.7	Alternative evidence structures	41
3.8	Evidence structure of closed loop design	42
3.9	Decision Tree	48
3.10	Decision analysis process	48
4.1	Joint affected by Rheumatoid arthritis	55
4.2	A hand affected by Rheumatoid arthritis	55
4.3	Selection Process: Flow Diagram for anti-TNF therapy in RA.	64
4.4	Network Diagram for RA analysis.	66
4.5	Forest Plot of HAQ results	73
4.6	Forest Plot of ACR results	74
4.7	Autocorrelation Plot	75
4.8	Trace plot	76
4.9	Posterior distribution of meta-regression parameter	76

5.1	Patient Pathway in BRAM model (taken from National Institute for Health and Clinical Excellence (2010))	86
5.2	Cost-effectiveness plane: Cost-effectiveness results of alternative models .	90
5.3	Scatter Plots from PSA for structural choices.	91
5.4	CEACs for alternative models	91
5.5	Expected Value of Perfect Information	93
5.6	EVPI of model choice	94
6.1	RA case study: Combined Network Diagram:	102
6.2	Three level hierarchical model	103
6.3	RA case study: Network diagram of observational evidence	105
6.4	RA case study: Inconsistency Check	108
6.5	RA case study: Adjusting for overprecision using OBS data as prior . . .	111
6.6	RA case study: Adjusting for overestimation using OBS data as prior . .	111
6.7	RA case study: Shrinkage from data	112
6.8	RA case study: Adjusting for overprecision in 3-level Model	113
6.9	RA case study: Adjusting for overestimation in 3-level Model: RA (2) . .	114
6.10	MS case study: Flowdiagram	116
6.11	MS case study: Evidence Network of RCT trials	118
6.12	MS case study: Evidence Network of observational trials	118
6.13	MS case study: Forest Plot of relative efficacies based on RCTs	120
6.14	MS case study: Forest Plot of results	122
6.15	MS case study: Inconsistency Check	123
6.16	MS case study: Adjusting for overprecision using OBS data as prior . . .	125
6.17	MS case study: Adjusting for overestimation using OBS data as prior . .	126
6.18	MS case study: Adjusting for overprecision in 3-level analysis	127
6.19	MS case study: Adjusting for overestimation in 3-level model MS	128
6.20	MS case study: Effect of shrinkage	128
7.1	MCDA logistic regression:	142
8.1	MCDA research plan	149

Chapter 1

Introduction

The aim of this thesis is to develop statistical methodologies which improve reimbursement decisions in healthcare. The allocation of healthcare resources is a very topical issue in the current economic climate, where cuts in the healthcare budget are inevitable. Pharmacoeconomics has to deal with the consequences of resource scarcity. Given a limited healthcare budget it is important to distribute scarce resources fairly and efficiently. A new healthcare intervention provides value for money when the cost per additional unit of health benefit does not exceed the currently least efficient technology, which is reimbursed. In most European countries health technologies are formally evaluated regarding their costs and benefits prior to reimbursement. The term describing this formal evaluation is Health Technology Assessment (HTA).

A HTA is a support tool aiding decision making in healthcare. Bayesian Methods are well suited for a decision theoretic approach to the problem. Figure 1.1 shows a schematic illustration of the decision making process and where the work of this thesis feeds in.

Applied to healthcare reimbursement decisions, the question mark represents the problem at hand, whether or not a particular intervention should be reimbursed. In order to evaluate this decision problem, available evidence needs to be assessed and a decision rule applied. $x(\cdot)$ represents the evidence associated with an intervention and its comparators. Based on current knowledge the decision maker has a prior distribution $\pi(M, \theta)$ on the model M and the parameters θ of the model. In an evidence synthesis relevant information X is collected and the prior is updated in light of the collected information. This yields a posterior distribution $f(\theta, M|X)$ of the model M and the parameter θ given the evidence X . In figure 1.1 this part of the process is marked blue. In order to take a decision based on the posterior distribution, a decision rule needs to be applied. A utility function $u(\cdot)$ indicates the utility of each possible outcome of the problem and allows for a ranking of the alternatives in terms of subjective, expected utility, $Rank(d_i)$ refers

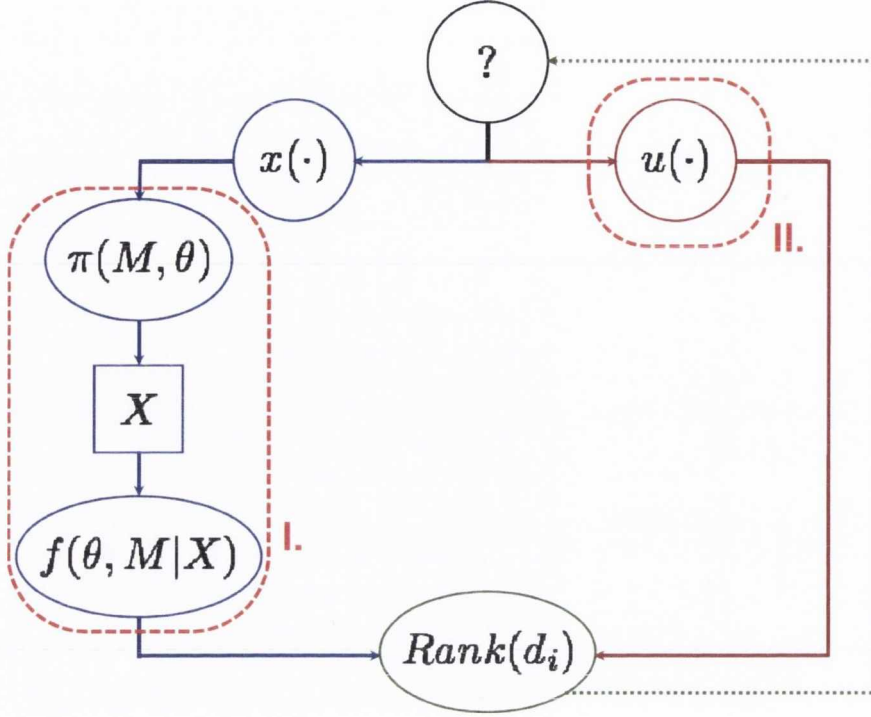


Fig. 1.1: Schematic illustration of Decision Making and where the work of this thesis feeds into the process.

to the rank of decision d_i . The decision should be taken such that subjective, expected utility is maximised; this refers to the red part of figure 1.1.

The work of this thesis feeds into both sides of this diagram. The impact of structural uncertainty in evidence synthesis has been assessed in relation to choice of outcome measure as well as structural choices within the model. While the current process only allows for the inclusion of randomized controlled trials (RCTs) as evidence, a methodology to allow for the inclusion of a wide variety of evidence sources has been developed. These aspects of the thesis fit into area I. of figure 1.1. The currently used utility function is the incremental net benefit; a measure combining the incremental costs and the incremental clinical effectiveness associated with an intervention. The work in this thesis proposes an alternative approach to utility, where other aspects influencing the decision on reimbursement (such as the overall budget impact, the quality of evidence or innovation) can be systematically included; this feeds into area II. of figure 1.1.

1.1 Evidence Synthesis

A key component of gathering the information for a HTA is the assessment of relative efficacy between alternative treatments. This is part of assessing the additional health

benefit a technology provides compared to relevant comparators and therefore provides the basis for estimating the cost per additional unit of health benefit. When direct evidence comparing alternative treatments is not available or sufficient, as is often the case, mixed treatment comparison (MTC) models provide a powerful methodology to estimate the relative efficacy between these treatments. MTCs are an extension of meta-analytical methods combining both, direct and indirect evidence. A major part of this thesis focusses on optimising and extending MTC methodologies with the aim of maximising the use of available evidence.

1.2 Multiple Criteria Decision Analysis

HTAs aim to aid informed decision making by assessing different aspects of health technologies. Clinical aspects, economic aspects, ethical aspects as well as the quality of evidence need to be assessed; one is therefore facing a multi-dimensional decision problem. A systematic approach to combining the multitude of objectives influencing healthcare decisions improves the transparency and coherence in the process. This is useful for the pharmaceutical industry on one side, because it is easier to predict the reimbursement decision at an earlier stage, but also benefits the decision maker, since it provides justification when decisions are challenged. In this thesis a multi criteria decision analysis (MCDA) approach is proposed for the Irish healthcare system to systematically incorporate relevant objectives.

1.3 Overview of Chapters

A brief outline of the research conducted follows.

Chapter 2: Health Technology Assessment

This chapter introduces the concept of HTA with a particular focus on the technical components. Following a motivational introduction, a brief outline of the HTA process in Ireland provides the conceptual context for this thesis. It follows a review of methodologies for economic evaluations, which comprises of measuring costs and health benefits and how to determine cost-effectiveness. Methodologies to assess the uncertainty in economic evaluations are introduced including probabilistic sensitivity analysis and value of information. A brief overview of other components of HTA is given to introduce the multiple dimensions of the decision problem.

Chapter 3: Statistical Methodology

This chapter details the statistical methods used throughout the analyses carried out for this thesis. All modelling is done in a Bayesian framework, which is particularly suitable for medical applications with an emphasis on decision making. The chapter begins with a detailed introduction to Bayesian inference. Models are fitted using MCMC techniques, which are described subsequently. Meta-analytical methods and extensions to mixed treatment comparisons are reviewed as well as an outline of Bayesian decision theory.

Chapter 4: MTC: Continuous versus Binary Outcome Measures

Dichotomized outcome measures result in a loss of power to detect relationships. While this is widely known in the statistical literature, this is of particular interest for MTC modelling for a number of reasons. Indirect comparisons suffer from an increased variance compared to direct comparisons, making it crucial to make the greatest use of available data. Furthermore, it has been pointed out that dichotomization can be used when designing trials to quantify a treatment effect, which is the primary aim of many clinical trials aiming for license approval. Further use of the trial data for pharmacoeconomic assessments is not taken into account. Using a case study in rheumatoid arthritis this chapter presents a MTC model on a range of efficacy measures illustrating the increased power of continuous outcome measures in such models compared to using binary measures.

Chapter 5: Impact of Structural Uncertainty in Economic Evaluations

All economic evaluations are subject to uncertainty caused by model parameters as well as the model itself. The real task of an HTA can hence be described as aiding decision making in the presence of uncertainty. The impact of structural uncertainty is widely neglected in evaluations, while different structural choices can influence the decision regarding cost-effectiveness as well as the uncertainty surrounding the decision. This chapter investigates the impact of structural choices in the MTC model on the decision of a subsequent cost-effectiveness analysis. Structural uncertainty is quantified using generalized value of information methods based on the parameterisation of structural choices.

Chapter 6: MTC: Including observational evidence

One of the incentives to combining evidence in a MTC model is to make an informed decision based on all available evidence. Nevertheless MTCs are typically restricted to

evidence from randomized controlled trials (RCTs), which are considered to be the gold standard of evidence, because the controlled setting minimizes potential bias. On the other hand, the experimental setting limits the external validity, while observational data reflects reality well. In this chapter three methods to systematically include evidence from different trial designs in a MTC model are developed. The effects of including evidence other than RCTs in a MTC and the advantages and disadvantages associated with the three methods are illustrated using two case studies. The application in rheumatoid arthritis from the previous chapters is extended and a further case study in multiple sclerosis illustrates the effects in a more complex evidence structure.

Chapter 7: MCDA in the Irish healthcare setting

The decision whether or not to reimburse a new technology depends on a number of criteria. A multi criteria decision analysis (MCDA) is a formal approach to synthesize multiple criteria in decision making, which improves the transparency, consistency and coherence compared to a case by case approach. In this chapter a descriptive approach is taken to identify and weight criteria based on guidelines and past decisions to inform a multi attribute value modelling approach to MCDA in the Irish healthcare setting. The work identifies criteria which have influenced reimbursement decisions in the past as well as their relative importance. This is work towards a normative approach which can be implemented in the Irish healthcare system.

1.4 Research Contributions

The following are the main contributions made by the research contained in this thesis:

1. The development of a MTC model for continuous outcome measures in rheumatoid arthritis. The impact of dichotomizing outcome measures in MTC modelling.
2. The inclusion of structural uncertainty in the economic modelling of anti-TNF treatment strategies in rheumatoid arthritis and the assessment of the impact of structural choices.
3. The incorporation of evidence from different trial designs is an important problem; the methodologies developed in this thesis provide a novel solution to this problem.
4. The development of a descriptive MCDA approach for HTA in the Irish healthcare setting quantifies the impact of different criteria on the reimbursement decision. This is a novel approach to decision making in the Irish healthcare setting.

Chapter 2

Health Technology Assessment

This chapter introduces the concept of Health Technology Assessment (HTA) with a particular focus on the technical components. A brief outline of the HTA process in Ireland is given to provide a conceptual context for this thesis. A review of methodologies for economic evaluations follows; models developed in chapters 4-6 aim to improve the process of evidence synthesis for such assessments. Other aspects of HTAs are outlined also; these form the basis for the multi criteria decision making approach to the Irish healthcare system discussed in chapter 7.

2.1 Introduction

The healthcare budget is the largest budget in Ireland; expenditure has increased from €5.7 billion in 2000 to €15.4 billion in 2009. €2.24 billion (15%) of this was spent on drugs in 2009. This growth rate is deemed unsustainable and the current financial situation makes budget cuts inevitable. The challenge is to maximise the benefit from the restricted budget. To ensure value for money, an evidence based, transparent decision making process is therefore very important.

A HTA aids decision making in healthcare by analysing new and existing interventions regarding multiple endpoints. The assessment thereby aims to derive the value of the technology relative to current clinical practice (Taylor and Taylor (2009)). A main focus lies in answering the following questions:

- How do the health outcomes of the technology compare with available treatment alternatives in terms of clinical effectiveness?
- Are these improvements in health outcomes commensurate with the additional costs of the technology?

HTA is a relatively new field, which has seen remarkable growth in the last two decades (Banta (2003)). Methods are now incorporated in national healthcare policy making processes regarding the reimbursement of drugs or other health technologies in Australia, Canada and most EU member states (Hutton et al. (2006)).

2.1.1 HTA in Ireland

Under the Health Act 2007, a statutory HTA agency was established as part of the Health Information and Quality Authority (HIQA). It is the agency's responsibility to set standards for quality and safety in health services including HTAs; published guidelines are available online (HIQA (2011, 2010a,b)).

The National Centre for Pharmacoeconomics (NCPE) was established in 1998. Funded by the Irish Department of Health and Children (DoH&C) the NCPE conducts all pharmacoeconomic assessments for the Health Service Executive (HSE) to promote cost-effective prescribing.

An agreement between the Irish Pharmaceutical Healthcare Association (IPHA) and the HSE allows the HSE to request an assessment of any medicines which may have a significant budget impact on the Irish healthcare system (IPHA (2010)).

Currently all new medicines are subject to a rapid review, products are then either reimbursed or referred to a full economic assessment (Tilson and Barry (2010)). A full assessment consists of an economic evaluation, typically in form of a cost utility analysis, a budget impact analysis, and a dynamic cost-effectiveness threshold allows for the inclusion of other aspects which may be of interest to the decision.

2.2 Economic Evaluation

An economic evaluation is a systematic process which compares the costs of an intervention with its consequences and therefore gives a guideline of whether the intervention is good value for money or not. The most common form of an economic evaluation is a cost utility analysis (CUA), which is the approach recommended in the national HTA guidelines.

Let T_1 and T_2 be two treatments which are to be compared. The mean economic cost of each treatment is $\bar{c}_i$ for $i = 1, 2$ and the mean clinical effects $\bar{e}_i$ for $i = 1, 2$.

The outcomes of interest in a CUA analysis are the incremental mean costs and the

incremental mean effects which are defined respectively as follows:

$$\begin{aligned}\Theta_c &:= \bar{c}_2 - \bar{c}_1 \\ \Theta_e &:= \bar{e}_2 - \bar{e}_1\end{aligned}\tag{2.1}$$

2.2.1 Measuring costs and effects

For the economic costs of a new drug all types of outgoings that affect the healthcare system need to be considered. This comprises not only the drug costs themselves, but also training for new equipment and staffing costs etc. Generally, costs that occur from a wider range, for example time away from work, are not considered in the analysis. These costs might be considered in alternative scenarios.

While it is straightforward to allocate a monetary value to the mean economic costs, it is less obvious how to measure the effects. A CUA uses a combination of mortality and quality of life: quality adjusted life years (QALYs).

Definition 2.1 *Quality adjusted life years (QALYs) are a scale used to quantify effectiveness. It includes both, the quality and the quantity of life. The score is based on the number of years that would be added by the intervention. An additional year of life at perfect health is allocated the value 1.0, death the value 0.0. Life years of not perfect health are given a value in between. The health status is assessed using a questionnaire such as EQ5D, “a standardised instrument for use as a measure of health outcome”, EuroQol Group (2011).*

The incremental costs and QALYs are often estimated using a Markov model. These simulation based models have the advantage over decision tree approaches of being able to model a time horizon of a life time. Evidence to inform inputs for these models may be sourced from a number of areas including natural history, epidemiology, efficacy and effectiveness and costs. Further modelling of the evidence is required, such as an evidence synthesis for effectiveness in order to combine different sources of data for use in the model. Further details on Markov models and meta-analysis for evidence synthesis are given in chapter 3. For the remainder of this chapter mean estimates for costs and QALYs are assumed to be known.

2.2.2 Decision rules

Once the values for the mean economic costs and mean clinical effectiveness for each treatment have been assigned, there are several ways of how to summarise the results of a CUA.

The standard has been the incremental cost-effectiveness ratio:

Definition 2.2 The incremental cost-effectiveness ratio (ICER) is defined by:

$$ICER := \frac{\Theta_c}{\Theta_e} = \frac{\bar{c}_2 - \bar{c}_1}{\bar{e}_2 - \bar{e}_1}$$

Being the ratio of incremental costs to incremental effectiveness the ICER represents the cost for one gained QALY by choosing treatment T_2 over treatment T_1 . Within a constraint budget, the healthcare payer must choose the intervention which will provide the most benefit for the least acceptable cost. A threshold K can be used to decide whether it is worth funding the treatment or not.

Definition 2.3 The threshold K represents the maximum acceptable price one is willing to pay for one additional life year at perfect health, i.e. for one QALY. It is called the willingness to pay (WTP). Beyond this threshold the intervention is not classed as good value for money. It was first introduced by Weinstein and Zeckhauser (1973).

There is a difficulty in interpreting the ICER which can be best explained looking at the cost-effectiveness plane in figure 2.1. The plane is divided into four quadrants corre-

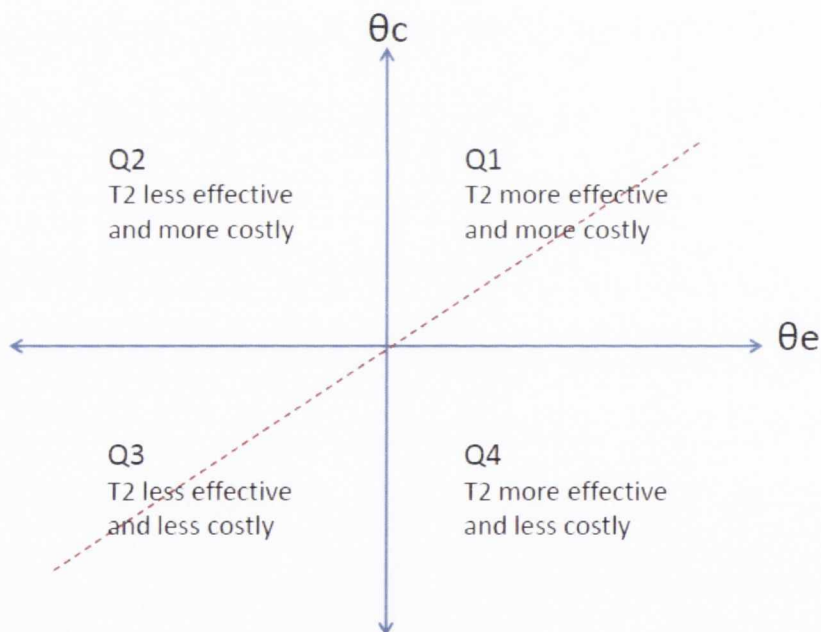


Fig. 2.1: Cost-effectiveness plane

sponding to the different signs of Θ_c and Θ_e . In quadrant $Q4$ one gains QALYs at a lower cost by using treatment T_2 , meaning T_2 dominates T_1 . In quadrant $Q2$ using treatment T_2 yields a loss of QALYs at a higher costs; T_1 dominates T_2 . In both quadrants the ICER is negative and the interpretation clear.

In quadrants $Q1$ and $Q3$ the ICER is positive. An ICER in $Q1$ means that treatment T_2

is more effective but costs more as well. Here the decision is taken using the threshold K . The dotted line in the graph has slope K , hence an ICER higher than K lies above the line and yields the choice of T_1 , while an ICER lower than K lies below the line and yields the choice of T_2 .

In quadrant $Q3$ treatment T_2 is less effective, but saves money. The question for the decision maker is whether the savings are high enough to justify the loss in QALYs; again the decision is taken using threshold K . However, in this case an ICER higher than K , which lies below the line, yields to the choice of treatment T_2 and an ICER lower than K , above the line, yields the choice of T_1 . This is where the confusion in the interpretation of the ICER arises. In $Q1$ high values for the ICER are supporting treatment T_1 , while in $Q3$ high values for the ICER are supporting treatment T_2 .

In order to avoid this confusion another summarising measure is introduced, the incremental net benefit.

Definition 2.4 *The incremental net benefit (INB) is a function of the willingness to pay threshold K and is defined by:*

$$INB(K) := K\Theta_e - \Theta_c$$

The INB converts the incremental effectiveness into a monetary value and then subtracts the incremental costs. Therefore it represents the incremental net monetary benefit.

Looking at figure 2.1 one can easily see that the INB is positive below the dashed line, where treatment T_2 is the preferred option and negative above the dashed line corresponding to the area where treatment T_1 is preferred. In terms of making decisions one would choose the treatment with the highest INB. In chapter 3.4 it will be shown that a function like the INB is called a utility function. The INB is one of many possible utility functions for this application. There is no need for it to be linear, a function as in figure 2.2 is a perfectly reasonable alternative. In this case the amount we are willing to pay per QALY depends on the total number of QALYs gained. While we are willing to pay a high amount for the first QALYs, we are not willing to pay as much for an additional QALY on top of an already high number of QALYs.

Remark 2.5 *Choosing the willingness to pay K requires a lot of care. K can neither be considered fixed nor random. Typically, the sensitivity to K is analysed using different values for K . The amount a healthcare system can spend per QALY gained obviously depends on the available budget and hence on the country, Spiegelhalter et al. (2004). The National Institute for Health and Clinical Excellence (NICE) in the UK considers*

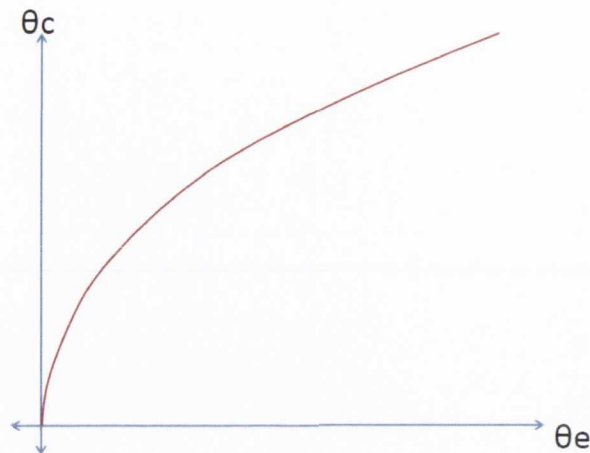


Fig. 2.2: Non linear utility curve

ICERs of less than £20,000 as cost-effective and discusses ICERs between £20,000 - £30,000. ICERs higher than this are approved only in very particular cases (National Institute for Health and Clinical Excellence (2004)).

Remark 2.6 In the threshold method the actual budget size is not considered. An alternative would be to line up the INBs of all new interventions and those with the highest INB would be successively picked until the budget is used. This, however, is not possible since not all INBs are available at the same time (Campbell (2010)). A more reasonable approach would be to consider opportunity cost by freeing up healthcare resources to pay for the more cost-effective option (Eichler et al. (2004)).

Remark 2.7 The INB is a utility function taking into account incremental costs and incremental QALYs. In section 2.4 at the end of this chapter a number of other criteria influencing the decision of reimbursement are introduced. Chapter 3 reviews methodologies of how multiple criteria influencing a decision can be included in a utility function. Such a methodology for the Irish healthcare system is developed in chapter 7.

2.3 Uncertainty in Economic Evaluations

As in all statistical analyses the estimates of the parameters of interest are surrounded by uncertainty. In this section different sources of uncertainty in a CUA are identified and ways of estimating and presenting the uncertainty are introduced.

Being aware of the uncertainty surrounding the estimates is important. However, economic evaluations require a decision theoretic approach and, as M. Campbell said in Campbell (2010), “in the end one has to make a decision. One usually cannot simply

‘not reject the null hypothesis’[...]”. The real task can hence be best described as making decisions in the presence of uncertainty. For further reading see O’Hagan and Luce (2003); Spiegelhalter et al. (2004); O’Hagan (2009).

2.3.1 Sources of uncertainty

Uncertainty in an economic evaluation arises from multiple sources. Adopting the categorisation proposed by Briggs (2000), one can distinguish between modelling uncertainty and parameter uncertainty, each of which can be further subcategorised.

Parameter uncertainty can relate to three different types of parameters, those regarding the choice of analytical method, those describing the patient sample, and those which could in principle be sampled from data. In modelling uncertainty one can distinguish between uncertainty based on the choice of the model as a whole and uncertainty which relates to the structure within a chosen model.

Bojke et al. (2006) give an overview of structural uncertainty in economic evaluations. A number of decision models commissioned by the NHS are reviewed with the aim of defining structural uncertainties. Model selection methods and model averaging are reviewed and a method to parameterize structural uncertainty as an extension of these methods is proposed to allow for the inclusion of structural uncertainty in value of information calculations. More details on these methods are described later in this chapter. Figure 2.3 gives a graphical representation of the categorisation and examples for each source of uncertainty.

2.3.2 Estimating uncertainty

The different types of uncertainty introduced in the previous section require different treatments. A brief outline of these actions is given in this section, for more detail more detail see Briggs (2000). A step by step guide to assessing uncertainty in economic evaluations is provided by Bilcke et al. (2011); starting with the identification and specification of uncertainties in the model, followed by the presentation of uncertainty in the model results and finally the assessment of which sources of uncertainty are the key drivers of the result.

Uncertainty relating to analytical methods, such as the applied discount rate, can be handled by well communicating the chosen method as a ‘reference case’. The sensitivity can be tested by employing a set of alternative methods, which are thought appropriate. In addition, this helps to compare different studies.

Patient sample characteristics do effect the economic evaluation, but are not uncertain-

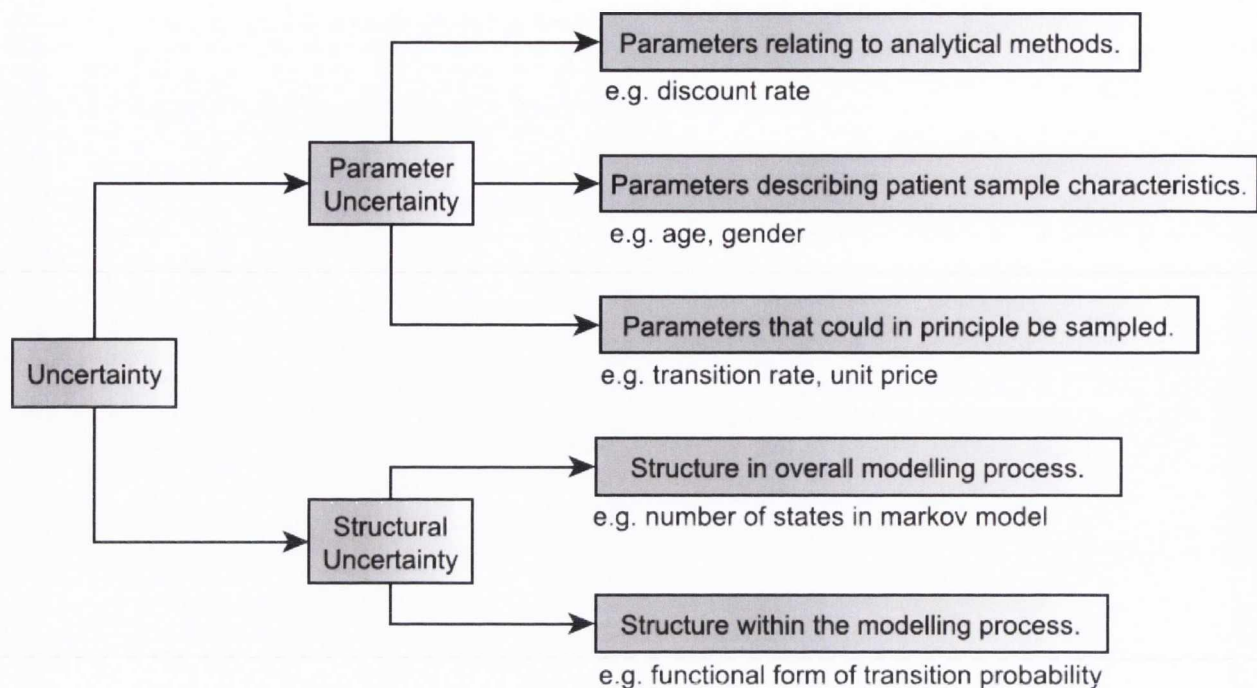


Fig. 2.3: Sources of uncertainty in an economic evaluation.

ties in itself. Issues relating to clinical margins (different reactions to an intervention due to clinical characteristics) or intensity margins (effect of different intensity levels of an intervention) should hence be dealt with outside of the formal analysis. Nevertheless, it is important to clearly specify the patient characteristics.

Uncertainty in the model structure can be addressed by modelling a set of alternative scenarios and investigating the impact on the decision, see for example Bojke et al. (2008) and a decision can be taken using model averaging (Jackson et al. (2010)), or parameterizing the set of options and address uncertainty in the same way as uncertainty in parameters which could in principle be sampled (Bojke et al. (2009)). Generally, one could deal with the overall model process uncertainty in a similar manner, however, one would need different teams of analysts in order to understand the surrounding uncertainty. Strong et al. (2012) have developed a method based on model discrepancy. The target outcome Z is modelled in terms of the model function $f(X)$, where X refers to the model inputs, and a discrepancy term δ_Z .

$$Z = f(X) + \delta_Z$$

The size of δ_Z is determined based on belief elicitation on sub-functional level. The advantage of this method is that structural uncertainty as a whole is considered rather than within a given set of model choices.

Uncertainty arising from parameters that could in principle be sampled is estimated by performing a sensitivity analysis.

One Way Sensitivity Analysis

In a one way sensitivity analysis one parameter at a time is varied over a range of different values and the model rerun to compare the outcomes. The method is useful to determine which parameters have most impact on the results of the model, e.g. drug costs, efficacy or survival rates. Results are often presented on a tornado diagram, figure 2.4. The basecase ICER is represented by a vertical line in the middle and extensions to the left or right indicate increase or decrease in ICER with varying parameter values. A one way

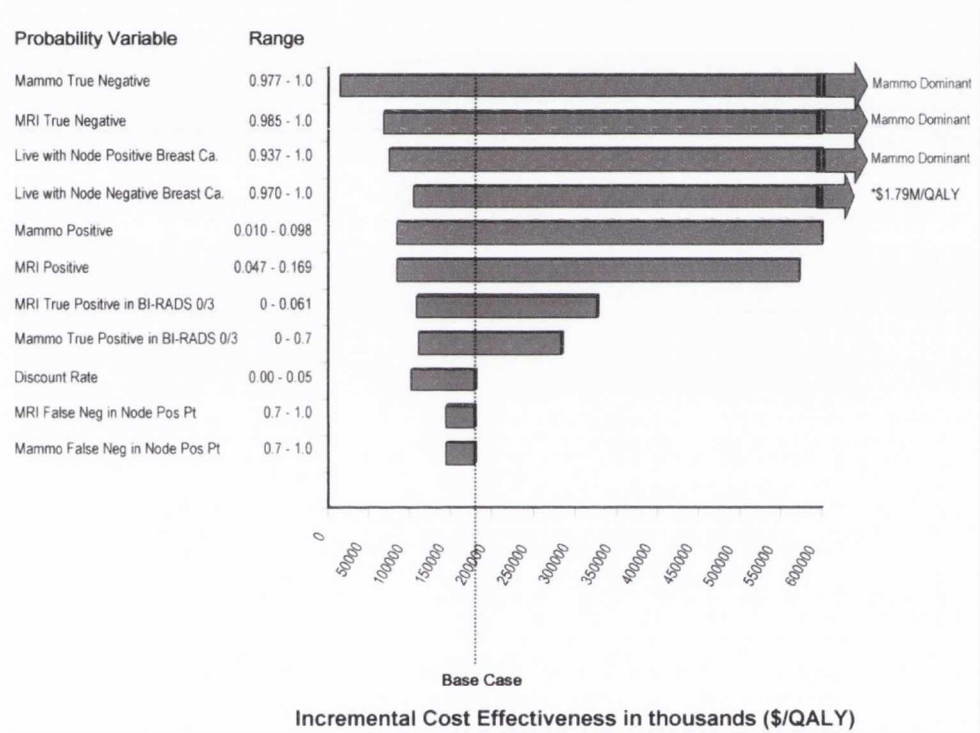


Fig. 2.4: Tornado plot demonstrating the parameters which increase or decrease the basecase ICER. (Taken from Moore et al. (2009))

sensitivity analysis is a rather superficial method, since joint impact of more than one varied input is not considered, even though it could be larger than the impact of just one varied variable. Furthermore, the true value of the input variable might not lie within the range of tested values. The preferred way of assessing uncertainty is a probabilistic sensitivity analysis (PSA); this is the recommended method by NICE (National Institute for Health and Clinical Excellence (2004)).

Probabilistic Sensitivity Analysis

A PSA allows exploration of the uncertainty associated with all parameters simultaneously. A probability density function is assigned to each input parameter representing the uncertainty in its true value. A random set of input values is drawn from these distributions and the model ran using these values. After repeating this process again and again a large sample from the output distribution is obtained, which can be used to infer about the uncertainty (Claxton (2008a)). A PSA is a very natural Bayesian analysis, since assigning probability distributions to unknown parameters is in effect a prior specification, this will be explained further in chapter 3.

2.3.3 Presenting Uncertainty

One way of visualising the results of a PSA is on a cost-effectiveness plane. A PSA provides a large sample from the joint posterior distribution of incremental effectiveness and cost $f((\Theta_e, \Theta_c)|data)$, from which a scatter plot can be produced. A schematic presentation can be found in figure 2.5.

Of interest is the probability of cost-effectiveness; the probability of treatment T_2 being

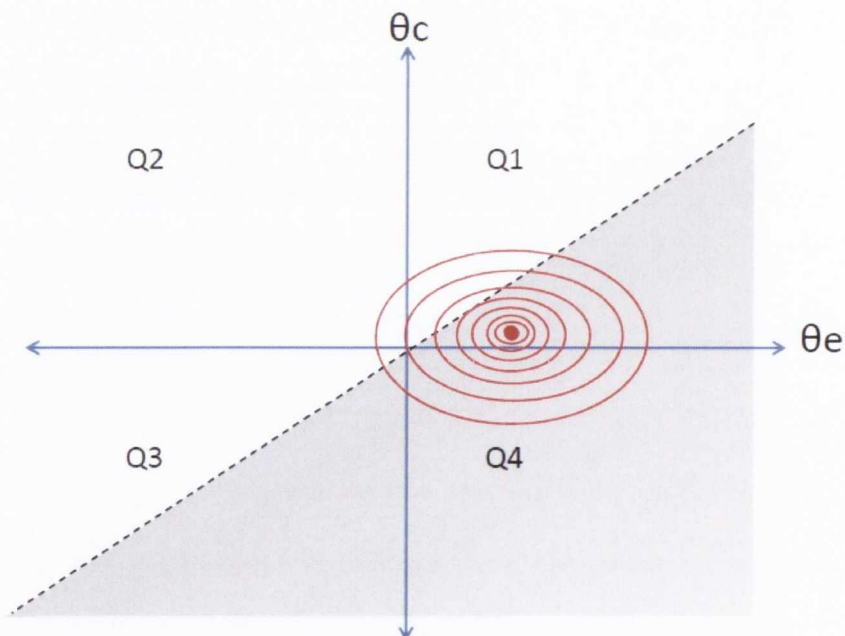


Fig. 2.5: PSA results on a cost-effectiveness plane.

cost-effective is the probability of the true values of (Θ_e, Θ_c) lying below the dashed line. Recalling that the INB is positive if and only if it lies below the line, naturally yields the decision theoretic approach of choosing treatment T_2 if the expected INB $E(INB(K))$

is positive.

Clearly, threshold K plays a keyrole in determining $E(INB(K))$. As mentioned in remark 2.5, the choice of threshold K is a complicated issue. Even though it is not a random quantity, it cannot be treated as a fixed value and therefore it makes sense to show how the outcome changes for different values of K .

One graph which has this required characteristic is the cost-effectiveness acceptability curve (CEAC). The CEAC plots the probability of a positive INB $P(INB > 0)$ against K .

Figure 2.6 shows the typical form of a CEAC. $P(INB > 0)$ on the cost-effectiveness plane



Fig. 2.6: CEAC Example

is the proportion of the joined distribution which lies in the shaded area of figure 2.5. The extreme case of $K = 0$ corresponds to a horizontal line, meaning $P(INB > 0)$ is equal to the probability of treatment T_2 being cheaper than treatment T_1 . The other extreme, $K = \infty$, corresponds to a vertical line and $P(INB > 0)$ is equal to the probability of treatment T_2 being more effective than treatment T_1 . In between these cases the shaded area in $Q1$ grows and the shaded area in $Q3$ shrinks. The CEAC in figure 2.6 is a typical form since in practise most probability functions are centered in $Q1$, in which case the CEAC increases.

In cases where more than 2 treatments are compared, the CEAC plots the probability for each of the treatments to yield the highest INB for given K .

The slope of the CEAC quantifies the uncertainty. A very steep slope means that one is quite certain for which values of K treatment T_2 is cost-effective and for which it is not; whereas a very flat slope means that the decision of cost-effectiveness is very sensitive to

the choice of K.

2.3.4 Value of Information

When faced with a high level of uncertainty in the process of deciding whether to adopt treatment T_1 or treatment T_2 the question arises whether funding further research will yield a more accurate decision. The benefits of further knowledge can be quantified as the potential cost associated with making the wrong decision. The benefit of knowing the true value θ of $\Theta = (\Theta_e, \Theta_c)$ is called value of perfect information, $VPI(\theta)$.

Let Ω_2 denote the shaded area in figure 2.5, where treatment T_2 is preferred, and Ω_1 the white area, where T_1 is the preferred treatment. Suppose the joint posterior distribution suggests that treatment T_2 is on balance the preferred treatment, i.e. $ENB(\Theta|data) > 0$.

In the case of $\theta \in \Omega_2$ the correct decision was taken and knowing θ would not bring any benefit. If, however, $\theta \in \Omega_1$ the decision was wrong and a loss of $-INB(\theta)$ occurs.

Definition 2.8 Value of information is defined as the potential loss when making the wrong decision.

$$VPI(\theta) = \max(-INB(\theta), 0).$$

The expected value of perfect information (EVPI) is defined as follows:

$$EVPI_2 = E[\max(-INB(\theta), 0)|data]$$

The EVPI can be calculated simulating values of Θ using MCMC (Claxton (2008a); Brennan et al. (2007)):

$$EVPI = \frac{1}{N} \sum_{n=1}^N \max_d INB(d, \theta_n) - \max_d \left(\frac{1}{N} \sum_{i=1}^N INB(d, \theta_n) \right)$$

The first part of this formula refers to the expected INB given perfect information. N is the number of PSA simulations; θ_n the parameter values of the n^{th} simulation. Given perfect information, the decision maximising the INB is taken in each simulation step. The average of these INBs is taken to calculate the expected INB given perfect information. The second part of the formula refers to the expected INB given current knowledge. Rather than choosing the decision which yields the highest INB for each simulation, the decision, which on average yields the highest INB is chosen. The EVPI is calculated as the difference in INB given perfect information and given current information. Further research is considered sensible if the involved costs do not exceed the EVPI.

The EVPI does not indicate the direction of required research. The expected value of

perfect parameter information (EVPPI) considers each parameter separately and thereby indicates where to invest in further research. Let $\theta = (\theta^i, \theta^c)$; assuming perfect information on only the subset of parameters θ_i , the EVPPI represents the expected gain of obtaining this information. The following formula can be used to estimate the EVPPI:

$$\frac{1}{K} \sum_{k=1}^K \max_d \left(\frac{1}{J} \sum_{j=1}^J INB(d, \theta_k^i, \theta_{j,k}^c) \right) - \max_d \left(\frac{1}{N} \sum_{i=1}^N INB(d, \theta_n) \right)$$

K values are sampled from the parameters of interest θ^i ; conditional of each θ_k^i J values of the remaining variables θ^c are sampled. The expected INB given perfect information on θ^i is calculated using nested loops. The second part of the equation representing the expected INB given current knowledge remains the same. The requirement of the outer and inner loop in the first part of the equation, makes the estimation of the EVPPI computationally expensive.

2.4 Other components of a HTA

So far it has been illustrated how to aid the decision of funding an intervention based on the incremental costs and clinical effects associated with the intervention compared to standard care. While the economic evaluation plays a dominant role in many healthcare decisions, the two dimensions of costs and effects do not stand alone. Some of the other aspects which are investigated in a HTA have been briefly mentioned in the beginning of this chapter. An overview of these aspects is given in this section.

Medical aspects

Medical criteria in a HTA encompass the efficacy and safety of an intervention. It has been shown how the relative efficacy of alternative treatments feeds into an economic evaluation. While a certain safety level needs to be demonstrated to gain licence approval from the European Medicines Agency (EMA) at an earlier stage, the costs and consequences of adverse events need to be accounted for in the economic evaluation.

Economic aspects

Each intervention is associated with a certain cost, including the actual drug cost, the cost of staff, training etc. In order to ensure value for money, the benefits of an intervention need to be balanced with the associated costs. The incorporation of efficacy and costs in an economic evaluation has been illustrated in the previous section.

Budget Impact

While an economic evaluation determines whether an intervention provides good value for money, the budget impact analysis (BIA) estimates the actual monetary impact the reimbursement would have on the payer. A technology may be classed as cost-effective but the budget impact prevents reimbursement. In 2008, for example, the cervical cancer vaccine was assessed in the Irish healthcare setting and was deemed cost-effective with a low ICER of €17,385 per life year gained (Usher et al. (2008)). However, the budget impact of €9.7 million was deemed unaffordable by the payer at that time (Tilson and Barry (2010)). Budget impact is mainly driven by the size of the eligible patient population.

Social/ Ethical aspects

The decision of reimbursement is also influenced by social or ethical circumstances. It is difficult to evaluate health states in children; the population is therefore generally willing to grant the benefit of the doubt when the patient population is children. In order to foster research one might increase the willingness to pay for very innovative treatments. The development of drugs for rare diseases (orphan drugs) needs to be encouraged, since it would otherwise not be commercially viable (Drummond et al. (2007)).

Quality of evidence

A decision based on the above analyses can only be as good as the underlying data. While Sensitivity analyses aim to capture the uncertainty surrounding the decision, it remains important to check the quality of the data to ensure unbiased estimates.

2.4.1 Combining components

Ireland adopts an implicit threshold approach. No threshold is officially stated, but it may be inferred from previous decisions made. The HSE is presented with summaries based on thresholds $K=€20,000$ and $K=€45,000$. Decisions are currently taken on a case by case method, a dynamic cost-effectiveness threshold allows for individual circumstances of each decision. This also allows for the consideration of the above components in the decision process.

A partially structured method to incorporate the different components may improve consistent and transparent decision making in healthcare, methods for multi criteria

decision analysis are reviewed in chapter 3 and applied to the Irish healthcare system in chapter 7.

2.5 Summary

HTAs aid resource allocation decisions in health care and are now incorporated in national policy processes in many countries including Ireland. This chapter introduces the concepts of an economic evaluation; how to measure costs and benefits and how to present the outcomes. Uncertainty surrounding the analysis can be assessed using one way sensitivity analysis and probabilistic sensitivity analysis. The outcomes of a probabilistic sensitivity analysis allow for the calculation of value of information estimates, which quantify uncertainty in monetary terms. This allows for the direction of further research. The value of a new intervention is not only determined by an economic evaluation, other aspects including ethical considerations, budget impact and the quality of available evidence also play a role.

Chapter 3

Statistical Methodology

Subsequent chapters make use of a variety of statistical methods. This chapter aims to provide a description of the theory of these methods. The chapter begins with an introduction to Bayesian inference. All modelling is done in a Bayesian framework, which is particularly suitable for medical applications with an emphasis on decision making (Ashby and Smith (2000); Ades et al. (2006)). Clinical evidence evolves over time; the use of a prior distribution allows for updating current knowledge as new data becomes available. Bayesian hierarchical models provide a natural way of synthesising evidence from different sources; which is necessary for an informed decision making process. A key component of decision making in medical applications is making predictions; predictions can be modelled in a straightforward manner using Bayesian statistics. Models are fitted using MCMC techniques, which are described in the second section. Subsequently, meta-analytical methods and extensions to mixed treatment comparisons are presented. Also presented is an outline of the Bayesian decision theory which is the framework for integrating MCDA techniques in the Irish healthcare system.

3.1 Bayesian Inference

Statistics is concerned with inference, the act of deriving a statement about some unknown features of a physical system based on available information (O'Hagan (1994)). For a long time the way to carry out inference was using so-called classical or frequentist methods, but since the middle of the 20th century the interest in an alternative, the Bayesian approach, has increased. It is now a widely accepted and established practice providing very natural and flexible methods to statistical problems. Bayesian inference and classical inference differ from each other in several major aspects.

Interpretation of probability: Probability is one of the core expressions in statistics.

However, the two approaches interpret it differently. The classical framework views probability as a property of an event. Every person estimating this probability will agree on the same value. A Bayesian on the other hand will speak of the probability for an event, rather than of an event. Probability is said to be subjective, reflecting someone's current belief for an event to happen based on his current knowledge. Therefore, different people will derive different probabilities for the same event.

Interpretation of parameters: Both approaches aim to make inference on a parameter of interest which represents a feature of the physical system. The definition of a parameter, however, differs. The classical parameter is an unknown, but fixed value, hence an estimate is given in form of a point estimate together with a range representing the uncertainty. The Bayesian parameter is treated as a random variable. The estimate is therefore given in the form of a probability distribution.

Information: Inference on a parameter is based on available information. Classical inference bases its inference solely on data which was collected for the analysis. Bayesian inference, on the other hand, combines this data with prior knowledge. For example, if the aim was to determine the proportion of Dubliners using a bike as means of transport, the classical analyst would ask a sample of Dubliners regarding their bike use and base inference on this data. A Bayesian would include his prior belief of the proportion, which could for instance be the proportion which was estimated for a comparable city.

In healthcare one typically learns from experience, it is essential to combine multiple sources of evidence and the statistical models are becoming increasingly complex. While standard methods have difficulties dealing with these issues, the flexible Bayesian perspective seems to be particularly suitable (Spiegelhalter et al. (2004)).

The foundation of Bayesian statistics lies in Thomas Bayes' theorem which was published in 1763, shortly after his death.

Theorem 3.1 *Consider two events A and B . Then*

$$P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)}$$

which follows directly from the product rule:

$$P(A, B) = P(A|B) \cdot P(B) \tag{3.1}$$

For the purposes in this thesis, the theorem needs to be rewritten in terms of random variables (RVs). Let $\theta = (\theta_1, \dots, \theta_p)$ be the parameter one is interested in making inference about and $X = (X_1, \dots, X_n)$ some observed data. Then:

$$f(\theta|X) = \frac{\pi(\theta) \cdot f(X|\theta)}{f(X)} \propto \pi(\theta) \cdot f(X|\theta). \tag{3.2}$$

$\pi(\theta)$ is called prior probability, $f(X|\theta)$ the likelihood and $f(\theta|X)$ the posterior probability. $f(X)$ is a normalizing constant given by:

$$f(X) = \begin{cases} \int \pi(\theta) \cdot f(X|\theta) d\theta & \text{(continuous RV)} \\ \sum_{\theta} \pi(\theta) \cdot f(X|\theta) & \text{(discrete RV)} \end{cases}$$

Using these terms Bayes' theorem can be interpreted as a formula for updating from prior knowledge to posterior knowledge by multiplying with $\frac{f(X|\theta)}{f(X)}$. It therefore describes how a probability changes as new information becomes available.

Lee (1997) provides an accessible introduction to Bayesian inference; for more comprehensive texts see O'Hagan (1994); Bernardo and Smith (2000).

3.1.1 Prior Distribution

In almost all situations one has some knowledge about the parameter of interest θ even before collecting and analysing new data. This knowledge can be based on an experts' opinion, previous studies may have analysed the same effect or the analyst's own knowledge of the physical system may provide a range of plausible values. This prior belief is summarised using a probability density function $\pi(\theta)$ and is called prior probability.

The prior is not unique nor necessarily plays an important role in the process of inference. A whole range of different priors is available to express different reasonable opinions, ranging from sceptical over non informative priors to enthusiastic priors. Depending on the aim of the analysis different prior distributions should be adopted, see for example Spiegelhalter (2004). The more data becomes available the less influence the prior has on the result.

3.1.2 Likelihood Function

The likelihood function $f(X|\theta)$ is given by the joint probability density of the observed data $X = (X_1, \dots, X_n)$ given the parameters. As a function of the data given θ it is a probability density. The likelihood is considered as a function of the parameter θ given the data, which does not integrate to 1. For each value of θ it tells how likely it is to observe data X .

3.1.3 Posterior Distribution

The prior distribution updated by the likelihood yields the posterior distribution $f(\theta|X)$; the updating process is given by Bayes' theorem:

$$\text{posterior} \propto \text{prior} \cdot \text{likelihood}$$

The division of the right hand side by the normalizing constant $f(X)$ ensures that the posterior integrates to 1. The posterior represents one's updated state of knowledge about the parameter as one has learned about the data. The posterior distribution contains all available information about the parameter of interest. Therefore the task of inference is to provide suitable summaries of the posterior. Any summaries of the posterior distribution may be of interest for inference; moments, quantiles or highest posterior density (HPD) intervals. All of these can be expressed in terms of posterior expectations of functions of θ and take the following form:

$$E[g(\theta|X)] = \frac{\int g(\theta)\pi(\theta)f(X|\theta)d\theta}{\int \pi(\theta)f(X|\theta)d\theta} \quad (3.3)$$

3.1.4 Prior Elicitation

Prior elicitation is the process of constructing a prior distribution which reflects one's background information. The two steps involved in the process are firstly the selection of a distribution and secondly the specification of the parameters of the chosen distribution. The choice of distribution is generally governed by mathematical tractability constraints. For this reason conjugate priors are often chosen.

Definition 3.2 *A prior distribution is conjugate if it belongs to the same class of distributions as the posterior distribution.*

Using a conjugate prior has the advantage that when updating the prior distribution to the posterior distribution, only the parameter values change. The class of conjugate priors is often large enough to provide a good approximation of the “true” prior (Lee (1997)). Non conjugate prior distribution typically yield a posterior distribution which is not of closed form. In cases where a conjugate prior is not an appropriate choice, numerical methods can be used to sample from the posterior distribution.

Choosing the parameters for the prior requires great care. Expert opinion can be used to elicit various properties of the parameter (e.g. mean, quantiles) which are then used to fit the distribution. When little information is available, the specification of a non-informative prior is required. A non-informative prior is defined such that no information regarding the parameter in question is added to the analysis. When the support of the parameter is finite, a uniform distribution over this range is a typical choice. For other classes of distributions it is possible to choose parameters which result in a very flat distribution, giving almost equal probabilities to a wide range of plausible values.

3.2 Monte Carlo Methods for Approximating Posterior Distributions

Monte Carlo methods refer to a class of numerical methods useful for obtaining numerical solutions to problems too complicated to solve analytically. Direct simulation techniques such as rejection sampling or importance sampling draw exact samples from the distribution of interest, however, such methods break down as the dimension of θ increases. Markov Chain Monte Carlo (MCMC) is an approximate sampling technique, which makes use of markov chains to draw random samples from a density which asymptotically approaches the density of interest. Using MCMC it is possible to solve problems of higher dimension. Gilks et al. (1996) provide a good introduction and overview to MCMC.

3.2.1 Monte Carlo Integration

Monte Carlo integration is a numerical integration technique using random numbers. It is useful for evaluating expressions of the form in equation 3.3. For simplicity assume the aim is to evaluate the expectation of $g(\theta)$, whose parameters have a density function $\pi(\theta)$:

$$E[g(\theta)] = \int g(\theta)\pi(\theta)d\theta \quad (3.4)$$

Monte Carlo integration draws independent samples $\{\theta_t, t = 1, \dots, m\}$ from $\pi(\cdot)$ and uses these to approximate the expectation:

$$E[g(\theta)] \approx \frac{1}{m} \sum_{t=1}^m g(\theta_t) \quad (3.5)$$

The law of large numbers ensures that this approximation becomes as accurate as necessary with increasing m (Robert and Casella (2004)). It is typically not feasible to draw an independent sample, but it actually suffices if samples are generated by a process which draws them through the support of the distribution in the correct proportions (Gilks et al. (1996)).

MCMC makes use of this through a markov chain with $\pi(\cdot)$ as the stationary distribution.

3.2.2 Markov Chains

Definition 3.3 A stochastic process is a set of random variables $\{\theta_t : t \in T\}$ where T is called the index set. Each $\{\theta_t\}$ takes a value in a set S , called the state space. This process is known as a discrete time stochastic process when the index set T is countable

Definition 3.4 A discrete time stochastic process $\{\theta_t : t \in T\}$ is called a Markov Chain with countable state space S if:

1. $P(\theta_n \in S) = 1 \quad \forall n \geq 0$

2. θ_{t+1} is sampled from a distribution which only depends on the current state θ_t :

$$P(\theta_{t+1}|\theta_0, \theta_1, \dots, \theta_t) = P(\theta_{t+1}|\theta_t)$$

This is known as the Markov property.

A Markov chain is homogeneous if the conditional distribution $P(\theta_{t+1}|\theta_t)$ is independent of t . In such cases the probability distribution $P(\cdot|\cdot)$ is called the transition kernel.

Definition 3.5 • A markov chain is irreducible, if for any starting state $s_0 \in S$ there is a positive probability to reach every region of the state space S .

- A state s has period k if any returns to state s must occur in multiples of k . If $k = 1$, a state is called aperiodic.
- A state $s \in S$ is recurrent, if a markov chain starting in s has a probability of 1 to return to state s . A recurrent state is positive recurrent if the expected time to return to this state is finite. A markov chain is called positive recurrent if all states are positive recurrent.
- A homogeneous markov chain called ergodic, if it is irreducible, aperiodic and positive recurrent.

An ergodic markov chain of random variables will converge to a unique stationary distribution $\Phi(\cdot)$ which does not depend on starting value θ_0 or t . Therefore, there exists an s such that $\{\theta_t, t > s\}$ are sampled from distribution $\Phi(\cdot)$. The so-called burn-in $\{\theta_t, t \leq s\}$ are discarded and the remaining sample is used to estimate $E[g(\theta)]$, where θ has distribution $\Phi(\cdot)$:

$$E[g(\theta)] \approx \frac{1}{m-s} \sum_{t=s+1}^m g(\theta_t) \quad (3.6)$$

This provides a solution to the problem in equation 3.5 if we can construct a markov chain with stationary distribution $\pi(\cdot)$.

3.2.3 Metropolis Hastings Algorithm

Using the Metropolis Hastings algorithm it is possible to construct a markov chain with the stationary distribution being the distribution of interest $\pi(\cdot)$. It was introduced by Hastings (1970) as a generalisation of the method by Metropolis et al. (1953).

The idea is quite simple. At each step $(t + 1)$ a sample candidate point Y is drawn

from a proposal distribution $q(\cdot|\theta_t)$. This sample candidate is accepted as θ_{t+1} with a probability $\alpha(\theta_t, Y)$:

$$\alpha(\theta_t, Y) = \min \left(1, \frac{\pi(Y)q(\theta_t|Y)}{\pi(\theta_t)q(Y|\theta_t)} \right) \quad (3.7)$$

The proposal distribution $q(\cdot|\cdot)$ is of arbitrary form, however, the rate of convergence to the stationary distribution $\pi(\cdot)$ will depend crucially on the relationship between $q(\cdot|\cdot)$ and $\pi(\cdot)$. $q(\cdot|\cdot)$ should be chosen such that it can be easily sampled from and evaluated to increase computational efficiency.

3.2.4 Gibbs Sampler

Throughout this thesis models are fitted using WinBUGs, a software package using MCMC methods including Gibbs sampling (Lunn et al. (2000)). Gibbs sampling was first introduced by Geman and Geman (1984). It is a special case of single component Metropolis Hastings, which consists of iteratively updating each component of $\theta = (\theta_1, \dots, \theta_p)$. Let $\theta_{t,i}$ be the state of the i th component of θ at the end of iteration t and

$$\theta_{t,-i} := \theta_{t+1,1}, \dots, \theta_{t+1,i-1}, \theta_{t+1,i+1}, \dots, \theta_{t,p}$$

be the set of all components of θ_t except $\theta_{t,i}$ at the i^{th} step of the $t+1^{th}$ iteration. Each iteration consists of p updating steps where $\theta_{t,i}$ is updated in step i of iteration $t+1$ using Metropolis Hastings algorithm. A candidate point Y_i for $\theta_{t+1,i}$ is drawn from the proposal distribution

$$q_i(Y_i|\theta_{t,i}, \theta_{t,-i}).$$

The candidate is accepted with probability

$$\alpha(\theta_{t,-i}, \theta_{t,i}, Y_i) = \min \left(1, \frac{\pi(Y_i|\theta_{t,-i})q_i(\theta_{t,i}|Y_i, \theta_{t,-i})}{\pi(\theta_{t,i}|\theta_{t,-i})q_i(Y_i|\theta_{t,i}, \theta_{t,-i})} \right)$$

where $\pi(Y_i|\theta_{t,-i})$ is the full conditional distribution for $\theta_{t,-i}$ under $\pi(\cdot)$. If Y_i is accepted, set $\theta_{t+1,i} = Y_i$, else $\theta_{t+1,i} = \theta_{t,i}$.

Gibbs sampling is a special case of this method drawing from the full conditional distribution:

$$q_i(Y_i|\theta_{t,i}, \theta_{t,-i}) = \pi(Y_i|\theta_{t,-i}) \quad (3.8)$$

which gives an acceptance probability of 1, because draws from the conditional distribution are always accepted.

3.2.5 Convergence Diagnostics

Inference based on MCMC simulations is only valid if the markov chain has converged to the target distribution. Monitoring convergence therefore is very important to determine

the number of burn-in iterations and to ensure valid inference. There are a number of convergence diagnostics, useful overviews of these methods can be found in Cowles and Carlin (1996); Brooks and Roberts (1998). All methods can fail to detect lack of convergence. The common advice is to apply a combination of diagnostic tools. This chapter will only discuss a small number of diagnostics, which are used in the remainder of the thesis.

Gelman and Rubin (1992) developed a diagnostic tool to test whether the markov chain has forgotten its starting point. The idea is to run multiple chains from overdispersed starting points and check that they are indistinguishable. Visually this can be done by checking if the 2 sequences can be distinguished on a time series plot. An example of a well behaved plot can be found in figure 3.1. A more quantitative approach is based

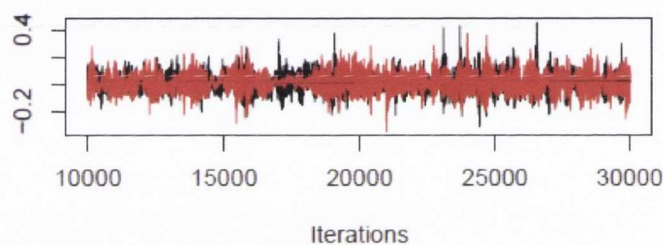


Fig. 3.1: Time series plot of 2 chains of a converged MCMC algorithm.

on the analysis of the variance. It is concluded that the algorithm has converged if the variance between the different chains does not exceed the variance within each individual sequence.

High autocorrelations within each chain indicate slow mixing and slow convergence. In order to explore the complete parameter space, a larger sample needs to be drawn. Autocorrelation can be assessed by looking at autocorrelation plots. Well behaved plots should show a decline in autocorrelation as lags decrease, an example can be found in figure 3.2. Autocorrelation can be improved by thinning, i.e. only using every k^{th} iteration of each chain. Raftery and Lewis (1992) developed a method to determine the number of burn-in iterations and the thinning given the desired accuracy.

3.2.6 Markov Models

Markov processes are often used for modelling random processes that evolve over time, and are particularly suitable to model chronic diseases (Sonnenberg and Beck (1993); Briggs and Sculpher (1998)). The state space consists of different stages in the disease; the transition probability defines the probability of moving between disease states over a

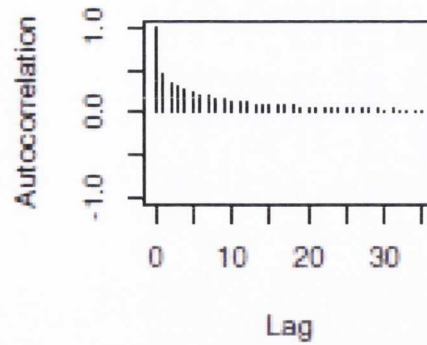


Fig. 3.2: Autocorrelation Plot.

discrete time period. Death is included as an absorbing state. Each state is associated with a cost and a QALY and the time spent in each state after running the model for a large number of cycles allows the estimation of total cost and QALY associated with the intervention. Figure 3.3 shows the diagram of a simple markov model. Patients that are

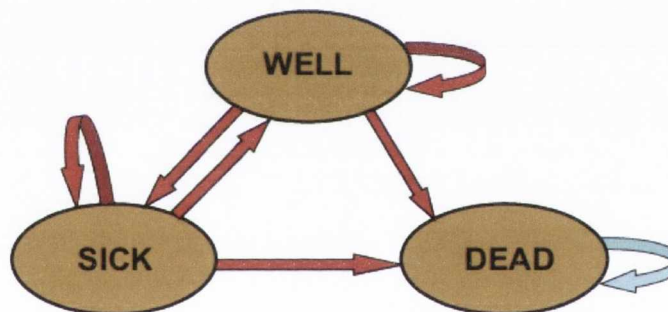


Fig. 3.3: Three stage markov model

well, can either remain well, become sick or die. Sick patients can either recover, remain sick or die.

3.3 Meta-Analysis

Methods for synthesizing evidence have recently been highlighted as one of the key priorities of research in pharmacoeconomics (Hughes (2012)). The statistical model for evidence synthesis is a meta-analysis (MA). This section provides an overview of current methodologies for meta-analysis, which will be extended in subsequent chapters to improve evidence synthesis in pharmacoeconomics.

Bayesian hierarchical models

A hierarchical model distinguishes between multiple levels of information, which allows for the modelling of hierarchical or clustered data (Goldstein (2010)). Many kinds of data are of this form, for example genetic data from a population can be clustered by family structures, i.e. youngest generation refers to level 1, their parents to level 2 etc. Meta-analysis is one form of hierarchical modelling, where level one refers to the effect observed in each individual study i and level 2 refers to the overall effect based on pooling the individual study effects.

Modelling hierarchical data requires the introduction of hyperparameters. Let y_{ij} be the i^{th} observation in cluster j of level 1; each cluster yields an estimate of interest θ_j . These θ_j are interpreted as an unobservable sample from the population distribution, which is defined in terms of hyperparameters Ψ , which in turn need to be assigned a prior distribution. This yields the following structure:

$$\begin{aligned} y_{ij} &\sim f(y_{ij}|\theta_j) \\ \theta_j &\sim \pi(\theta|\Psi) \\ \Psi &\sim \pi(\Psi) \end{aligned} \tag{3.9}$$

The joint prior distribution of (θ, Ψ) is

$$\pi(\theta, \Psi) \sim \pi(\theta|\Psi)\pi(\Psi)$$

yielding the joint posterior distribution

$$\begin{aligned} f(\theta, \Psi) &\propto \pi(\theta, \psi)f(y|\Psi, \theta) \\ &\propto \pi(\theta, \psi)f(y|\theta) \end{aligned} \tag{3.10}$$

For further reading see Gelman and Hill (2007), a general introduction is provided by Greenland (2000).

3.3.1 Simple Meta Analysis

Most of the time a single study group is far too small to allow a definite decision regarding the involved error. MA solves this problem by combining trials evaluating the same intervention (Egger et al. (2002)).

A big issue MA has to deal with is making sure that the resulting estimate remains unbiased. Even if one only combines data from randomized controlled trials (RCTs) the problems of heterogeneity between study populations and publication bias remain.

There are two ways of conducting a meta-analysis: A fixed effects setting and a random effects setting. Furthermore, one can build a model either in a classical or in a Bayesian

environment. In the following, an explanation of the differences of these types and how to figure out which is the most suitable way for a particular situation is provided. A motivational introduction is given in Rosenthal and DiMatteo (2001). A guideline and an overview of methods can be found in Sutton, Duval, Tweedie, Abrams and Jones (2000); further reading can be found in Sutton and Abrams (2001) and Sutton and Higgins (2008).

A fixed effects model can be considered if there is no heterogeneity between the study populations.

Model 3.6 Assume there are N studies measuring the effect of interest. In a fixed effects (FE) meta analysis the treatment effect estimated in each study is assumed to be the same across all studies. Mathematically this relates to the following model:

$$\begin{aligned}\Delta_c[i] &\sim N(\mu_c[i], \tau_c[i]), & \Delta_t[i] &\sim N(\mu_t[i], \tau_t[i]) \\ \mu_c[i] &= \nu[i] \\ \mu_t[i] &= \nu[i] + \delta[i] \\ \delta[i] &= d\end{aligned}$$

where $\Delta_c[i]$ and $\Delta_t[i]$ refer to the observed treatment effect in control and treatment arm of study i and $\tau_c[i]$ and $\tau_t[i]$ are the corresponding precisions. d is the estimated overall treatment effect.

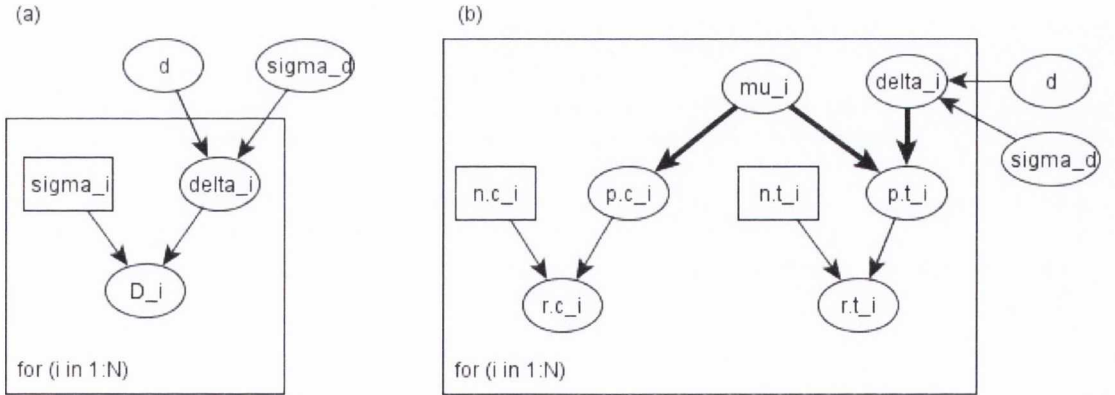


Fig. 3.4: DAG for (a) continuous and (b) binary meta-analysis model.

If homogeneity within the study populations cannot be assumed, a random effects meta-analysis should be conducted. This model was first introduced by DerSimonian and Laird (1986).

Model 3.7 A random effects (RE) model assumes that each study i , $i = 1, \dots, N$, estimates its own effect $\delta[i]$. These study estimates together form the estimate of a common population mean d . The model therefore allows for the study populations to differ slightly from another. In mathematical terms the model takes the following form:

$$\begin{aligned}\Delta_c[i] &\sim N(\mu_c[i], \tau_c[i]), & \Delta_t[i] &\sim N(\mu_t[i], \tau_t[i]) \\ \mu_c[i] &= \nu[i] \\ \mu_t[i] &= \nu[i] + \delta[i] \\ \delta[i] &\sim N(d, \sigma_\delta^2)\end{aligned}$$

The difference to the fixed effects model is that the study effects $\delta[i]$ are assumed to be normally distributed with a common mean d rather than constant. σ_δ^2 is the between trial variance parameter. Figure 3.4 (a) shows the corresponding directed acyclic graph (DAG) for this model.

Definition 3.8 A directed acyclic graph (DAG) allows for the graphical representation of a model, which fully describes the joint distribution of all unknown quantities. Each quantity is represented by a node; square nodes refer to constants, oval nodes to random quantities. Arrows show stochastic dependencies between nodes.

The acceptance of heterogeneity within the study populations is an advantage and applies to most of the applications, however, since more variability is allowed for in the model, results are less precise. However, where between trial heterogeneity is small, both models give virtually identical results.

The above models assume a continuous outcome measure in the trials. Clinical trials very often report binary outcome measures, in such situations the model formulation changes slightly.

Remark 3.9 Please note that for binary outcome measures the first three lines in each model is replaced by

$$\begin{aligned}r_c[i] &\sim \text{Bin}(n_c[i], p_c[i]), & r_t[i] &\sim \text{Bin}(n_t[i], p_t[i]) \\ \text{logit}[p_c[i]] &= \nu[i] \\ \text{logit}[p_t[i]] &= \nu[i] + \delta[i]\end{aligned}$$

where $n_c[i]$, $n_t[i]$ and $r_c[i]$ and $r_t[i]$ are the number of patients enrolled and the number of responders in control and treatment arm of study i , respectively. δ_i in this case is the log odds ratio of treatment arm compared to the control arm in study i . The corresponding DAG is shown in figure 3.4 (b).

The pooled estimate d is calculated as a weighted average of the individual study effects. There are several possibilities of calculating these estimates.

The inverse variance method uses the inverse of the within study variance as weights, yielding the following estimate for the treatment effect:

$$\hat{d} = \frac{\sum (w_i \cdot \Delta_i)}{\sum w_i}$$

where $w_i = \frac{1}{\sigma_i^2}$ the inverse of the within study variance.

For binary outcome measures the Mantel Haenszel method provides an estimate for the overall odds ratio (OR) and the associated variance parameter (Mantel and Haenszel (1959); Robins et al. (1986)). Another method for pooling odds ratios is the Peto method (Yusuf et al. (1985)). It uses an inverse variance approach but utilizes an approximate method of estimating the log odds ratio; different weights are used also. Methods by Warn et al. (2002) allow for the estimation of other outcomes such as risk ratios or risk differences.

The models described in 3.6 and 3.7 are classical models. Both can easily be implemented in R; using packages ‘meta’ or ‘rmeta’.

Conducting a Bayesian analysis has the advantage of being able to include the whole variety of available data (Higgins and Whitehead (1996)). The subjectivity of the Bayesian approach is sometimes seen as a disadvantage, since eliciting prior knowledge typically is not a trivial task. However, when sufficient data is available, the impact of the prior distribution becomes very small and both approaches yield very similar results. The mathematical models for fixed effects and random effects remain the same as in 3.6 and 3.7, one simply needs to define additional prior distributions on d and σ_δ .

Assessing Heterogeneity

As mentioned earlier fixed effects can be used if the populations of the studies do not systematically differ from one another. If one cannot assure this, random effects should be used. Most of the time it is not obvious whether there is heterogeneity among the populations. There is no reason why they should be different, but one cannot be certain. In this situation, how can one decide which model to use?

In the classical framework there exist statistics to test for heterogeneity (Higgins and Thompson (2002)). When conducting the analysis in R, the summary function automatically displays two of these: The I^2 statistic and the Q statistic. The I^2 statistic was developed by Higgins et al. (2003) and provides the percentage of variation across studies that are due to variation rather than chance.

Definition 3.10 The I^2 statistic is defined as

$$I^2 = 100\% \frac{Q - df}{Q}$$

where $df=N-1$ are the degrees of freedom when combining N studies and Q is the Q -statistic. The Q -statistic is the weighted sum of squared differences of individual study effects to the pooled estimate.

It is χ^2 -test with $N - 1$ degrees of freedom. The χ^2 -test has low power when few studies are combined. In the Bayesian framework there are no such statistics. Alternatively, the deviance information criterion (DIC) can be calculated (Spiegelhalter et al. (2002)).

Definition 3.11 The DIC is a goodness of fit test, a generalization of the Akaike information criterion. It is defined as follows:

$$DIC = p_D + \bar{D}$$

where $\bar{D}$ is the expected deviance $D(\theta) = -2\log(f(X|\theta)) + C$. C is a constant, which cancels out when comparing two models and $f(X|\theta)$ is the likelihood function. $\bar{D}$ favours a good fit; more complicated models are penalized by the effective sample size $p_D = \bar{D} - D(\bar{\theta})$, where $\bar{\theta}$ is the expectation of θ .

Since the fixed effects model is nested within the random effects model, a decision can be taken according to best fit. In addition, the residual deviance $\bar{D}$ on its own can be used as an absolute measure of fit (Dias et al. (2011)). Software packages such as WinBUGs include functions to calculate the DIC.

The estimate for between trial variance σ_δ^2 also provides a helpful measure.

In cases where trials are homogeneous, random and fixed effects models provide similar estimates, since the between trial variance parameter is very small.

Meta-Regression

Meta-regression extends the above models to relate the observed effect to a characteristic observed in the involved trials (Thompson and Higgins (2002)). For example, in placebo controlled trials high disease severity at baseline may yield a higher treatment effect than lower disease severity. Meta-regression aims to explain the heterogeneity caused by such effects. A meta-regression model for a random effects MA was developed by Thompson

and Sharp (1999). The model in 3.7 changes to:

$$\begin{aligned}
\Delta_c[i] &\sim N(\mu_c[i], \tau_c[i]), & \Delta_t[i] &\sim N(\mu_t[i], \tau_t[i]) \\
\mu_c[i] &= \nu[i] \\
\mu_t[i] &= \nu[i] + \delta[i] + \beta(x[i] - \bar{x}) \\
\delta[i] &\sim N(d, \sigma_\delta^2)
\end{aligned} \tag{3.11}$$

where $x[i]$ is the value of the covariate in study i , $\bar{x}$ its mean value. β represents the change in effect size per unit increase in x and d now represents the effect size estimated at $\bar{x}$.

Meta-regression results need to be interpreted with care. The relationship described is an observational association across the trials, even when the underlying trials are randomized. Another problem is caused when the relationship is based on average values, which is often the case when dealing with published trials. A small number of studies does not allow for robust estimations. Therefore the meta-regression results should be taken as explanatory rather than used to actually estimate any relations between treatment effect and covariates.

Outcome measures

Measuring the efficacy of an intervention is not straight forward. For many medical conditions there are disease specific measures of disease activity and even for one condition there often exist a range of different measures. Improvement is typically measured as the difference in disease activity at two timepoints. Improvement can be either measured on a continuous scale or using a binary outcome. For continuous measures of improvement the outcome of a meta analysis is the difference in improvement:

$$\Delta = \Delta_A - \Delta_B$$

Binary measures categorize patients as either responders, when their improvement is larger than a given value, or as non-responders, when their improvement is less than this given value. For binary data there are three common outcome measures of a meta-analysis: odds ratios (OR), risk ratios (RR) and risk difference (RD).

$$\begin{aligned}
OR &= \frac{p_A}{1-p_A} / \frac{p_B}{1-p_B} \\
RR &= p_A / p_B \\
RD &= p_A - p_B
\end{aligned} \tag{3.12}$$

Issues related to different summary measures for binary outcomes are highlighted by Deeks (2002). MA models as well as MTC models, which are described in the following section, can be fitted for all of these outcome measures.

Publication Bias

Often, clinical trials which report a conclusive result are more likely to be accepted for publication than articles with negative or non-conclusive results. This yields a bias in the overall published literature, which is referred to as publication bias. The existence of publication bias has been confirmed in Dickerson et al. (1986) and Easterbrook et al. (1991). In an attempt to decrease this problem, prospective registration in a public database of randomized controlled trial was introduced.

It is important to assess the effect of publication bias on the results and conclusions of meta-analyses. The simplest method to detect publication bias is by looking at a funnel plot (Sutton, Abrams and Jones (2000)). Figure 3.5 shows a typical funnel plot of 35 simulated studies and the same data with 5 missing studies illustrating the effect of publication bias. The trim and fill method introduced by Duval and Tweedie (2000)

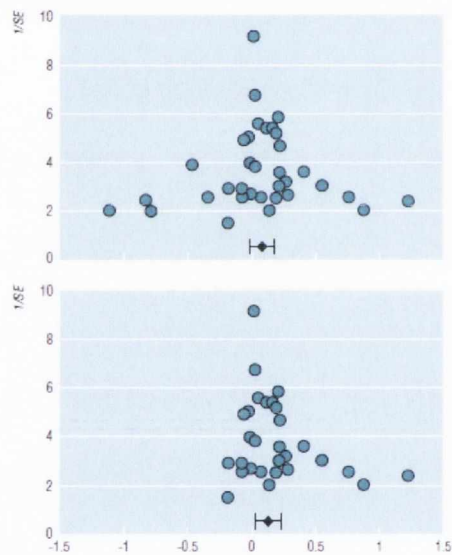


Fig. 3.5: Funnel plot for 35 simulated studies (top) and same data with five missing studies (bottom). (Taken from Sutton, Abrams and Jones (2000).)

allows the estimation of the effect size of publication bias by estimating number and effect of missing studies.

Unfortunately the above methods do not work well when small number of studies are combined. In such situations data bases must be searched for unpublished trials.

3.3.2 Mixed Treatment Comparisons

For many diseases a range of treatments is available. Estimates of their relative efficacy are of interest to inform clinical practice as well as economic evaluations. Two treatments

A and B are ideally compared using direct evidence in the form of RCTs. However, due to financial or ethical reasons such trials might not be available. For the approval of a new intervention it is often sufficient to conduct a trial with placebo or standard of care as comparator rather than using existing alternatives, therefore drug companies usually have no interest in conducting head to head trials for a direct comparison to rivaling treatments. An ethical issue might arise when an existing drug is known to be effective making it unethical to conduct a trial for a new treatment where the control arm is given placebo. In these cases the new treatment is compared to the existing treatment and no placebo controlled trials are available. In situations where direct evidence is not available or insufficient one therefore relies on indirect evidence to estimate the relative treatment effects of interest.

The structure of available evidence can be described using diagrams.

Definition 3.12 *An evidence network diagram illustrates the structure of available evidence. The different interventions are represented by nodes, solid lines between nodes represent direct evidence and a dotted line refers to indirect comparisons of interest. For simplicity, these dotted lines are often omitted in complex structures. For additional information the solid lines can be labeled with the size and number of trials involved in this direct comparison. Figure 3.6 (b) shows a simple example of an evidence network.*

A mixed treatment comparison (MTC) is a generalization of a meta-analysis to include more than two treatments. Figure 3.6 shows the extension of the evidence structure of a meta-analysis to the simplest case of an indirect comparison. In a meta-analysis a range

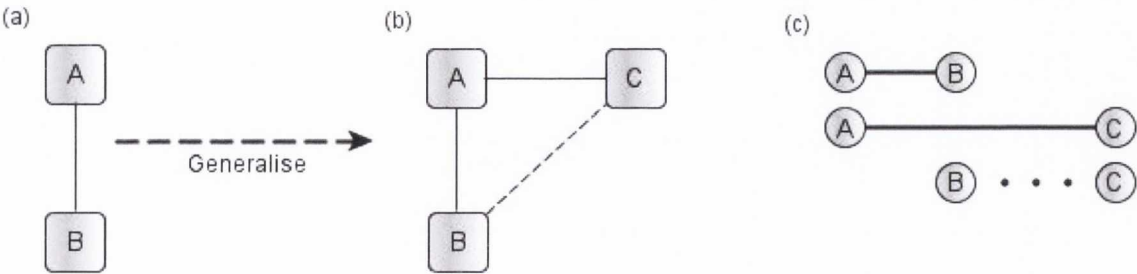


Fig. 3.6: Indirect comparison: Simplest case

of trials comparing the same two treatments A and B are combined to an overall (A vs. B) estimate, see figure 3.6 (a). The simplest case of an indirect comparison extends this network to a third treatment C. Evidence now consists of trials comparing drug A and drug B and also a set of trials comparing drug A and drug C, figure 3.6 (b). Due to the

common link to drug A it is possible to infer something about B versus C; it is defined to be the difference between them, as illustrated in figure 3.6 (c).

While direct evidence is preferred, an indirect comparison can be conducted when none or not sufficient direct evidence is available. According to Lu and Ades (2004) there is nothing improper about the careful use of statistical methods for indirect comparisons, but there are concerns that the indirect estimate might be more biased and have a higher variance than a direct estimate. It is important, therefore, to be aware of the assumptions made and the issues concerning a MTC.

Since MTC modelling is an extension of meta-analysis, the same assumptions are required and the same concerns apply. In addition, MTC modelling requires the assumption of treatment exchangeability.

Definition 3.13 *Treatment exchangeability is a similarity assumption between the trials included for the analysis. For the simplest case in figure 3.6 this means that the effect of drug B in the A vs. B trials would have been the same if one replaced drug C by drug B in the A vs. C trials.*

When evidence networks are of closed loop design, meaning the existence of a closed circle of direct evidence within the network, as can be found in figure 3.8, it is possible to determine the level of consistency among the trials (Lu and Ades (2006); Dias et al. (2010)). Dias et al. propose a node-splitting method which allows the comparison of estimation based on direct evidence and estimation based on indirect evidence. When analysing networks without closed loops, the analyst relies on clinical expertise when selecting the trials to ensure the assumption is fulfilled.

While indirect evidence provides further knowledge about the comparison of interest, the variance of an indirect estimate is typically larger than the variance of a direct comparison of similar size. For the simplest case, the variance of the indirect estimate is defined as the sum of the variances of the two direct estimates informing the comparison.

$$\tilde{Var}(Bvs.C) = \hat{Var}(Avs.B) + \hat{Var}(Bvs.C)$$

It is therefore particularly important to optimize the use of available evidence. Different methodologies have been developed to conduct a mixed treatment comparison, Wells et al. (2009) have published a summary of these methods.

Bucher method

The Bucher method is the first adjusted approach to mixed treatment comparison introduced by Bucher et al. (1997). Adjusted methods use effects relative to a comparator

rather than absolute efficacy and hence maintain the controlled environment of controlled trials. The Bucher method has been developed to calculate odds ratios as outcome measures for the simplest case scenario in figure 3.6 (b). Two meta analyses are conducted separately for the two sets of trials yielding the overall effect measures $\log(OR_{Avs.B})$ and $\log(OR_{Avs.C})$. The indirect comparison estimate is defined as:

$$\log(OR_{Bvs.C}) = -\log(OR_{Avs.B}) + \log(OR_{Avs.C}) \quad (3.13)$$

This way the method can only deal with the simplest case structure, however, it can easily be extended to more complicated networks, such as the star and ladder designs in figure 3.7 using pair wise contrasts. Models using different outcome measures such as risk ratios or risk differences also exist.

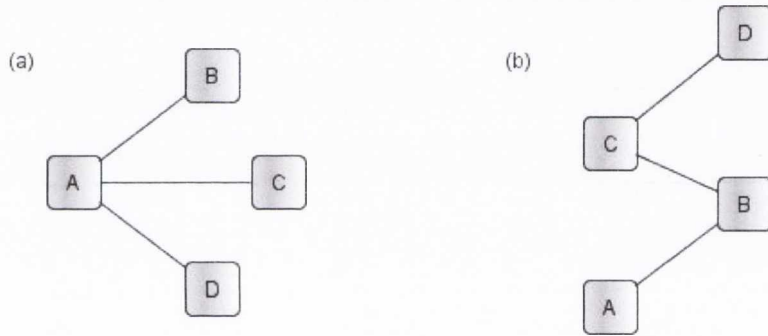


Fig. 3.7: Alternative evidence structures: (a) Star and (b) Ladder design

Lumley's network analysis

When more than one common link is available between two treatments, Lumley's network analysis is applicable. The simplest case of a closed loop evidence structure can be found in figure 3.8. This method is an extension of the Bucher method and allows the determination of a level of consistency of the network. In figure 3.8 treatment A and D are two common links for an indirect comparison of treatment B and C. The procedure is as before, but two estimates: $\log(OR_{Bvs.C})$ can be calculated using each of the common links. These allow to determine the amount of agreement within the two estimates and hence indicate the reliability of the result. In his paper Lumley (2002) gives a quantification of the incoherence of a network. Lumley defines incoherence to be the disagreement between indirect estimates, which cannot be explained by measurement error and between trial heterogeneity. In addition to the random effects allowing for between

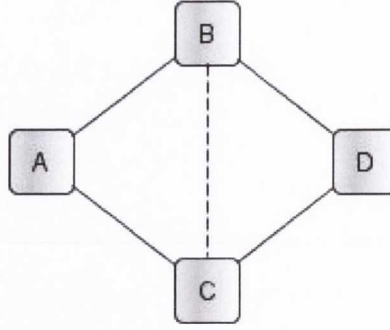


Fig. 3.8: Evidence structure of closed loop design: Example for Lumley’s method

trial heterogeneity, he introduces an additional random effects parameter measuring the change of effect based on different comparator links. In the absence of heterogeneity this parameter is expected to be close to zero; its variance is a measure of consistency.

Bayesian MTC

Bayesian hierarchical models provide a powerful tool to conduct a mixed treatment comparison. First introduced by Lu and Ades (2004), it is the most flexible approach to MTC being able to deal with all types of connected evidence structures. Indirect comparisons are performed simultaneously where no direct evidence is available and borrowing strength across the network ensures an optimal use of the data, Caldwell et al. (2005). Jansen et al. (2008) provides a good introduction to MTC modelling.

The underlying mathematical model is an extension of the meta analysis model described in 3.7.

$$\begin{aligned}
 \Delta_c[i] &\sim N(\mu_c[i], \tau_c[i]), & \Delta_t[i] &\sim N(\mu_t[i], \tau_t[i]) \\
 \mu_c[i] &= \nu[i] \\
 \mu_t[i] &= \nu[i] + \delta[i] \\
 \delta[i] &\sim N(d[t[i]], \tau_\delta) \\
 IC[k, l] &= d[k] - d[l] \quad k \neq l \\
 d[k] &\sim [-, -] \quad \tau_\delta \sim [-, -] \quad \nu[i] \sim [-, -]
 \end{aligned} \tag{3.14}$$

As before $\Delta_c[i]$ and $\Delta_t[i]$ refer to the observed treatment effects in the control and treatment group of study i with respective precision $\tau_c[i]$ and $\tau_t[i]$. $\delta[i]$ is modelled to be the relative treatment effect in study i . Assuming random effects, the study specific true effects are drawn from a normal distribution with a mean specific to the treatment $d[k]$; $t[i]$ refers to the treatment given in study i . The between trial precision parameter τ_δ is assumed to be the same for all comparisons. Each comparison between treatment k and

treatment l , $IC[k, l]$, can be written in terms of basic parameters $d[k]$. Prior distributions are required for the study specific baseline effects $\nu[i]$, the basic parameters and between trial precision τ_δ .

The above model assumes a continuous measure of treatment efficacy, the corresponding model for binary measures has the following form:

$$\begin{aligned}
r_c[i] &\sim \text{Bin}(n_c[i], p_c[i]), & r_t[i] &\sim \text{Bin}(n_t[i], p_t[i]) \\
\text{logit}[p_c[i]] &= \nu[i] \\
\text{logit}[p_t[i]] &= \nu[i] + \delta[i] \\
\delta[i] &\sim N(d[t[i]], \tau_\delta) \\
IC[k, l] &= d[k] - d[l] \quad k \neq l \\
d[k] &\sim [-, -] \quad \tau_\delta \sim [-, -] \quad \nu[i] \sim [-, -]
\end{aligned} \tag{3.15}$$

In recent years Bayesian MTC models have developed to be a widely used tool in medical applications, examples include Cooper et al. (2006); Welton et al. (2009). Its value has been recently highlighted (Cooper et al. (2011)).

Meta-regression

The methods of Bayesian MTC modelling have been extended to allow for meta-regression (Nixon et al. (2007a)). In an attempt to explain heterogeneity among trial populations the methods introduced earlier for simple meta-analysis have been generalized to MTC modelling. This yields the following model, a combination of 3.11 and 3.14:

$$\begin{aligned}
\Delta_c[i] &\sim N(\mu_c[i], \tau_c[i]), & \Delta_t[i] &\sim N(\mu_t[i], \tau_t[i]) \\
\mu_c[i] &= \nu[i] \\
\mu_t[i] &= \nu[i] + \delta[i] + \beta(x[i] - \bar{x}) \\
\delta[i] &\sim N(d[t[i]], \tau_\delta) \\
IC[k, l] &= d[k] - d[l] \quad k \neq l \\
d[k] &\sim [-, -] \quad \beta \sim [-, -] \quad \tau_\delta \sim [-, -] \quad \nu[i] \sim [-, -]
\end{aligned} \tag{3.16}$$

Results from meta-regression in MTCs need to be interpreted with even more care than when applied to simple MA. In addition to the methods discussed earlier, the issue of overfitting needs to be addressed. Overfitting occurs when the distributions of covariates do not overlap across the treatments (Kreif et al. (2012)). In the simplest case, for instance, if the covariates take generally higher values in the A vs. B trials than in the A vs. C trials, a meta-regression will confound the treatment effect by overfitting the regression to explain some of the effect due to the treatment rather than due to the covariate. Plotting the covariate distribution can help to assess this issue. When

one common comparator exists in all trials, an alternative is to plot the observed effect against the mean value of the covariates for each trial. Trials analysing the same drugs should not be clustered.

Meta-regression provides a useful tool to explain heterogeneity due to known covariates. However, estimates should be taken as explanatory rather than for inference.

Measuring Inconsistency

The assumption of treatment exchangeability is difficult to test. All methods to estimate inconsistency in a network rely on the comparison of different paths between two treatments. Inconsistency occurs when different paths yield different efficacy estimates. Therefore, only evidence of closed loop design allows for the estimation of inconsistency. When the evidence network has no closed loops, one has to rely on clinical expertise to select the trials for the analysis. Methods have been developed to quantify inconsistency in networks of closed loop design.

Lu and Ades (2006) adopt Lumley's concept of incoherence, but distinguish between statistical heterogeneity between trials estimating the same comparison and inconsistency occurring between groups of trials estimating different comparisons. Inconsistency degrees of freedom (ICDF) are defined by the number of independent evidence cycles in the network. Models making a consistency assumption, such as defined in line 3 in equation 3.14, are compared to models relaxing this assumption by introducing inconsistency factors (ICF) to this relation. Each ICF is associated with a loop in the network and its size is an estimate of the inconsistency in the loop. A similar approach of comparing models with and without inconsistency factors is taken by Salanti et al. (2008). The node-splitting method by Dias et al. (2010) separates the evidence based on direct evidence and indirect evidence for network edges with potential inconsistency within one model. Posterior distributions are obtained for the treatment effect based on either form of evidence, which allows for a direct comparison. An increase in model fit when splitting a node indicates inconsistency. As before, the ICDF determine the maximum number of inconsistencies in a network.

3.4 Decision Theory

Making decisions is part of every day life; some decisions are simple, like choosing what to have for lunch, others are more complex, like deciding whether or not to buy a house. While a lot of decisions are made intuitively, some require more thought. In very complex cases it helps to formalise the process of decision making in order to provide guidance

through the net of possibilities and uncertainties.

Healthcare is one of the areas where a lot of complex decisions have to be made which often have a large impact and affect the whole population. Since budget is limited, deciding to spend money on one treatment effectively means taking it away somewhere else. In the current economic climate in Ireland this issue recently came to public attention (Lynch (2012)). An informed decision making approach is therefore important to ensure best possible care for the population. This section provides an introduction to the theory of decision making, for further reading see DeGroot (1970); Lindley (1985); French et al. (2009).

One is faced with a decision problem when a choice among alternative actions has to be made yielding uncertain consequences.

The first task is to identify all possible actions and all uncertain events related to these actions. Both lists have to be exhaustive and exclusive. One cannot exhaust the list of possible decisions by simply adding the negation of one decision, since this formulation is generally too vague. For example, when deciding whether to go to the cinema, the negation of not going naturally implies doing something else, like going for a dinner or having a relaxed evening at home. It might be impossible to consider all possibilities, but one should create a reasonable exhaustive list, which covers all relevant possibilities. The list has to be exclusive since the aim is to choose one element of the list as the decision. For example, consider the decision of what to wear for the day. A list of all clothes is not exclusive, since several pieces have to be chosen to create a complete outfit. The list of all possible outfits that can be created from your clothes is an exclusive list. Define the list of all possible actions as

$$d = (d_1, d_2, \dots, d_m).$$

Uncertain events are the other major part in decision making. If there was no uncertainty it would be possible to identify the best choice of action. For example, if we knew it was going to rain, there would be no debate on whether to take an umbrella. The list of uncertain events

$$\theta = (\theta_1, \theta_2, \dots, \theta_n)$$

has to be exclusive and exhaustive to ensure that one and only one event of the list will occur.

The next step is to attach a probability to each of these uncertain events describing how likely it is for each event to occur.

$$p(\theta) = p(\theta_1), p(\theta_2), \dots, p(\theta_n).$$

Table 3.1: Decision Table for the number of possible actions $m = 3$ and the number of uncertain events $n = 4$.

	θ_1	θ_2	θ_3	θ_4
d_1	$u(C_{11})$	$u(C_{12})$	$u(C_{13})$	$u(C_{14})$
d_2	$u(C_{21})$	$u(C_{22})$	$u(C_{23})$	$u(C_{24})$
d_3	$u(C_{31})$	$u(C_{32})$	$u(C_{33})$	$u(C_{34})$
	$p(\theta_1)$	$p(\theta_2)$	$p(\theta_3)$	$p(\theta_4)$

Once decision d_i is taken and θ_j has occurred all uncertainty is removed and the decision will produce a definite result, a foreseeable consequence, namely:

$$C_{ij}$$

The term utility is introduced as a means of describing preferences between consequences. Each consequence is associated with a utility

$$u(C_{ij}).$$

Based on this information it is possible to build a decision table to summarise all available knowledge. A decision table for the number of possible actions $m = 3$ and the number of uncertain events $n = 4$ is displayed in table 3.1.

The last step before deciding on one action is to combine probabilities for events with utilities of consequences, which is done in the expected utility of decision d_i .

$$\overline{u}(d_i) := \sum_{j=1}^n u(C_{ij})p(\theta_j) \tag{3.17}$$

Now an informed decision can be made by choosing the action with the highest expected utility. Figure 1.1 in chapter 1 shows how this utility function feeds into the reimbursement decision making process.

In cases where it is known that event θ_k will occur one simply has to choose action i , which yields the highest utility given the occurrence of θ_k :

$$\max_i \quad u(C_{ik})$$

Decision Tree Analysis

An alternative to a decision table is provided by a decision tree, a clear and straightforward concept to break down a very complex decision into small decisions. Such a decision tree exists for all decision problems and is particularly suitable for problems that can be

divided into sequential problems.

A decision tree grows in chronological order from left to right; it consists of decision nodes from where several branches representing different decisions origin, and random nodes, the starting points for branches representing the different uncertain events that might occur. Decision and random nodes alternate starting with a decision node and finishing with a random node. In order to distinguish between them, a circle refers to a random node and a square to a decision node.

To each branch representing an uncertain event a probability has to be assigned. Since time passes from left to right it is important to consider that at any node information left of that node is available and has to be included into the probability.

The utilities from the decision table can be assigned to the terminal branches.

Now the decision tree can be analysed, which is done starting off at the terminal branches. This makes sense since one could not take a decision in the beginning of the tree without considering the consequences that are involved. For each random node, working from right to left, the expected utility is calculated using the assigned probabilities of the following branches and the utilities assigned to the end of these branches. At each decision node one can now choose the decision resulting in the higher expected utility, alternative decision branches are discarded. Once the base of the tree is reached the best decisions and their expected utilities are determined.

A decision tree for a simple decision problem can be found in figure 3.9.

3.4.1 Bayesian Decision Theory

The decision analysis process is a dynamic process, graphically represented in figure 3.10. A formal decision problem can be evaluated and a recommendation as to which decision is expected to yield the highest utility can be made applying the above process. However, as new information becomes available, perception about probabilities may change and the decision process needs to be refined in light of the additional evidence.

Decision theory therefore fits well within the Bayesian framework, where probability distributions can be updated as more information becomes available. Generally it is not practical to remove uncertainty from the problem, however, one might reduce the uncertainty by obtaining relevant information X . Using Bayes' law the probabilities $p(\theta)$ can be updated as information X becomes available:

$$p(\theta|X) \propto p(\theta) \cdot p(X|\theta)$$

Not only probabilities for events, but also utilities of consequences may change over time. Houlding and Coolen (2011) discuss ideas of adaptive utility in a Bayesian decision mak-

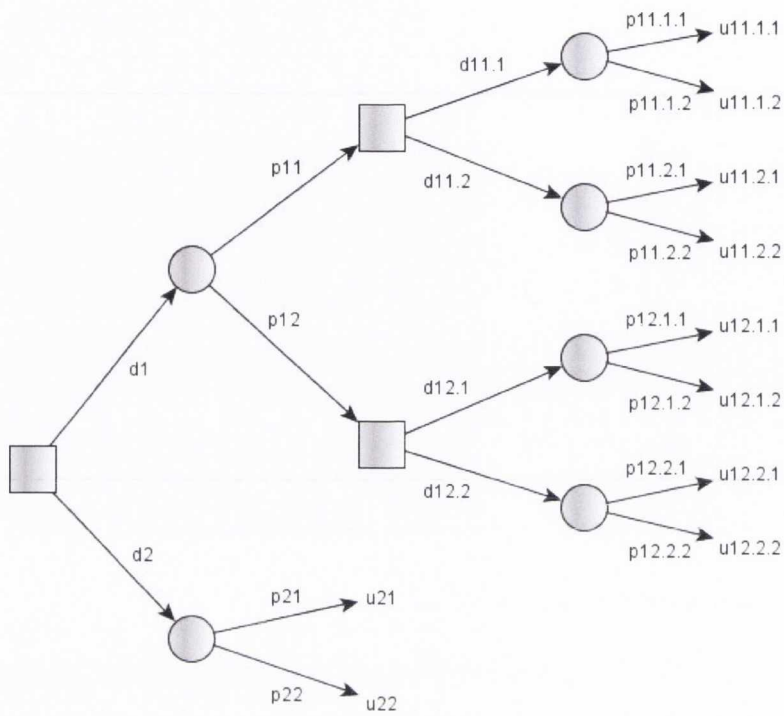


Fig. 3.9: Decision Tree

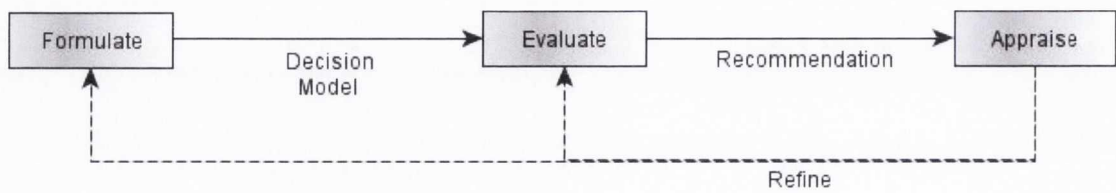


Fig. 3.10: The dynamic nature of a decision analysis process.

ing framework.

Value of information methods described in the previous chapter can be used to evaluate the expected gain from collecting additional data.

3.4.2 Multiple Criteria Decision Analysis

Many complex decision problems are driven by multiple, typically conflicting, objectives. French (2009) for example discusses a multi-attribute approach for the event of a nuclear accident, where each action is associated with conflicting objectives such as costs, public acceptability, stress, cancer-cases etc., which need to be jointly assessed to determine the overall consequences and utility of each action. In chapter 2 healthcare reimbursement decisions have also been identified as depending on multiple conflicting objectives. Costs and budget impact need to be minimised; clinical benefits maximised, while adverse events should be minimised, etc. Multiple Criteria Decision Analysis (MCDA) has been developed to aid decision making in such situations. In this section MCDA literature and current methods for applications in HTA are reviewed. For a general introduction see French et al. (2009).

When p criteria are taken into account the consequence C_{ij} associated with action i and event j is no longer a unidimensional, but becomes a p dimensional vector

$$C_{ij} = (C_{ij1}, \dots, C_{ijp}),$$

where each C_{ijk} represents the achievement against one of the attributes which are considered.

Multi Attribute value modelling

The utility term introduced earlier in this section is generalized to a p -dimensional space to represent the preference of consequences.

$$u(C_{ij}) = u(C_{ij1}, \dots, C_{ijp}) : R^n \rightarrow R$$

The utility function is additive if it can be written as:

$$u(C_{ij}) = u(C_{ij1}, \dots, C_{ijp}) = \sum_{k=1}^p u_k(C_{ijk})$$

where $u_k(C_{ijk})$ are the marginal utility functions. In order for the utility function to be additive the attributes need to be mutually preferentially independent.

Definition 3.14 A subset of attributes is called preferentially independent of the remaining attributes, if the preference between any pair of attributes from this subset does not depend on the levels of achievements of the attributes outside the subset. All attributes are said to be mutually preferentially independent, if this is true for all subsets.

Preferential independence is an important concept; if attributes are not mutually preferentially independent, it is not possible to define preferences for different levels of achievements of one attribute independently of another. It is usually possible to choose attributes in a way that mutual preferential independence holds.

Definition 3.15 γ_{kl} is called a constant relative tradeoff between C_k and C_l if an increase of γ_{kl} in C_k compensates a decrease of 1 in C_l .

An additive utility function is linear if it can be written as:

$$u(C_{ij}) = (C_{ij1}, \dots, C_{ijp}) = \sum_{k=1}^p w_k \cdot C_{ijk}$$

The w_k are called weights and the constant relative tradeoff between attributes is defined as $\gamma_{kl} = \frac{w_k}{w_l}$. The INB introduced in chapter 2 is an example of an additive utility function with 2 attributes (costs and QALYs), where threshold K is the relative tradeoff between costs and QALYs. For visual effects it may be useful to plot curves or surfaces of indifference, which indicate combinations of attribute scorings where one is indifferent about the choice of action, e.g. the ICER on the cost-effectiveness plane in case of the INB.

To utilize multi attribute value modelling for decision making the alternatives need to be assessed against all criteria. In order to justify adding up the criteria the scores need to be normalized to the same scale; typically to values from 0 to 1 or from 0 to 100 are chosen, where the alternative with the worst outcome scores 0 and the alternative with the highest outcome scores 1 or 100. The weights need to take into account the relative importance of the criteria, but also the different scores within each criteria, when the lowest and highest scoring alternative are given 0 and 1 or 100 respectively. This may result in a low weight for an important criteria, when all alternatives score similarly on this criteria. Assessing the weights is not a trivial task, it is hence important to perform a sensitivity analysis on the relative weights between criteria. In the previous chapter the CEAC was an indication for the sensitivity to changes in the relative weight between costs and QALYs.

As before, the alternative with the highest expected utility is the preferred action.

$$\max_i \bar{u}(d_i) := \max_i \sum_{j=i}^n u(C_{ij})p(\theta_j) = \sum_{j=i}^n \sum_{k=1}^p w_k \cdot C_{ijk}p(\theta_j)$$

Outranking

The outranking approach is based on pairwise comparisons of alternative actions. Comparing two actions d_h and d_i can result in either one action outranking the other, in which case this action is the preferred decision to take, they may be indifferent or incomparable. Action d_h is said to outrank action d_i if there is sufficient evidence to justify that action d_h is at least as good as action d_i . Clearly this is the case, if

$$u_k(C_{hjk}) \geq u_k(C_{ijk}) \quad \forall k = 1, \dots, p.$$

However, typically one action does not score better than another action on all criteria. Therefore a more detailed definition of outranking is needed.

Definition 3.16 *Concordance* of d_h is the evidence in favour for d_h outranking d_i . It is defined as

$$CC(d_h, d_i) = \frac{\sum_{k \in Q(d_h, d_i)} w_k}{\sum_{k=1}^p w_k}$$

where $Q(d_h, d_i)$ is the set of criteria where d_h performs at least as good as d_i ; the w_i are the relative weights given to the criteria.

Definition 3.17 *Discordance* of d_h is the evidence against d_h outranking d_i . There is said to be discordance of d_h outranking d_i if

$$\exists k : u_k(d_i) - u_k(d_h) > t_k.$$

t_k represents a level of minimal difference in attribute k which is deemed as no indifference; it is called the discordance threshold.

Using the definitions of concordance and discordance it is possible to define the outranking relation between two alternative actions d_h and d_i .

Definition 3.18 Action d_h is said to outrank action d_i if there is no discordance and

$$CC(d_h, d_i) \geq \bar{C}C$$

where $\bar{C}C$ is called the concordance threshold and defines the minimum difference in concordance deemed as relevant.

If d_h outranks d_i and d_i outranks d_h the two actions are said to be indifferent.

If neither choice outranks the other d_h and d_i are said to be incomparable.

Compared to multi attribute value modelling the assignment of weights is easier, since they only need to represent the relative importance of the criteria. Furthermore there is no need to normalize the scales on which the criteria are scored. However, it is difficult to determine values for the discordance and the concordance threshold. For further reading see Figueira et al. (2005).

Goal Programming

An attribute together with a direction, such as “minimize cost”, is an objective. A goal g_k is defined as a level to be achieved for an objective, for instance a maximum cost of €15,000. The goal programming approach is based on satisfying goals. There are two main approaches of goal programming, lexicographic goal programming and weighted goal programming, see Tamiz et al. (1998).

For lexicographic goal programming the objectives need to be sorted according to priority and deviations from the goal are minimised sequentially.

Weighted goal programming assigns weights to unwanted deviations of the goals according to their relative importance. The aim is then to minimize the sum over the weighted deviations:

$$\min \quad \text{DEV} = \sum_{k=1}^p (w_k^+ \delta_k^+ + w_k^- \delta_k^-)$$

such that

$$f_k(x) - \delta_k^+ + \delta_k^- = g_k \quad \text{for } k = 1, \dots, p$$

where δ_k^+ and δ_k^- are the positive and negative deviations from the goal and $f_k(x)$ is the objective.

Using a hypothetical case study, Thokala and Duenas (2012) have illustrated how each of these techniques described above can be used for HTAs in the UK.

NICE are currently implementing the idea of value based pricing (VBP), a pricing strategy which aims at setting the price for an intervention based on the value of the product to the NHS and the patient. This requires the assessment of what comprises value in a medical intervention. Health gain alone in form of the QALY does not capture all aspects of value; a number of other criteria need to be considered also. VBP and MCDA are closely related in that MCDA provides a methodology of assessing decision problems with multiple outcome criteria. Discussions as to what constitutes value are ongoing (Drummond et al. (2013)). For further information on value based pricing see Claxton (2008b); Towse (2010); Orr et al. (2011).

Chapter 4

MTC: Continuous versus Binary Outcome Measures

The work in this chapter has been done in collaboration with Dr. Roisin Adams, Chief II Pharmacist at the NCPE, who has provided information on the pharmaceutical background as well as conducted the systematic literature review and data collection. The clinical results have been published in Schmitz, Adams, Walsh, Barry and FitzGerald (2012) and the methodological aspect is published in “BMC Medical Research Methodology” (Schmitz, Adams and Walsh (2012*b*)).

4.1 Motivation

Dichotomized outcome measures are widespread in medical research. The perceived advantage of simplicity in the interpretation comes at a cost, however (Royston et al. (2006)). The loss of information results in a loss of power to detect relationships, furthermore the type I error rate may be inflated (Austin and Brunner (2004)), and there is a risk of underestimating the variance parameter. These issues have been discussed in the literature relating to a number statistical analyses (Chen et al. (2007); Sauerbrei et al. (2007); Breitling and Brenner (2010); Dawson and Weiss (2012)).

The work in this chapter demonstrates the consequences of dichotomized outcome measures in the context of MTC modelling. It has been pointed out that dichotomization can be used when designing trials to quantify a treatment effect (Senn and Julious (2009); Julious et al. (1997)). Quantifying a treatment treatment effect is the primary aim of many clinical trials aiming for license approval. The anticipated efficacy level is taken as a cut-off point and results allow a straightforward interpretation. Pharmacoeconomic assessments of healthcare interventions are now a formal component of decision mak-

ing in many countries, as described in chapter 2. Evidence syntheses typically rely on published clinical trials to estimate the relative efficacy among alternative agents to inform decision making. This has created an additional use of clinical trials demonstrating treatment efficacy, for which dichotomized measures suffer from disadvantages compared to the underlying continuous effect measure. It has been described in chapter 3 that the variance of an indirect comparison is larger than the variance of a comparable direct comparison; the variance increases with every indirect link in the chain. It is therefore crucial to make the greatest use of available data. The loss of power due to dichotomization adds to the increased variance in indirect comparisons. This is especially problematic for MTCs (vs. pairwise meta-analysis), since they typically have large standard errors due to the indirect nature of the comparison.

Using a case study in rheumatoid arthritis this chapter presents a MTC model on a range of efficacy measures illustrating the increased power of continuous outcome measures in such models compared to using binary measures. For the presentation, the checklist suggested by Spiegelhalter et al. (2004) is followed to structure the analysis into ‘background’, ‘methods’ and ‘results’ and further subcategories.

4.2 Background

4.2.1 The intervention

Rheumatoid arthritis (RA) is a chronic, progressive and disabling auto-immune disease, causing swelling and damaging cartilage and bone around the joints, see figure 4.1. Any joint may be affected but it is commonly the hands, feet and wrists. Figure 4.2 shows a hand affected by RA. According to the National Rheumatoid Arthritis Society (2011) common symptoms are joint swelling, pain, morning stiffness, poor sleep, fatigue and weight loss.

Over the past decade enhanced understanding of the molecular pathogenesis has led to the development of biologic agents that target specific parts of the immune system. These innovative treatments have altered the path and face of RA and outcomes for patients and society. Tumour necrosis factor alpha antagonists (anti-TNF- α) are the first of the biologic treatment groups used in RA. There are currently five anti-TNF agents licensed for RA in Europe; adalimumab, certolizumab, etanercept, golimumab and infliximab. All of these agents have demonstrated considerable efficacy in placebo controlled randomized controlled trials (RCTs) in patients who have had an inadequate response to conventional Disease Modifying Anti-Rheumatic Drugs (DMARDs) such as methotrexate (MTX) or

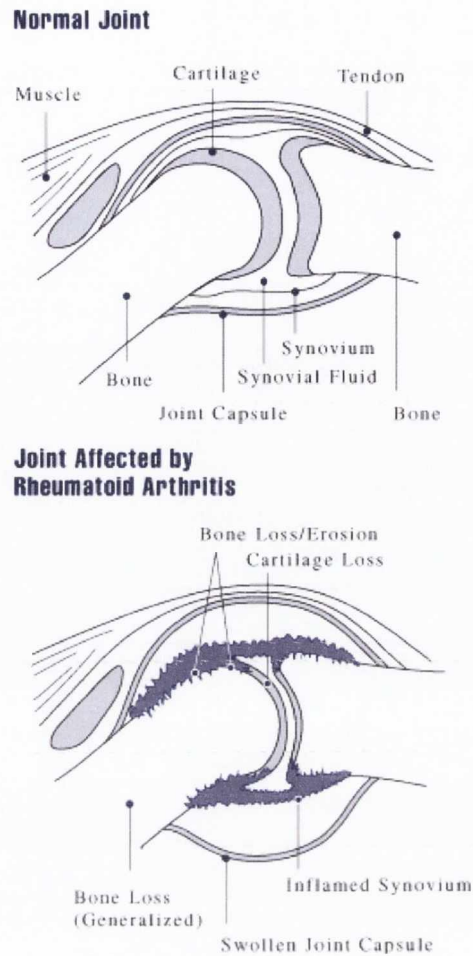


Fig. 4.1: Joint affected by Rheumatoid Arthritis (image taken from <http://www.niams.nih.gov/hi/topics/arthritis/rahandout.htm>)



Fig. 4.2: A hand affected by Rheumatoid Arthritis (image taken from http://en.wikipedia.org/wiki/Rheumatoid_arthritis)

sulphasalazine.

While there is a wealth of RCT evidence available for these agents compared to either placebo or conventional DMARDs, there are currently very limited head-to-head RCTs of anti-TNF agents. Despite this, some estimate of relative efficacy in order to inform choice of agent is needed.

Anti-TNF agents are of high economic impact; in 2009 adalimumab and etanercept represented 30% of the High Tech Drug Reimbursement scheme (€84.3 million), which makes up 12% of the entire drug expenditure.

4.2.2 Aim of Analysis

In the present analysis a MTC model is fitted to estimate all pair-wise comparisons among the five TNF- α inhibitors and placebo, for a range of outcome measures using a Bayesian approach, as has been outlined in chapter 3.

Nixon et al. (2007b) have developed a MTC model to fit the binomial American College of Rheumatology (ACR) outcome measure; it is possible to include trials with multiple treatment arms and to adjust for study level covariates. Jansen et al. (2008) have presented a MTC model on the outcome of the Health Assessment Questionnaire (HAQ), which is measured on a continuous scale. However, they did not allow for multiple treatment arms or the inclusion of baseline characteristics. The present analysis extends Jansen's methods to allow for these.

The aim of this analysis is to demonstrate the advantages of using continuous measures in MTC models compared to using binary measures. While binary measures are useful for demonstrating a certain level of efficacy in clinical trials, the analysis will show that the loss of information when changing from the underlying continuous scale to the binary outcome measure results in a loss of power to detect differences between treatments in MTC analyses. Fitting models for a continuous improvement measure and a discretised version thereof as well as for binary outcome measures and a continuous version of these illustrates the enhanced power to detect differences of continuous measures compared to binary measures.

The chosen outcome measures are based on the ACR criteria and the HAQ score, details are described in the next section.

4.2.3 Efficacy Measures

There are a number of scales on which to measure disease severity and treatment efficacy. These alternative methods can differ in the medical aspects they measure as well as their

Table 4.1: Measuring improvement in RA: ACR (American College of Rheumatology); HAQ (Health Assessment Questionnaire); DAS28 (Disease Activity Score). Note: *HAQ 20 and HAQ 50 are not validated response measures, they have been created here for the purpose of illustrating the loss of power in binary vs. continuous measures.

	ACR	HAQ	DAS28
continuous	meanACR ACRhybrid	HAQ %-improvement	DAS28
binary	ACR 20 ACR 50 ACR 70	HAQ 20* HAQ 50*	

scale.

In order to estimate the relative efficacy between treatments one has to decide on a measure of disease activity and improvement. Commonly used measures in RA are the ACR criteria, the Disease Activity Score (DAS28) and the HAQ score. Table 4.1 summarises the efficacy measures chosen for this analysis. While other measures exist in RA, they are outside the scope of this work and are discussed elsewhere (Saag et al. (2008)).

Definition 4.2.1 *The ACR response criteria is a binary combination measure including the number of tender and swollen joints, patient’s global assessment, physician’s global assessment, pain, degree of disability and level of acute-phase reactant. In order to achieve a ACR 20, ACR 50 or ACR 70 result, an improvement of 20%, 50% or 70% respectively is required in the swollen and tender joint counts as well as in 3 of the 5 remaining dimensions (Felson et al. (1995); Van Riel and van Gelstel (2000)).*

Continuous measures based on the ACR criteria have been introduced including the meanACR and the ACR hybrid measure (Felson and American College of Rheumatology (2007)).

Definition 4.2.2 *The meanACR measures the mean % improvement of the seven ACR core set measures.*

Definition 4.2.3 *The ACR hybrid measure combines the ACR 20, ACR 50 and ACR 70 with the meanACR. A patient’s ACR hybrid outcome is the same as the meanACR, but restricted by his binary ACR response. For example, the outcome for a patient who is an ACR 20 responder, but not an ACR 50 responder is restricted to the interval $[0.2, 0.5)$.*

Definition 4.2.4 The DAS28 score is also a combination measure, but is measured on a continuous scale from 0 to 10. It incorporates swollen joint count (*swollen28*), tender joint count (*tender28*) (both out of a total of 28 defined joints), an evaluation of the patients general health (*GH*) and the erythrocyte sedimentation rate (*ESR*). The score is obtained using the formula

$$DAS28 = 0.56 \cdot \sqrt{tender28} + 0.28 \cdot \sqrt{swollen28} + 0.70 \cdot \ln(ESR) + 0.014 \cdot GH$$

(Prevo et al. (1995)).

Definition 4.2.5 The HAQ score represents the result of a self-report questionnaire in which patients rate their ability to perform daily life activities such as washing one's hair or getting in and out of a car. Values range from 0 to 3 in steps of 0.125, where high values indicate a more severe disease status. The improvement in HAQ score is measured on a continuous scale. The full questionnaire is available at "<http://aramis.stanford.edu/HAQ.html>".

For the purpose of demonstrating the enhanced power of continuous measures in MTC models a discretised version of the HAQ score was defined:

Definition 4.2.6 Binary measures based on the HAQ score, HAQ 20 and HAQ 50, are defined analogously to ACR 20 and ACR 50. A patient achieves a HAQ 20 outcome, if his HAQ score has improved by at least 20%; a 50% improvement is required for HAQ 50.

4.3 Methods

4.3.1 Data

A systematic literature review following the PRISMA method (Moher et al. (2009)), was performed to identify trials meeting the inclusion criteria. The search included published studies up to and including October 2010 in PubMed, Embase and the Cochrane Database. Rheumatological inflammatory diseases other than RA, such as ankylosing spondylitis, psoriatic arthritis and connective tissue diseases were excluded from the search. The inclusion criteria were RCTs, patients with established RA who have had an inadequate response to MTX and who have been treated for at least 24 weeks (where 24 week data were not available, data within 6 weeks either before or after 24 weeks were used). Both monotherapy and combination therapy were included with an explicit term in the statistical model allowing for the additional effect of MTX. The outcome measures chosen were those described in the previous section based on the ACR criteria, the HAQ

score and the DAS28. Unfortunately the DAS28 was reported in too few trials to fit a MTC model (DAS28 was only reported in 6 trials, not representing each of the treatments). The total number of patients, number of respondents achieving ACR response, and the mean improvement and standard deviations (SDs) in the case of the HAQ were extracted. Authors were contacted in cases where the required data were not reported. Where no access to the requested data was provided, the following methodology was applied: in cases where the mean was not reported, the median was used; in the absence of SDs, interquartile ranges (IQRs) were used to estimate SDs using a normal approximation, and, in the remaining cases, the maximum of clinical trial SDs was used. The doses of biological agents included are those included in the RCTs. Demographic data including age, dose of MTX, mean disease duration, baseline HAQ score and previous number of DMARDs were recorded.

4.3.2 Statistical Model

A Bayesian MTC model is fitted to the data for each of the outcome measures of interest (HAQ, HAQ 20, HAQ 50, ACR 20, ACR 50, ACR 70 and ACRcont), where ACRcont refers to the continuous measure based on the ACR response.

The MTC model was introduced in chapter 3; the continuous version in model 3.14 and the binary version in 3.15.

For the use of the model for RA data three adjustments have to be made. The model should allow for more than one treatment arm per study; adjustments have to be made for the concurrent treatment with MTX. Furthermore, the improvement should be modelled relative to baseline disease activity; the reason being that disease activity at baseline influences the effectiveness of the intervention (Hyrich et al. (2006)). Also, the results of this analysis provide the basis for an economic analysis, which requires the percentage improvement in HAQ score as input parameters. A step by step approach is taken to allow for these issues in the above models.

As a first step the allowance for multiple treatment arms is included. Let index j refer to treatment arm j of study i . $t[i,j]$ refers to the treatment given in arm j of study i and $bi[i,j]$ refers to the baseline treatment associated with treatment arm j of study i . The

continuous mathematical model then takes the following form:

$$\begin{aligned}
\Delta[i, j] &\sim N(\mu_{\Delta}[i, j], \tau_{\Delta}[i, j]) \\
\mu_{\Delta}[i, j] &= \nu[i] + \delta[i, t[i, j]] \\
\delta[i, b[i, j]] &= 0 \\
\delta[i, t[i, j]] &\sim N(d[t[i, j]] - d[bi[i, j]], \tau_{\delta}) \\
d[1] &= 0 \\
IC[k, l] &= d[k] - d[l] \quad k \neq l \\
d[k] &\sim [-, -] \quad \tau_{\delta} \sim [-, -] \quad \nu[i] \sim [-, -]
\end{aligned} \tag{4.1}$$

The HAQ improvement in each arm j of study i , $\Delta[i, j]$, has a normal distribution with precision $\tau_{\Delta}[i, j]$, given in the trials. The model makes the assumption that effects in treatment arms of the same study, where the same treatment is given, are constant; i.e. fixed effects are assumed for within trial study effects. Therefore it holds $\delta[i, j] = \delta[i, t[i, j]]$. This line has been omitted in the model for simplicity and is incorporated in line 2 of equation 4.1. None of the trials incorporates more than two different treatments; therefore it is not necessary to allow for between treatment correlation. This will be incorporated in chapter 6. Between trials the model assumes random effects, as before, and each comparison $IC[k, l]$ can be written in terms of basic parameters $d[k]$. $d[1] = 0$ defines treatment 1 to be the basic treatment; all other $d[k]$ describe the effects relative to treatment 1. Prior distributions need to be defined for basic parameters $d[k]$ and between trial precision τ_{δ} and $\nu[i]$.

The binomial model allowing for multiple treatment arms takes the following form:

$$\begin{aligned}
r[i, j] &\sim Bin(n[i, j], p[i, j]) \\
logit(p[i, j]) &= \nu_i + \delta[i, t[i, j]] \\
\delta[i, b[i, j]] &= 0 \\
\delta[i, t[i, j]] &\sim N(d[t[i, j]] - d[bi[i, j]], \tau_{\delta}) \\
d[1] &= 0 \\
IC[k, l] &= d[k] - d[l] \quad k \neq l \\
d[k] &\sim [-, -] \quad \tau_{\delta} \sim [-, -] \quad \nu[i] \sim [-, -]
\end{aligned} \tag{4.2}$$

Changes from equation (3.15) to (4.2) are analogous to changes in the continuous case. A LOR $\delta[i, j]$ is obtained for each treatment arm relative to the baseline arm; treatment arms of the same study, where the same treatment is given, are assumed to estimate the same treatment effect; everything else remains the same.

In a next step the model is extended to allow for the concurrent treatment with MTX. The effect due to MTX is assumed to be additive, meaning the relative effect between

two arms where no MTX is given is the same as the relative effect between two arms where MTX is given in both arms:

$$(A + \text{MTX}) \text{ vs. } (B + \text{MTX}) \equiv A \text{ vs. } B \quad (4.3)$$

The continuous model takes the following form:

$$\begin{aligned} \Delta[i, j] &\sim N(\mu_\Delta[i, j], \tau_\Delta[i, j]) \\ \mu_\Delta[i, j] &= \nu[i] + \alpha[i, t[i, j]] + b \cdot I[i, j] \\ \alpha[i, b[i, j]] &= 0 \\ \alpha[i, t[i, j]] &\sim N(a[t[i, j]] - a[bi[i, j]], \tau_\alpha) \\ a[1] &= 0 \\ IC[k, l] &= a[k] - a[l] \quad k \neq l \\ a[k] &\sim [-, -] \quad b \sim [-, -] \quad \tau_\alpha \sim [-, -] \quad \nu[i] \sim [-, -] \end{aligned} \quad (4.4)$$

The difference to the previous model described in equation (4.1) is the splitting of treatment effect into a part representing the effect due to the drug of interest (in the RA case the anti-TNF effect) $\alpha[i, t[i, j]]$ and a part explaining the effect which is due to the concurrent treatment with MTX, $b \cdot I[i, j]$ indicates whether MTX was given in treatment arm j of study i . By multiplying b with the indicator the assumptions of an additive effect as explained above is implemented. In this way, only treatment arms, where MTX is given in either the treatment arm or in the respective control arm inform parameter b . As previously the effect in arms of the same study is assumed constant, $\alpha[i, j] = \alpha[i, t[i, j]]$ (this line has been omitted in the equation for simplicity; the effect is incorporated in line 2 of equation 4.4.). Between trial random effects are assumed for α and fixed effects for the MTX effect. One could just as well assume random effects for b , but the RA data does not provide sufficient evidence to inform a between trial variability for b . All comparisons $IC[k, l]$ can be expressed in terms of basic parameters $a[k]$. Prior distributions need to be specified for $a[k]$, for b , for the between trial standard deviation σ_α and for $\nu[i]$. The binary version of this model is mathematically described as follows:

$$\begin{aligned} r[i, j] &\sim \text{Bin}(n[i, j], p[i, j]) \\ \text{logit}(p[i, j]) &= \nu[i] + \alpha[i, t[i, j]] + b \cdot I[i, j] \\ \alpha[i, b[i, j]] &= 0 \\ \alpha[i, t[i, j]] &\sim N(a[t[i, j]] - a[bi[i, j]], \tau_\alpha) \\ a[1] &= 0 \\ IC[k, l] &= a[k] - a[l] \quad k \neq l \\ a[k] &\sim [-, -] \quad b \sim [-, -] \quad \tau_\alpha \sim [-, -] \quad \nu[i] \sim [-, -] \end{aligned} \quad (4.5)$$

Again, changes follow analogously to changes made in the continuous case.

In a last step the model is extended such that the drug effect $\mu_\Delta[i, j]$ depends on baseline disease activity. To model relative percentage improvement μ_Δ is modelled as a multiplier to the HAQ score at baseline $\lambda[i, j]$. All other aspects of model (4.4) remain unchanged.

$$\begin{aligned}
\Delta[i, j] &\sim N(\mu_\Delta[i, j], \tau_\Delta[i, j]) \\
\mu_\Delta[i, j] &= \lambda[i, j] \cdot \delta[i, j] \\
\delta[i, j] &= \nu[i] + \alpha[i, t[i, j]] + b \cdot I[i, j] \\
\alpha[i, b[i, j]] &= 0 \\
\alpha[i, t[i, j]] &\sim N(a[t[i, j]] - a[b[i, j]], \tau_\alpha) \\
a[1] &= 0 \\
IC[k, l] &= a[k] - a[l] \quad k \neq l \\
a[k] &\sim [-, -] \quad b \sim [-, -] \quad \tau_\alpha \sim [-, -] \quad \nu[i] \sim [-, -]
\end{aligned} \tag{4.6}$$

The binomial measures selected for this study are by definition relative to baseline disease activity, since the cut-off between response and no response is a % improvement relative to baseline. It is therefore not necessary to adjust the model defined in equation 4.5 further.

These final models in equations 4.5 and 4.6 enable the estimation of relative efficacy of a number of treatments while allowing for multiple treatment arms, concurrent treatment with additional medication and dependence of treatment effect on baseline disease activity for both, continuous and binary outcome measures. In the continuous case differences in efficacy are estimated; odds ratios are estimated in the binary case. Using methods by Warn et al. (2002) the model can be modified to calculate other measures, such as risk ratios.

Model 4.5 has been developed by Nixon et al. (2007b) for the binary case.

4.3.3 Prior Distributions

The models described above require prior distributions for the basic treatment effects, for the effect due to MTX, for the between trial precision and for the baseline response. Vague priors have been chosen for all of these. For $d[k]$ in models (4.1) and (4.2), for $a[k]$ and b in (4.4) - (4.6) as well as for the baseline response $\nu[i]$ a normal distribution centred at no treatment effect 0 with a very low precision 0.0001 have been chosen:

$$\begin{aligned}
d[k] &\sim N(0, 0.0001) \\
a[k] &\sim N(0, 0.0001) \\
b &\sim N(0, 0.0001) \\
\nu[i] &\sim N(0, 0.0001)
\end{aligned} \tag{4.7}$$

The same prior can be used for both continuous and binary models, since the improvement difference and the log odds ratios take values on the real line where 0 represents no difference. The priors do not favour any of the drugs and allows for a wide range of likely values in favour for each. It is a widely used vague prior on treatment effects.

Rather than on the between trial precision parameter directly, the prior was defined on between trial standard deviation $\sigma_{\delta/\alpha} = \sqrt{1/\tau_{\delta/\alpha}}$. A uniform distribution was chosen, as proposed by Gelman (2006).

$$\sigma_{\delta/\alpha} \sim \text{dunif}(0, 2) \quad (4.8)$$

To test whether the range of the uniform prior was chosen appropriately the analysis has been rerun using wider ranges ($\text{dunif}(0, 5)$ and $\text{dunif}(0, 7)$) and estimates did not change. Spiegelhalter et al. (2004) suggest a half normal distribution as an alternative choice for the prior distribution of the heterogeneity parameter. The choice of a half normal distribution ($\text{dnorm}(0, 4)I(0,)$) did not change estimates noticeably.

4.3.4 Computation / Software

The MTC models were fitted in WinBUGs, a MCMC software using Gibbs sampling (Lunn et al. (2000)). The complete code and input data is provided in appendix A. Computational feasibility allowed for a large number of iterations. Each model discarded 50,000 burnin iterations and was run with 100,000 iterations and two chains. Convergence was analysed in R using the CODA package (Plummer et al. (2006)); the effective sample size was checked, and visual inspection of the autocorrelation and the chains confirmed convergence.

4.4 Results

4.4.1 Data

The systematic literature review identified sixteen RCTs meeting the inclusion criteria, figure 4.3 shows the flow diagram of the selection process. Five trials met the inclusion criteria for adalimumab, four for infliximab, two for etanercept, two for golimumab and three for certolizumab. MTX was included in 11 of the studies; no MTX was given in four of the trials. One trial contained arms of combination and arms of monotherapy. Figure 4.4 shows the network of available evidence. The data is summarized in table 4.2, baseline demographics in table 4.3.

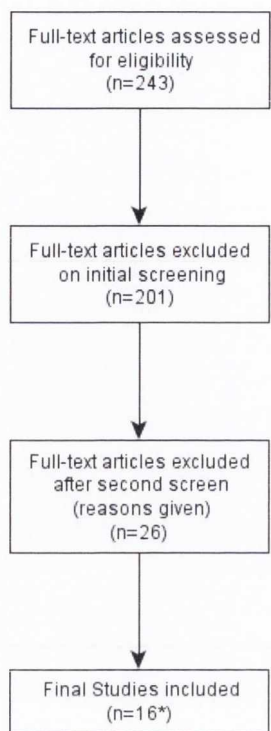


Fig. 4.3: Selection Process: Flow Diagram for anti-TNF therapy in RA.

Table 4.2: Data extracted from RCTs: Number of patients N; improvement in HAQ score Δ HAQ; number of ACR20, ACR50, ACR70 responders; HAQ score at baseline HAQ_b ; + indicates additional treatment with MTX; NA = not available. Ada=adalimumab; Inf=infliximab; Eta=etanercept; Gol=golimumab, Cert=certolizumab, P=placebo. *Median; †estimated from IQR; ‡data provided by authors following request.

Trial	Arm	N	Δ HAQ	ACR20	ACR50	ACR70	HAQ_b
Weinblatt et al. (2003)	P+	62	0.27 (0.6)	9	5	3	1.64
	Ada+	69	0.54 (0.6)	33	22	7	1.52
	Ada+	67	0.62 (0.6)	45	37	18	1.55
	Ada+	73	0.59 (0.5)	48	31	14	1.55
Keystone et al. (2004)	P+	200	0.24 (0.5)	59	19	5	1.45
	Ada+	207	0.56 (0.5)	131	81	43	1.44
	Ada+	212	0.60 (0.5)	129	87	37	1.48
Van De Putte et al. (2004)	P	110	0.07 (0.5)	21	9	2	1.88
	Ada	112	0.39 (0.6)	44	23	11	1.88
	Ada	106	0.29 (0.6)	38	20	9	1.88
	Ada	103	0.49 (0.5)	55	36	19	1.84
	Ada	113	0.38 (0.6)	52	25	14	1.83
Miyasaka (2008)	P	87	-0.1 (0.6)	12	5	1	1.39
	Ada	87	0.2 (0.5)	25	14	9	1.57
	Ada	91	0.2 (0.6)	40	22	11	1.64
	Ada	87	0.4 (0.6)	44	28	13	1.77
Kim et al. (2007)	P+	63	0.2 (0.5)	23	9	5	1.3
	Ada+	65	0.5 (0.6)	40	28	14	1.4

Continued on next page

Table 4.2 – continued from previous page

Trial	Arm	N	Δ HAQ	ACR20	ACR50	ACR70	HAQ_b
Maini et al. (1999)	P+	88	0.3 (0.5) $\ddagger$	18	7	0	1.8
	Inf+	86	0.3 (0.5) $\ddagger$	45	22	7	1.8
	Inf+	86	0.5 (0.5) $\ddagger$	47	25	9	1.8
	Inf+	87	0.5 (0.6) $\ddagger$	51	26	15	1.8
	Inf+	81	0.4 (0.5) $\ddagger$	49	21	9	1.5
Westhovens et al. (2006)	P+	363	NA	87	33	16	1.5
	Inf+	360	NA	199	110	48	1.5
	Inf+	361	NA	205	119	54	1.5
Zhang et al. (2006)	P+	86	0.45 (NA)	42	22	12	1.6
	Inf+	87	0.76 (NA)	66	38	20	1.6
Schiff et al. (2008)	P+	110	NA	49	22	10	1.8
	Inf+	165	NA	98	61	40	1.7
Moreland et al. (1999)	P	80	0.03 (NA)	9	4	1	1.7
	Eta	76	0.58 (NA)	39	18	7	1.7
	Eta	78	0.62 (NA)	46	31	12	1.6
Weinblatt et al. (1999)	P+	30	0.4 (NA)	8	1	0	1.5
	Eta+	59	0.7 (NA)	42	23	9	1.5
Keystone et al. (2009)	P+	133	0.13* (0.4) $\dagger$	37	18	7	1.25*
	Gol	133	0.13* (0.7) $\dagger$	47	26	15	1.38*
	Gol+	89	0.38* (0.5) $\dagger$	53	33	18	1.38*
	Gol+	89	0.5* (0.5) $\dagger$	53	29	16	1.38*
Kay et al. (2008)	P+	35	NA	13	2	0	1.3
	Gol+	35	NA	21	13	3	1.7
	Gol+	34	NA	19	10	6	1.8
Keystone et al. (2008)	P+	199	0.18 (NA)	27	15	6	1.7
	Cert+	393	0.60 (NA)	231	146	84	1.7
	Cert+	390	0.63 (NA)	237	156	79	1.7
Smolen et al. (2009)	P+	127	0.14 (0.5)	11	4	1	1.6
	Cert+	246	0.5 (0.5)	141	80	39	1.6
	Cert+	246	0.5 (0.5)	142	81	26	1.6
Fleischmann et al. (2009)	P	109	-0.07 (0.4) $\ddagger$	10	4	0	1.6
	Cert	111	0.39 (0.7) $\ddagger$	51	25	9	1.4

The MTC models described previously are fit to the data; the continuous model for the improvement in HAQ score and the binary model for the ACR 20, ACR 50 and ACR 70 outcome measures.

The aim of this analysis is to demonstrate the increased power of continuous outcome measures to detect differences in MTC models compared to binary measures. To explore this hypothesis the HAQ score was discretised to a binary outcome measure and the ACR criteria were transformed into a continuous measure, ACRcont; the resulting measures were analysed in addition to the trial reported outcome measures and results compared. The rationale for doing this was that comparing HAQ and ACR outcomes is not comparing like with like; they measure quite different aspects of disease and additional significant findings using the HAQ score could be because of this rather than there

Table 4.3: Mean Baseline Characteristics of trials: Mean Age (Years), mean duration of disease DoD (Years), mean number of previous DMARDS, Dose of MTX (mg). NR=not reported; NU=not used.

Trial	Age (Yrs)	DoD (Yrs)	DMARDS (no.)	Dose MTX (mg)
Weinblatt et al. (2003)	56	12	3	17
Keystone et al. (2004)	56	11	2	17
Van De Putte et al. (2004)	53	11	4	NU
Miyasaka (2008)	55	7	NU	NU
Kim et al. (2007)	49	7	NU	16
Maini et al. (1999)	53	8	3	15
Westhovens et al. (2006)	52	7	NR	15
Zhang et al. (2006)	48	8	NR	NR
Schiff et al. (2008)	49	8	NU	16
Moreland et al. (1999)	52	12	3	NU
Weinblatt et al. (1999)	50	13	3	19
Keystone et al. (2009)	51	6	NU	15
Kay et al. (2008)	53	7	NR	NR
Keystone et al. (2008)	52	6	1	14
Smolen et al. (2009)	52	6	1	13
Fleischmann et al. (2009)	54	10	2	NU

being any effect of a continuous vs. a binary outcome measure. Estimating a discrete version of the HAQ and a continuous measure of the ACR allows for the comparison of binary and continuous outcomes while the measure is kept fixed.

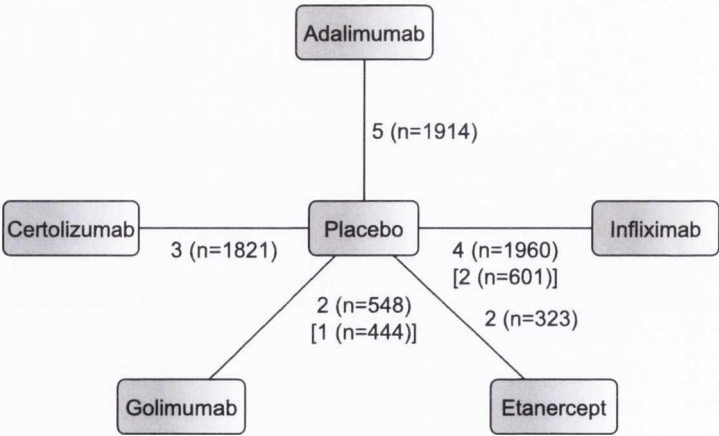


Fig. 4.4: Network Diagram for RA analysis: Edges are labelled with the number of studies and the total number of patients included in these studies. Numbers in square brackets refer to HAQ evidences where this differs from ACR evidence.

Discretised HAQ: HAQ 20 and HAQ 50

For the purpose of comparing continuous and binary outcome measures based on the same data, the HAQ 20 and HAQ 50 measures defined in definition 4.2.6 are estimated based on the continuous HAQ improvement. These discretised HAQ outcomes are not validated outcome measures for RA and are therefore not reported in the trials.

The trials report mean and standard deviation of HAQ improvement. The number of HAQ 20 and HAQ 50 responders in each trial arm was estimated as follows. For each trial arm, the HAQ improvement of 1000 patients was generated by reference to normal distribution with mean and SD given from the data in that group. From these patients the proportion of HAQ 20 and HAQ 50 responders was calculated and applied to the number of patients in each trial arm. Data is summarised in table 4.4.

Table 4.4: Input Data for discretised HAQ score and continuous ACR: Number of patients N; number of HAQ20 and HAQ50 responders; continuous ACR response ACRcont; + indicates additional treatment with MTX; NA=not available.
Ada = adalimumab; Inf = infliximab; Eta = etanercept; Gol = golimumab, Cert = certolizumab, P = placebo.

Trial	Arm	N	HAQ 20	HAQ 50	ACRcont(SD)
Weinblatt et al. (2003)	P+	62	30	12	0.08 (0.22)
	Ada+	69	45	24	0.35 (0.20)
	Ada+	67	47	26	0.46 (0.22)
	Ada+	73	52	26	0.42 (0.21)
Keystone et al. (2004)	P+	200	95	32	0.13 (0.22)
	Ada+	212	141	81	0.42 (0.21)
	Ada+	207	154	82	0.41 (0.21)
Van De Putte et al. (2004)	P	110	29	4	0.09 (0.20)
	Ada	112	47	16	0.31 (0.19)
	Ada	106	60	19	0.30 (0.19)
	Ada	103	57	20	0.38 (0.22)
	Ada	113	60	21	0.33 (0.20)
Miyasaka (2008)	P	87	23	8	0.07 (0.17)
	Ada	87	37	10	0.28 (0.19)
	Ada	91	39	16	0.33 (0.20)
	Ada	87	48	18	0.37 (0.21)
Kim et al. (2007)	P+	63	28	10	0.18 (0.27)
	Ada+	65	43	24	0.42 (0.22)
Maini et al. (1999)	P+	88	40	9	0.09 (0.19)
	Inf+	86	40	12	0.35 (0.19)
	Inf+	86	54	19	0.36 (0.20)
	Inf+	87	52	23	0.38 (0.21)
	Inf+	81	48	20	0.37 (0.19)
Westhovens et al. (2006)	P+	363	NA	NA	0.12 (0.23)
	Inf+	360	NA	NA	0.37 (0.20)
	Inf+	361	NA	NA	0.38 (0.21)
Zhang et al. (2006)	P+	86	52	28	0.27 (0.31)

Continued on next page

Table 4.4 – continued from previous page

Trial	Arm	N	HAQ 20	HAQ 50	ACRcont(SD)
Schiff et al. (2008)	Inf+	87	61	39	0.46 (0.20)
	P+	110	NA	NA	0.22 (0.29)
	Inf+	165	NA	NA	0.41 (0.22)
Moreland et al. (1999)	P	80	25	8	0.06 (0.16)
	Eta	76	47	26	0.34 (0.19)
	Eta	78	53	31	0.40 (0.21)
Weinblatt et al. (1999)	P+	30	16	8	0.10 (0.17)
	Eta+	59	45	29	0.43 (0.19)
	P+	133	49	13	0.14 (0.25)
Keystone et al. (2009)	Gol	133	56	29	0.30 (0.20)
	Gol+	89	51	24	0.40 (0.22)
	Gol+	89	60	31	0.39 (0.21)
Kay et al. (2008)	P+	35	NA	NA	0.14 (0.20)
	Gol+	35	NA	NA	0.39 (0.20)
	Gol+	34	NA	NA	0.38 (0.21)
Keystone et al. (2008)	P+	199	83	35	0.07 (0.20)
	Cert+	393	263	146	0.40 (0.22)
	Cert+	390	249	137	0.41 (0.22)
Smolen et al. (2009)	P+	127	46	10	0.04 (0.14)
	Cert+	246	163	66	0.38 (0.21)
	Cert+	246	155	67	0.38 (0.20)
Fleischmann et al. (2009)	P	109	19	2	0.04 (0.14)
	Cert	111	60	34	0.33 (0.19)

Continuous ACR: ACRcont

While the continuous ACR measures were proposed by the Felson and American College of Rheumatology (2007), they were not assessed in any of the trials. Therefore a continuous ACR measure, ACRcont, is generated based on the ACR 20, ACR 50 and ACR 70 outcomes reported in the trials. This enables the comparison of the outcomes of the binary measures with those resulting from the continuous scale. ACRcont takes a value between 0 and 1 representing the percentage improvement in the dimensions which are combined for the binary ACR criteria.

The trial data allows the categorisation of patients into non-responders (group 1) (patients not achieving an ACR 20 response), patients achieving ACR 20 but not ACR 50 response (group 2), patients achieving ACR 50 but not ACR 70 response (group 3) and ACR 70 responders (group 4). Assuming a mean ACRcont response m_i for each group of patients these numbers are used to generate the mean μ and standard deviation σ of

Table 4.5: Mean improvement estimated from patient level data.

	<i>group 1</i>	<i>group 2</i>	<i>group 3</i>	<i>group 4</i>
meanACR	0.17	0.41	0.57	0.72
ACRhybrid	0.10	0.41	0.59	0.77
Control Group	0.00	0.35	0.60	0.85

ACRcont for each trial arm:

$$\mu(\text{ACRcont}) = \frac{1}{N} \left(\sum_{i=1}^4 N_i \cdot m_i \right) \quad (4.9)$$

where N_i refers to the number of patients in each group and $N = \sum_{i=1}^4 N_i$.

$$\sigma(\text{ACRcont}) = \sqrt{\frac{1}{N} \sum_{i=1}^4 N_i \cdot (m_i - \mu(\text{ACRcont}))^2} \quad (4.10)$$

Patient level data from a RA cohort allowed the estimation of m_i for the four groups (Adams et al. (2010)). Mean ACRcont response m_i can be based on either continuous ACR measure described previously, depending upon which underlying measure one is trying to recreate. For the main analysis the meanACR measure is used; results based on the ACRhybrid are discussed in the sensitivity analysis. Table 4.5 summarizes the mean ACRcont responses for the four groups based on both outcome measures. Improvement in the core sets is measured by a drop in score; therefore overall improvement is restricted by an upper bound of 100%. To achieve a symmetric measuring scale overall worsening has also been restricted to -100% in each of the criteria, as proposed by the Felson and American College of Rheumatology (2007).

The patient level estimates are based on patients receiving biologic treatment; estimates can therefore be applied to the treatment arms of the studies. Unfortunately no patient level data representing the control groups is available. Given the lack of further information the analysis assumes interval midpoints as mean ACRcont estimates for the control groups. Group 1 is assumed to have a response in $[-1, 0.2)$, group 2 in $[0.2, 0.5)$, group 3 in $[0.5, 0.7)$ and the response of group 4 lies in $[0.7, 1]$. The interval for non responders is very broad covering a wide range of extreme values. The baseline analysis therefore assumes a conservative mean value of 0.00 for this group. This yields the control group estimates in table 4.5. Alternative scenarios assuming the same mean response for the control group as estimated for the treatment group are analysed in the sensitivity analysis. The estimated input data for the continuous ACR model based on meanACR is summarised in table 4.4.

4.4.2 Efficacy Results

The models calculate the relative efficacy among all anti-TNF agents and placebo. Relative efficacy for the continuous measures (HAQ improvement and ACRcont) was measured as difference in improvement. Relative improvement was modelled as a multiplier to baseline HAQ score. Odds ratios were calculated for the binary measures ACR 20, ACR 50, ACR 70 and HAQ 20 and HAQ50. The alternative risk ratio scale was analysed in a sensitivity analysis, since different scales for binomial outcome measures can affect the outcome of MTC analyses (Caldwell et al. (2012)). Thirteen trials were included for the HAQ analysis and sixteen for the ACR analysis.

The MTC results are summarised in table 4.6 for all HAQ outcomes and in table 4.7 for ACR outcomes. Mean and 80% credible intervals are reported to summarize the posterior distribution for the relative efficacy between anti-TNF agents as well as the between study standard deviation σ . The 80% level was chosen to shift away from the misunderstood acceptance-rejection dichotomy of random significance levels (Schmitz, Adams, Walsh, Barry and FitzGerald (2012)). Results with a probability of $\geq 90\%$ to favour one drug over the other are marked with a *. A graphical summary can be found in figure 4.5 for the HAQ measures and figure 4.6 for the ACR measures. Binomial results are plotted on the log scale.

Table 4.6: Results from HAQ analysis: Mean estimate with 80% credible intervals for each pair-wise comparison. Results with a probability of $\geq 90\%$ in favour of one drug over the other are marked with a *. The outcome measure is % improvement for the continuous HAQ outcome. In the case of HAQ 20 and HAQ 50 the measures are odds ratios. Ada=adalimumab; Inf=infliximab; Eta=etanercept; Gol=golimumab, Cert=certolizumab, P=placebo, σ =between trial SD

<i>Comparison</i>	<i>HAQ</i>	<i>HAQ20</i>	<i>HAQ50</i>
Ada vs P	0.20 (0.17,0.23)*	2.6 (2.2,3.1)*	3.1 (2.1,4.4)*
Inf vs P	0.10 (0.06,0.15)*	1.6 (1.2,2.1)*	2.0 (1.2,3.5)*
Eta vs P	0.31 (0.25,0.37)*	3.6 (2.6,5.2)*	4.1 (2.2,7.7)*
Gol vs P	0.23 (0.17,0.29)*	2.8 (2.0,4.1)*	4.2 (1.9,8.8)*
Cert vs P	0.25 (0.22,0.29)*	3.3 (2.7,4.0)*	5.0 (3.0,7.6)*
Inf vs Ada	-0.10 (-0.15,-0.05)*	0.6 (0.4,0.8)*	0.7 (0.3,1.3)
Eta vs Ada	0.11 (0.04,0.17)*	1.4 (0.9,2.1)	1.3 (0.7,2.7)
Eta vs Inf	0.21 (0.13,0.28)*	2.3 (1.4,3.6)*	2.0 (0.9,4.9)
Gol vs Ada	0.02 (-0.05,0.08)	1.1 (0.7,1.6)	1.4 (0.6,3.3)
Gol vs Inf	0.12 (0.05,0.20)*	1.8 (1.1,2.8)*	2.1 (0.8,5.4)
Gol vs Eta	-0.09 (-0.17,0.00)*	0.8 (0.5,1.3)	1.0 (0.4,2.7)
Cert vs Ada	0.05 (0.00,0.09)*	1.2 (0.97,1.6)	1.6 (0.8,2.8)
Cert vs Inf	0.15 (0.10,0.21)*	2.0 (1.5,2.9)*	2.4 (1.1,4.7)*
Cert vs Eta	-0.06 (-0.13,0.01)	0.9 (0.6,1.3)	1.2 (0.5,2.5)
Cert vs Gol	0.03 (-0.04,0.09)	1.1 (0.8,1.7)	1.2 (0.4,2.7)
σ	0.03 (0.01,0.05)	0.1 (0.0,0.3)	0.4 (0.1,0.6)

Table 4.7: Results from ACR analysis: Mean estimate with 80% credible intervals for each pair-wise comparison. Results with a probability of $\geq 90\%$ to favour one drug over the other are marked with a *. Outcome measures are odds ratios for ACR 20, ACR 50 and ACR 70 and % improvement for ACRcont. Ada=adalimumab; Inf=infliximab; Eta=etanercept; Gol=golimumab, Cert=certolizumab, P=placebo, σ =between trial SD

<i>Comparison</i>	<i>ACR 20</i>	<i>ACR 50</i>	<i>ACR 70</i>	<i>ACRcont</i>
Ada vs P	4.2 (3.4,5.3)*	5.6 (4.3,7.2)*	7.0 (4.3,10.9)*	0.27 (0.25,0.29)*
Inf vs P	3.5 (2.8,4.4)*	3.4 (2.7,4.4)*	3.7 (2.3,5.5)*	0.24 (0.22,0.27)*
Eta vs P	8.8 (5.5,13.9)*	12.1 (5.7,23.2)*	30.9 (4.2,123.8)*	0.32 (0.28,0.35)*
Gol vs P	3.3 (2.3,4.8)*	4.2 (2.7,6.5)*	6.5 (2.8,13.3)*	0.25 (0.21,0.28)*
Cert vs P	10.6 (8.0,14.1)*	9.5 (6.7,13.5)*	14.3 (6.8,27.0)*	0.32 (0.30,0.35)*
Inf vs Ada	0.8 (0.6,1.1)	0.6 (0.4,0.9)*	0.5 (0.3,1.0)*	-0.03 (-0.06,0.00)*
Eta vs Ada	2.1 (1.2,3.4)*	2.2 (0.96,4.2)	4.4 (0.6,19.0)	0.05 (0.01,0.09)*
Eta vs Inf	2.5 (1.5,4.1)*	3.5 (1.5,6.9)*	8.3 (1.1,37.7)*	0.08 (0.03,0.12)*
Gol vs Ada	0.8 (0.5,1.2)	0.8 (0.4,1.2)	0.9 (0.4,2.2)	-0.02 (-0.06,0.02)
Gol vs Inf	0.9 (0.6,1.5)	1.2 (0.7,2.0)	1.7 (0.7,4.1)	0.01 (-0.04,0.05)
Gol vs Eta	0.4 (0.2,0.7)*	0.3 (0.2,0.8)*	0.2 (0.0,1.8)	-0.07 (-0.12,-0.02)*
Cert vs Ada	2.5 (1.8,3.6)*	1.7 (1.1,2.6)*	2.1 (0.9,4.7)	0.05 (0.02,0.09)*
Cert vs Inf	3.0 (2.1,4.3)*	2.8 (1.8,4.3)*	3.8 (1.7,8.3)*	0.08 (0.05,0.12)*
Cert vs Eta	1.2 (0.7,2.0)	0.8 (0.4,1.8)	0.5 (0.1,3.6)	0.01 (-0.04,0.05)
Cert vs Gol	3.2 (2.0,5.1)*	2.3 (1.3,3.9)*	2.2 (0.8,6.4)	0.08 (0.03,0.12)*
σ	0.2 (0.0,0.4)	0.2 (0.0,0.4)	0.5 (0.0,0.8)	0.03 (0.01,0.04)

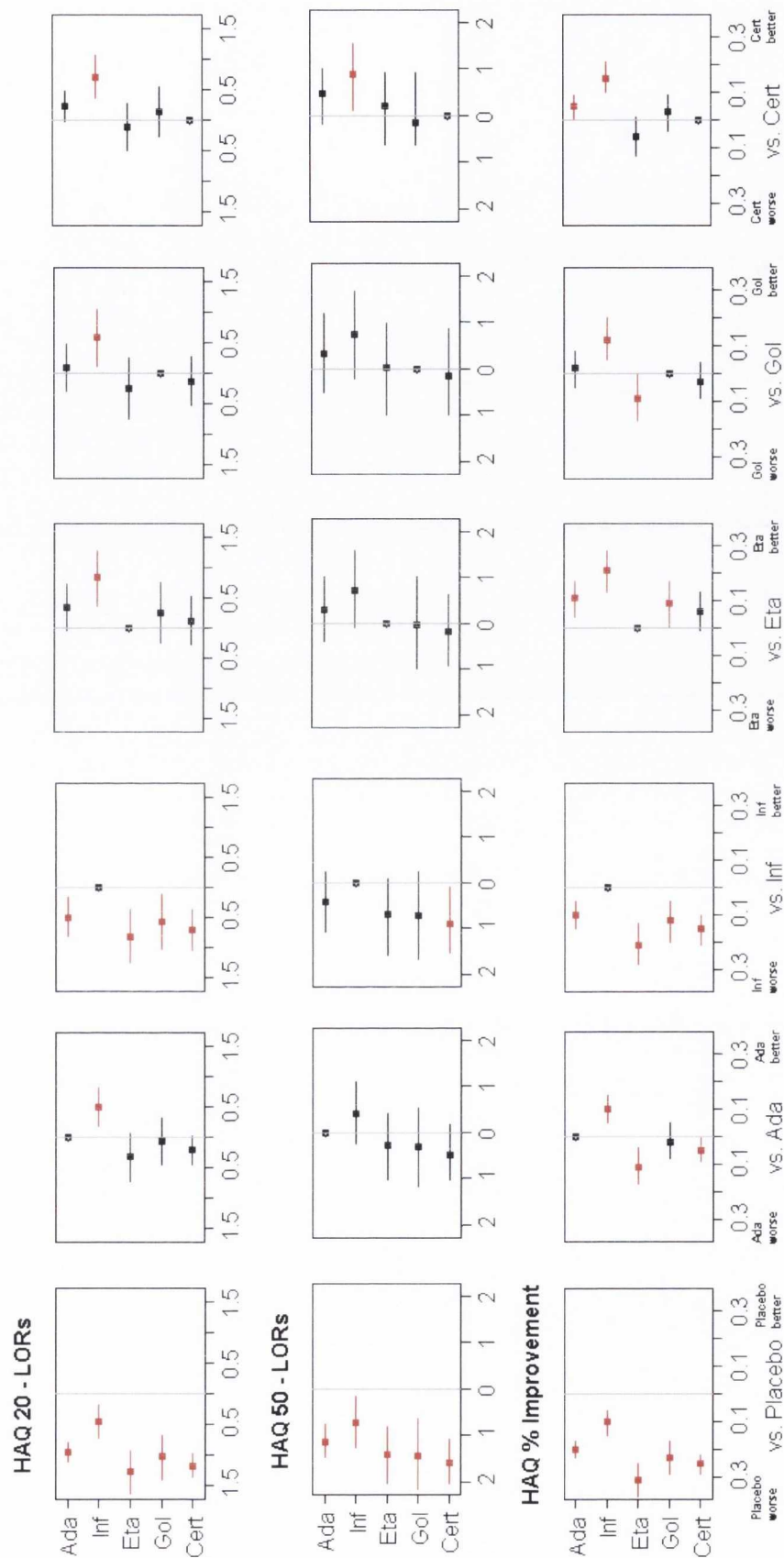


Fig. 4.5: Forest Plot of HAQ results: Plots pairwise LORs for HAQ 20 and HAQ 50 outcome and improvement differences for % improvement in HAQ score of anti-TNF agents against placebo and one another.

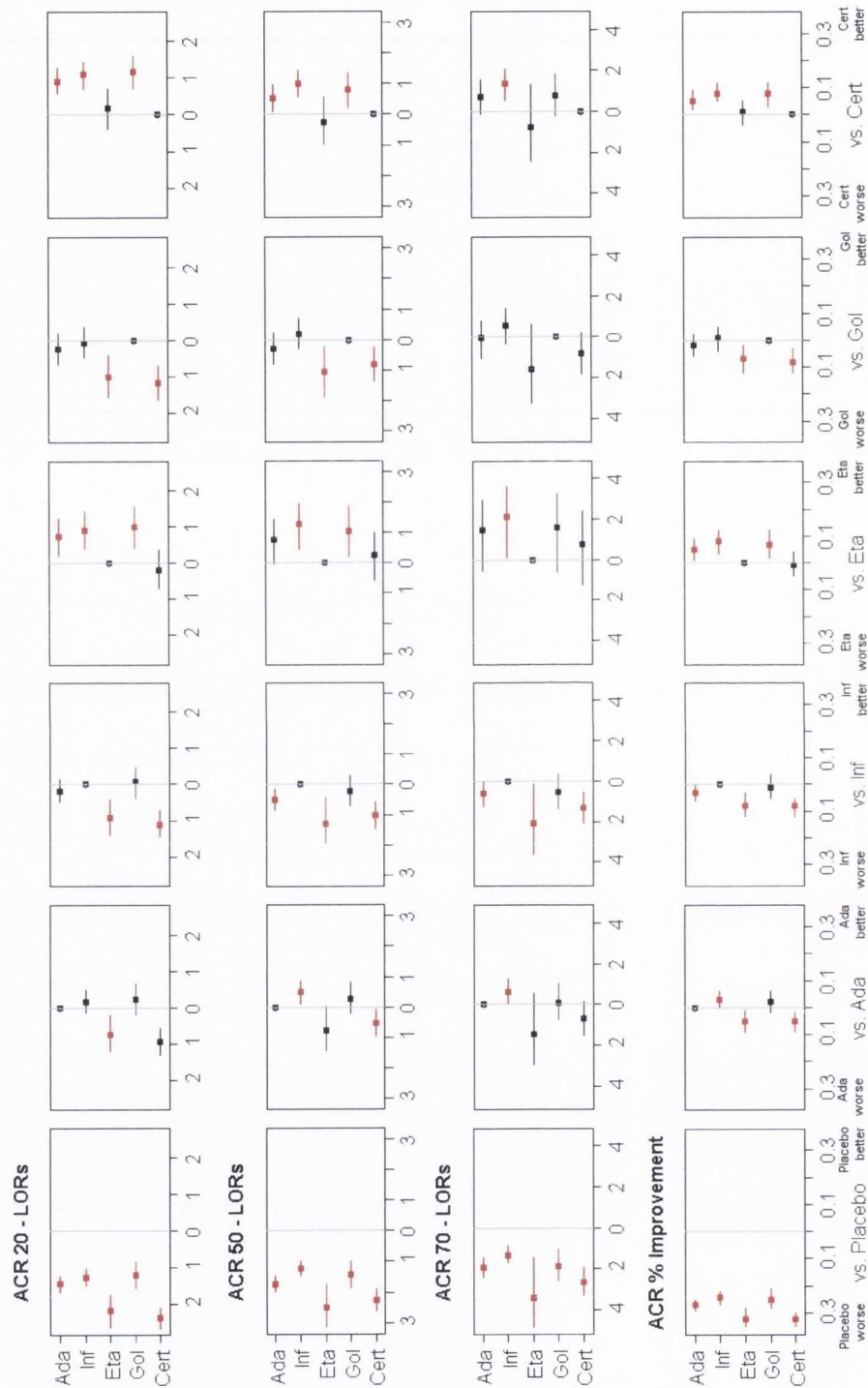


Fig. 4.6: Forest Plot of ACR results: Plots pairwise LORs for ACR 20, ACR 50 and ACR 70 outcome and improvement differences for ACRcont of anti-TNF agents against placebo and one another.

The key point is that all differences between agents detected by the binary HAQ measures are also detected by the continuous HAQ improvement measure, while not all differences detected in the continuous HAQ model are detected using the binary HAQ data; the differences detected by the binary HAQ measures are subsets of the differences detected by the continuous HAQ improvement measure. This is also demonstrated in the ACR analysis. It illustrates the point made that a continuous measure of effect has a greater power to detect a difference between treatments in an evidence synthesis.

4.4.3 Convergence

Convergence was analysed using a combination of visual inspection of the chains and diagnostic tools, introduced in chapter 3.

The effective sample size was checked, the autocorrelation plots showed a decrease with increasing lags, as can be seen for a small selection of the parameters in figure 4.7. The two chains are indistinguishable, indicating good mixing, an example of chain plots of the ACR 20 analysis is shown in figure 4.8. Raftery and Lewis diagnostics indicate an appropriate number of burnin and iterations.

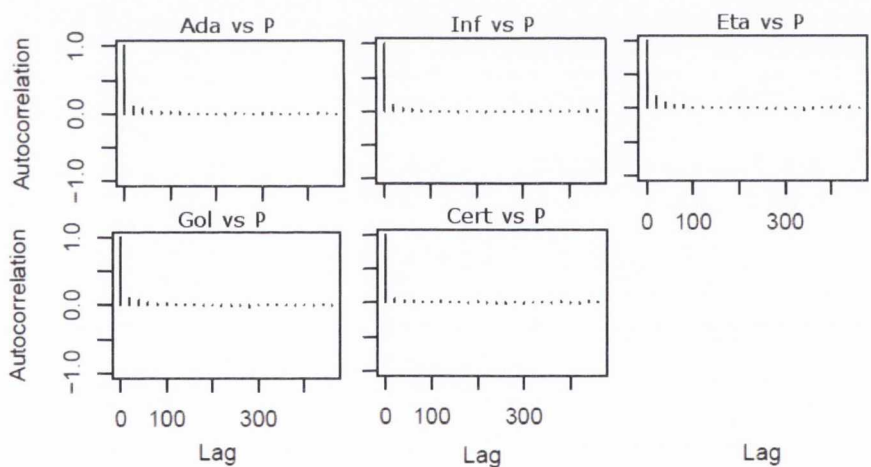


Fig. 4.7: Autocorrelation plot of a subset of the efficacy parameters from HAQ analysis.

4.5 Sensitivity Analysis

Meta-regression

The trials included for the analysis were conducted over a time period of 10 years. The model was extended to a meta-regression to include potential confounding parameters

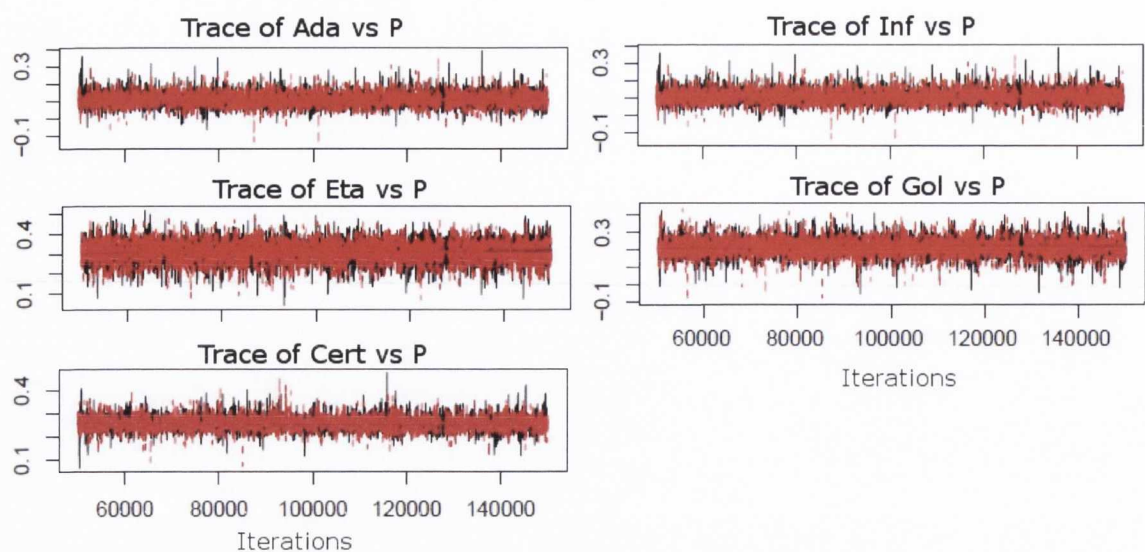


Fig. 4.8: Well behaved trace plot of a subset of the efficacy parameters from HAQ analysis.

such as year of publication, duration of disease, number of previous DMARDs and age. None of these were found to have a significant impact throughout the outcome measures and therefore were not included in the main analysis. The posterior distributions of the regression parameters are plotted in figure 4.9.

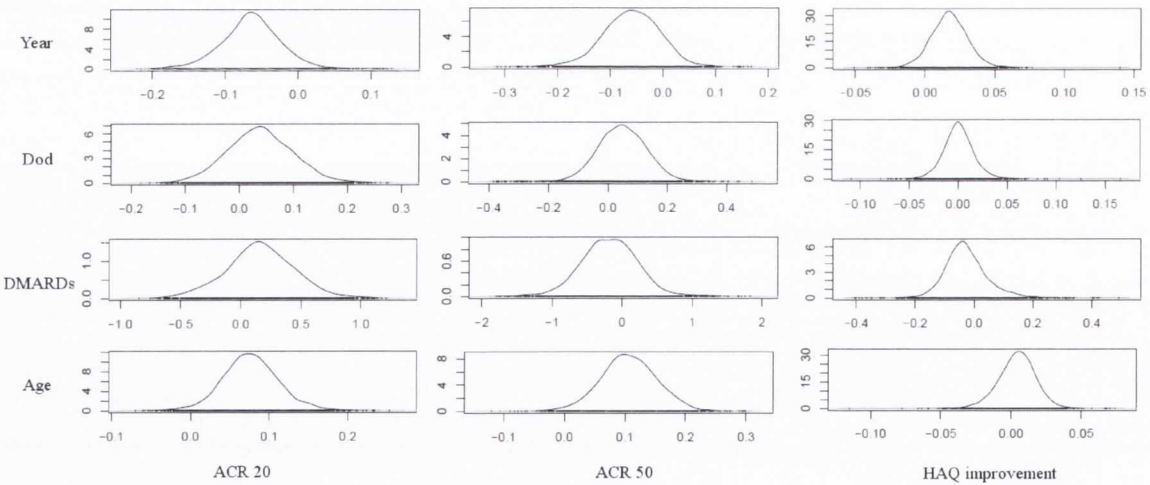


Fig. 4.9: Posterior distribution of meta-regression parameter for ACR 20, ACR 50 and HAQ improvement.

Prior distributions

Different prior distributions on the between trial variance parameter allowing for a wider range of values were tested, but did not alter the conclusions. The distributions tested were:

$$\sigma \sim duni f(0, 5)$$

$$\sigma \sim duni f(0, 7)$$

Varying the precision parameter of the normal prior distributions for the remaining parameters between 1,000 and 100,000 did not alter the results either.

Underlying continuous response

The continuous ACR response underlying the binary ACR 20, ACR 50 and ACR 70 measures is estimated using the mean ACRcont response for each of the four responder groups. For the baseline analysis meanACR was used as the underlying continuous effect. In a sensitivity analysis a model was fitted using the ACR hybrid as underlying effect (ACRhybrid). This did not have a significant impact on the results, however, adalimumab was not found to be superior to infliximab in this scenario. There was substantial uncertainty regarding the mean ACRcont response in the control arms, as no patient level data was available. An additional model has been fitted assuming the mean response in each group to be same for treatment arms and control arms (meanACR-meanACR and ACRhybrid-ACRhybrid). While this did influence the biologic treatment effect versus control, the relative efficacy between biologic treatments was not greatly affected. Again, these models did not detect a difference between adalimumab and infliximab. Outcomes of the various MTCs are summarized in table 4.8.

Scale for binary outcome measures

Caldwell et al. (2012) have shown how different scales commonly used to analyse binary data have influenced the outcomes of a MTC, using an example in childhood nocturnal enuresis. Rather than ease of interpretation, which is commonly used, they propose a combination of measure of fit (DIC) and epidemiological reasoning for selecting the scale of the outcome measure. An analysis based on risk ratios (RRs) in addition to the analysis based on ORs is therefore conducted. The OR model can be adapted to estimate RRs making a few changes; for further details see Warn et al. (2002). Results of this analysis are summarised in table 4.9. The MTC results based on the RR scale confirm the findings from the OR scale. All differences detected by the binary HAQ and ACR measures are also detected in their continuous counterpart. More differences between

Table 4.8: Results from sensitivity analysis: Mean estimate with 80% credible intervals based on alternative ACRcont measures: ACRhybrid, meanACR-meanACR, ACRhybrid-ACRhybrid. Results with a probability of $\geq 90\%$ to favour one drug over the other are marked with a *. Ada=adalimumab; Inf=infliximab; Eta=etanercept; Gol=golimumab, Cert=certolizumab, P=placebo, σ =between trial SD

Comparison	ACR hybrid	meanACR-meanACR	ACRhybrid-ACRhybrid
Ada vs. P	0.25 (0.23,0.27)*	0.13 (0.11,0.15)*	0.13 (0.11,0.15)*
Inf vs. P	0.22 (0.20,0.25)*	0.12 (0.10,0.14)*	0.12 (0.10,0.14)*
Eta vs. P	0.30 (0.26,0.34)*	0.17 (0.14,0.21)*	0.17 (0.14,0.21)*
Gol vs. P	0.23 (0.19,0.27)*	0.12 (0.09,0.16)*	0.12 (0.09,0.16)*
Cert vs. P	0.30 (0.28,0.33)*	0.17 (0.15,0.19)*	0.17 (0.15,0.19)*
Inf vs. Ada	-0.02 (-0.06,0.01)	-0.01 (-0.04,0.02)	-0.01 (-0.04,0.02)
Eta vs. Ada	0.06 (0.01,0.10)*	0.04 (0.01,0.09)*	0.04 (0.00,0.08)*
Eta vs. Inf	0.08 (0.03,0.12)*	0.05 (0.01,0.09)*	0.05 (0.01,0.09)*
Gol vs. Ada	-0.02 (-0.07,0.02)	-0.01 (-0.05,0.03)	-0.01 (-0.05,0.03)
Gol vs. Inf	0.00 (-0.05,0.05)	0.00 (-0.04,0.04)	0.00 (-0.04,0.04)
Gol vs. Eta	-0.08 (-0.13,-0.02)*	-0.05 (-0.09,0.00)*	-0.05 (-0.09,0.00)*
Cert vs. Ada	0.06 (0.02,0.09)*	0.04 (0.01,0.07)*	0.04 (0.01,0.07)*
Cert vs. Inf	0.08 (0.04,0.12)*	0.05 (0.02,0.08)*	0.05 (0.02,0.08)*
Cert vs. Eta	0.00 (-0.05,0.05)	0.00 (-0.04,0.04)	0.00 (-0.04,0.04)
Cert vs. Gol	0.08 (0.03,0.13)*	0.05 (0.01,0.09)*	0.05 (0.01,0.09)*
σ	0.03 (0.02,0.04)	0.03 (0.01,0.04)	0.03 (0.01,0.04)

Table 4.9: Results from sensitivity analysis: Mean estimate with 80% credible intervals for binary analysis based on risk ratios. Results with a probability of $\geq 90\%$ to favour one drug over the other are marked with a *. Ada=adalimumab; Inf=infliximab; Eta=etanercept; Gol=golimumab, Cert=certolizumab, P=placebo, σ =between trial SD

Comparison	HAQ 20	HAQ 50	ACR 20	ACR 50	ACR 70
Ada vs. P	1.6 (1.4,1.8)*	2.3 (1.7, 3.0)*	2.3 (1.9, 2.8)*	3.5 (2.8, 4.4)*	4.3 (3.1, 6.0)*
Inf vs. P	1.2 (1.0, 1.4)*	1.6 (1.0, 2.5)*	1.9 (1.6, 2.3)*	2.4 (2.0, 3.0)*	2.7 (2.0, 3.6)*
Eta vs. P	1.7 (1.4, 2.1)*	2.5 (1.4, 4.1)*	3.6 (2.4, 5.2)*	5.7 (3.0, 9.7)*	5.7 (2.3, 12.3)*
Gol vs. P	1.7 (1.3, 2.1)*	3.1 (1.6, 6.0)*	1.9 (1.4, 2.5)*	2.9 (1.9, 4.1)*	3.6 (2.0, 6.2)*
Cert vs. P	1.8 (1.6, 2.1)*	3.4 (2.2, 5.0)*	4.9 (3.8, 6.3)*	5.9 (4.2, 7.9)*	7.3 (4.4, 11.7)*
Inf vs. Ada	0.8 (0.6, 0.9)*	0.7 (0.4, 1.2)	0.8 (0.6, 1.1)	0.7 (0.5, 1.0)*	0.6 (0.4, 1.0)*
Eta vs. Ada	1.1 (0.9, 1.4)	1.1 (0.6, 2.0)	1.5 (1.0, 2.3)*	1.7 (0.9, 2.9)	1.3 (0.5, 3.1)
Eta vs. Inf	1.5 (1.1, 1.9)*	1.6 (0.8, 3.0)	1.9 (1.3, 2.9)*	2.4 (1.2, 4.2)*	2.1 (0.8, 4.9)
Gol vs. Ada	1.1 (0.8, 1.4)	1.4 (0.7, 2.7)	0.8 (0.6, 1.1)	0.8 (0.5, 1.3)	0.8 (0.4, 1.7)
Gol vs. Inf	1.4 (1.0, 1.8)*	1.9 (0.9, 4.3)	1.0 (0.7, 1.4)	1.2 (0.8, 1.7)	1.3 (0.7, 2.5)
Gol vs. Eta	1.0 (0.7, 1.3)	1.2 (0.6, 2.9)	0.5 (0.3, 0.8)*	0.5 (0.3, 1.0)*	0.6 (0.2, 1.9)
Cert vs. Ada	1.2 (1.0, 1.4)*	1.5 (0.9, 2.4)	2.1 (1.5, 2.9)*	1.7 (1.2, 2.5)*	1.7 (0.9, 3.1)
Cert vs. Inf	1.5 (1.2, 1.9)*	2.2 (1.2, 3.9)*	2.6 (1.9, 3.7)*	2.4 (1.7, 3.5)*	2.7 (1.5, 4.7)*
Cert vs. Eta	1.1 (0.8, 1.3)	1.4 (0.7, 2.7)	1.4 (0.9, 2.3)	1.0 (0.6, 2.1)	1.3 (0.5, 3.4)
Cert vs. Gol	1.1 (0.8, 1.4)	1.1 (0.5, 2.2)	2.6 (1.8, 4.0)*	2.1 (1.2, 3.4)*	2.0 (0.9, 4.2)
σ	0.1 (0.0, 0.2)	0.4 (0.1, 0.7)	0.3 (0.1, 0.4)	0.2 (0.0, 0.4)	0.3 (0.0, 0.4)

anti-TNF agents were detected in the continuous analysis. This confirms the enhanced sensitivity to change of continuous measures in MTC models compared to using binary outcome measures.

4.6 Limitations

Doses across treatments and treatment arms varied, which has not been accounted for in the analysis. Generalising the model to include a meta-regression for dose would raise difficulties of comparability; doses across treatments are hard to compare as well as within one treatment when the same dose is given, but with different frequency.

The analysis only includes RCT evidence. However, there exists a large body of observational data collected via registries and open label studies. The inclusion of such evidence is possible within this framework, a methodology is developed in chapter 6 of this thesis. Patient level data allowed the estimation of mean ACR response within each of the responder groups. While this provided some additional information towards estimating the continuous ACR response measure, it did not replace the full information lost. The estimated ACR_{cont} measure is still subject to high uncertainty. Patient level data of the RCTs included in the analysis would allow the exact calculation of the continuous ACR measure.

The evidence network has a star design, which does not allow testing of the assumption of treatment exchangeability. Baseline characteristics across the trials were compared and the model was extended to a meta-regression to explain the potential influence of differing demographics.

4.7 Conclusions

This analysis illustrates the enhanced sensitivity to change of continuous measures in MTC models compared to using binary outcome measures, which was recently highlighted for epidemiological studies (Bennette and Vickers (2012)). While binary measures work well for demonstrating a certain level of response, which is the primary aim of many clinical trials, the information lost when categorizing the underlying continuous response can have significant impact on the results of a mixed treatment comparison.

MTC models are often utilized when faced with a choice of agents rather than for demonstrating efficacy. In the above application, for example, it is already known that all of the anti-TNF agents provide improvement in the treatment of RA. What remains is to determine whether some of these agents work better than others. Therefore it would be

of interest to know whether an agent provides a 30% improvement or a 60% improvement. When looking at the ACR 20 outcome, no difference would be seen between these two response rates.

The above analyses have shown two things. Firstly, the enhanced power to detect differences of a continuous measure as opposed to a binary measure was shown for both, the HAQ and the ACR measures. Secondly, the choice of cut-off level (e.g. 20%, 50% or 70%) has shown to have a strong impact on the results. Different significant results were detected when choosing different cut-offs; even the sequence of treatments when ordering from best to worse changes.

It is likely that the level of heterogeneity plays an important role in the impact of dichotomizing outcome measures. In order to provide guidelines to clinical practise, it is of great interest to investigate this relationship further. A simulation study allowing for different levels of heterogeneity would allow for further insight into this relationship.

It may be argued that binary outcomes are clear and easy to interpret. However, where binary outcome measures are required, results based on continuous measures can be transformed subsequently using cut-off points (Hasselblad and Hedges (1995)).

From a clinical point of view these results show that a MTC model based on continuous outcome measures provides greater precision of estimates of efficacy than those based on binary measures. This is of great benefit when carrying out an economic evaluation.

This analysis confirms the effectiveness of anti-TNF agents in the treatment of RA. The study illustrates the enhanced ability to detect differences between treatments when using continuous measures and shows the high dependency of MTC outcomes on the cut-off level when using binary data, which may be inappropriate in such methodologies.

Rather than steering away from binary measures, the analysis aims to advocate the additional reporting of underlying continuous effect measures in clinical trials to facilitate further analyses. The importance of the involvement of statisticians in the choice of clinical measures has been recently highlighted (Senn and Julious (2009)).

Chapter 5

Impact of Structural Uncertainty in Economic Evaluations

The work in this chapter extends the analysis carried out in the previous chapter. This work was done in collaboration with Dr. Pelham Barton from the Health Economics Unit at the University of Birmingham and Dr. Roisin Adams, Chief II pharmacist at the NCPE. Pelham provided access to and information about the BRAM model, which was used for the economic modelling. Roisin provided the pharmaceutical knowledge about RA and adjusted the BRAM model to the Irish healthcare setting. Some of this work appeared in the ISPOR connections (Adams et al. (2011)).

5.1 Motivation

As described in chapter 2 results of an economic evaluation are subject to different levels of uncertainty. While the impact of parameter uncertainty has been researched extensively and one-way as well as probabilistic sensitivity analyses are an integrated part of most economic evaluations, the impact of structural uncertainty is less widely investigated (Strong et al. (2012)). A review by Bojke et al. (2009) has shown that only 13 of 241 published HTA reports by the National Health Service in the UK between 1997-2005 have taken into account some form of structural uncertainty. Structural uncertainty can relate to the inclusion or exclusion of relevant comparators or health states. For example, a relevant comparator may be in the process of licensing, and should be included in the analysis as a sensitivity analysis. Every model requires certain assumptions to be made regarding the disease process, it may not always be clear which health states should or should not be included in the model. Alternative statistical techniques may be applied to different parts of the model; for example the choice of a parametric survival model.

This also adds to the overall structural uncertainty.

It has been shown that different structural choices can have an impact on the decision regarding cost-effectiveness as well as the uncertainty surrounding the decision (Bojke et al. (2008)). It is therefore important to communicate the impact of structural uncertainty to the decision maker as part of the analysis of uncertainty.

In this chapter the impact of structural choices in the MTC model on the decision of a subsequent cost-effectiveness analysis is investigated. Structural uncertainty is quantified using generalized EVPI methods based on the parameterisation of structural choices.

5.1.1 Application in Rheumatoid Arthritis

The analysis in this chapter extends the application in RA from the previous chapter 4. The continuous MTC model is adapted to account for different structural choices regarding fixed or random effects for the biologic treatment effects and the MTX effect, yielding four separate models. HAQ improvement relative to baseline is the chosen measure of efficacy and the MTC analysis is based on the trials identified in the literature search described in the previous chapter. The four MTC models yield four sets of relative effect estimates of anti-TNF agents; these sets are used in a subsequent CUA to estimate the incremental costs and QALYs for each treatment based on each MTC model.

5.1.2 Aim of Analysis

The aim of this chapter is to determine the impact that structural choices in the MTC model have on the decision of a subsequent cost-effectiveness analysis. For a scenario analysis a number of MTC models making different structural assumptions are fitted. A CUA is subsequently run for each set of parameter estimates. Results can be directly compared in terms of impact on reimbursement decision. Generalized EVPI methods based on the parameterisation of structural choices allow for the quantification of structural uncertainty.

5.2 Methods

5.2.1 Structural Choices in MTC

The aim of this analysis is to determine the impact different MTC model assumptions have on the result of the cost-effectiveness analysis. Four MTC models, which differ in the choice of fixed and random effects on the anti-TNF efficacy and the MTX effect, are fitted. Changes occur in line 5 of equation 4.6. The fixed effects for the MTX effect are

incorporated in line 3 of the same equation.

Model 1 assumes fixed effects on both, the anti-TNF efficacy $a[k]$ and the MTX effect b :

$$\begin{aligned}\alpha[i, t[i, j]] &= a[t[i, j]] - a[bi[i, j]] \\ \beta[i] &= b\end{aligned}\tag{5.1}$$

Model 2 assumes fixed effects for the MTX effect and random effects for the anti-TNF parameter. This allows for heterogeneity among the biologic treatment effects in the different studies, while the MTX effect is assumed to be fix.

$$\begin{aligned}\alpha[i, t[i, j]] &\sim N(a[t[i, j]] - a[bi[i, j]], \tau_\alpha) \\ \beta[i] &= b\end{aligned}\tag{5.2}$$

Model 3 assumes random effects for the MTX effect and fixed effects for anti-TNF effects. Here, the effect of MTX is allowed to differ among the studies, while the effect of the biologics is assumed to be fixed.

$$\begin{aligned}\alpha[i, t[i, j]] &= a[t[i, j]] - a[bi[i, j]] \\ \beta[i] &\sim N(b, \tau_\beta)\end{aligned}\tag{5.3}$$

Model 4 assumes random effects for both, the treatment effect and the MTX effect; each of the effects are sampled from a bivariate normal distribution, parameterized by mean and variance-covariance matrix:

$$\begin{pmatrix} \alpha[i, t[i, j]] \\ \beta[i] \end{pmatrix} \sim BVN \left(\begin{pmatrix} a[t[i, j]] - a[bi[i, j]] \\ b \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix} \right)\tag{5.4}$$

The data does not allow for the estimation of the full variance-covariance matrix. Homogeneous variance is therefore assumed, as proposed by Lu and Ades (2004). Due to the dependence of $\alpha[i, t[i, j]]$ and $\beta[i]$ on $\nu[i]$ it can be shown that the covariance of $\alpha[i, t[i, j]]$ and $\beta[i]$ is equal to $\sigma = \sqrt{\frac{1}{\tau}}$, which implies the correlation to be fixed at 0.5.

5.2.2 Economic Model

The Birmingham Rheumatoid Arthritis Model (BRAM) 2009 was adapted to the Irish healthcare setting to estimate the cost-effectiveness of the agents. The BRAM is an individual sampling model, which assesses the cost-effectiveness of adding a TNF-alpha inhibitor to a sequence of DMARDs when compared with the same sequence of DMARDs without a TNF-alpha inhibitor; the comparator chosen for the analysis is leflunomide. RA is a chronic disease characterized by periods of response to treatment followed by unpredictable loss of therapeutic effect and switches to alternative treatments. The model, therefore, compares sequences of treatments rather than single agents.

As patients progress through sequenced therapy, the HAQ score is modelled to deteriorate over time, with periods of response to treatment bringing the benefit of a slower rate of disease progression, figure 5.1. At each initiation of a new therapy, patients who

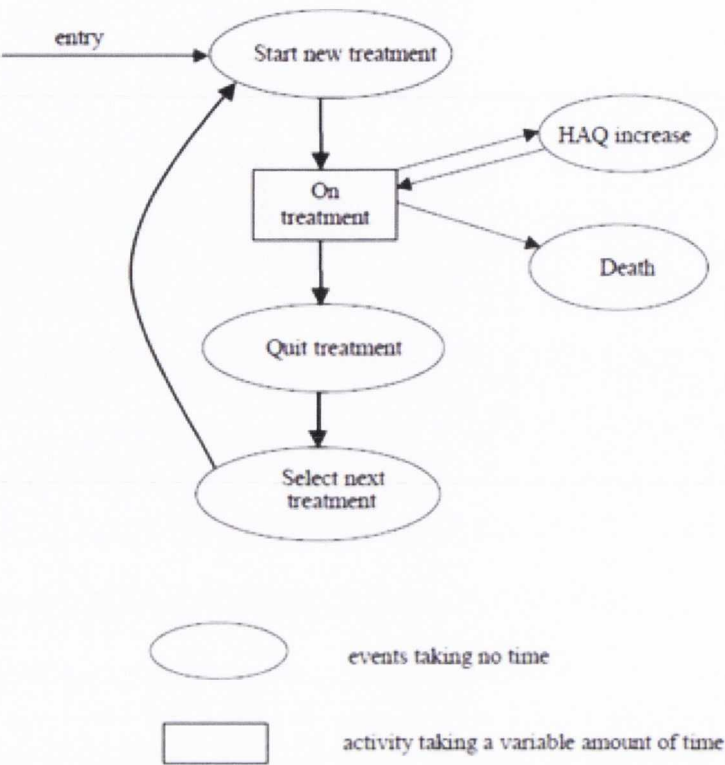


Fig. 5.1: Patient Pathway in BRAM model (taken from National Institute for Health and Clinical Excellence (2010))

respond are modelled to also receive a one-time reduction in HAQ score, a benefit that is retained until loss of efficacy occurs. This one time reduction for each of the anti-TNF agents is based on the MTC results.

Beta distributions are fitted for each of the drugs included in the analysis based on the mean and the standard deviation of the efficacy estimates derived in the MTC. Sampling from these distributions determines the one time reduction in HAQ score for the simulated patients. The beta distribution provides a flexible fit and takes values in (0,1), which is suitable for modelling a percentage improvement.

There is no specific cycle length as this varies depending on length of treatment response. A weibull survival model was fitted to registry data. The individual length of treatment response is modelled by sampling from these distributions.

Each health state, here determined by the HAQ score (values 0 to 3 in steps of 0.125), is associated with a specific cost and a specific quality of life. Costs include drug acquisition

as well as costs of screening, monitoring and administration of treatment. Quality of life data comes from an Irish patient cohort (Adams et al. (2010)).

For the present analysis, 1000 patients are randomly simulated to experience several alternative sequences of therapies. The analysis was conducted over a lifetime horizon, during which each patient accumulates costs and QALYs. Mean costs and QALYs associated with each treatment strategy are calculated based on these simulated patient histories.

For further details on the BRAM model see National Institute for Health and Clinical Excellence (2010).

5.2.3 Prior distribution

The models 5.1 - 5.4 require prior distributions for the baseline treatment effects and for the between trial standard deviation σ . As in the previous chapter, vague priors were chosen for all of these. For the biologic treatment effect parameter $a[k]$ and for the effect due to MTX b a normal distribution centered at no treatment effect 0 with a very low precision 0.0001 is chosen:

$$\begin{aligned} a[k] &\sim N(0, 0.0001) \\ b &\sim N(0, 0.0001) \end{aligned} \tag{5.5}$$

The priors do not favour any of the drugs and allow for a wide range of likely values in favour for each. A uniform distribution was chosen for the between trial standard deviation paramter $\sigma_{\alpha/\beta} = \sqrt{1/\tau_{\delta/\alpha}}$.

$$\sigma_{\alpha/\beta} \sim \text{dunif}(0, 2) \tag{5.6}$$

5.3 Results

5.3.1 Efficacy Results

Using the RCT data for HAQ improvement from table 4.2 the MTC models yield four sets of estimates for all pairwise comparisons; results are summarized in table 5.1. The mean estimates of the four models do not differ greatly; a more consistent difference can be found in the standard deviations. Models 1 and 3, which assume fixed effects for the between trial variance of the biologic treatment effect, have lower standard deviations than model 2 and 4, where random effects are assumed for the biologic treatment effect. A very high estimate for the between trial standard deviation σ is obtained in model 3. Model 3 allows for a between trial variation of the MTX treatment effect. However, the MTX treatment effect is only informed by trials where MTX is given in either the

Table 5.1: Results of four alternative MTC models: Mean % HAQ improvement (SD) for each comparison

Comparison	Model 1	Model 2	Model 3	Model 4
Ada vs. P	0.20 (0.01)	0.21 (0.02)	0.20 (0.01)	0.21 (0.02)
Inf vs. P	0.09 (0.02)	0.10 (0.04)	0.09 (0.02)	0.10 (0.04)
Eta vs. P	0.32 (0.04)	0.31 (0.05)	0.32 (0.04)	0.31 (0.05)
Gol vs. P	0.22 (0.03)	0.23 (0.05)	0.22 (0.03)	0.22 (0.05)
Cert vs. P	0.25 (0.02)	0.25 (0.03)	0.25 (0.02)	0.25 (0.03)
Inf vs. Ada	-0.12 (0.02)	-0.11 (0.04)	-0.12 (0.03)	-0.11 (0.04)
Eta vs. Ada	0.12 (0.04)	0.11 (0.06)	0.12 (0.04)	0.11 (0.06)
Eta vs. Inf	0.23 (0.05)	0.21 (0.07)	0.23 (0.05)	0.21 (0.07)
Gol vs. Ada	0.02 (0.03)	0.02 (0.05)	0.02 (0.03)	0.02 (0.05)
Gol vs. Inf	0.14 (0.04)	0.13 (0.06)	0.14 (0.04)	0.12 (0.06)
Gol vs. Eta	-0.09 (0.05)	-0.09 (0.07)	-0.09 (0.05)	-0.09 (0.07)
Cert vs. Ada	0.04 (0.02)	0.05 (0.04)	0.04 (0.02)	0.05 (0.04)
Cert vs. Inf	0.16 (0.03)	0.15 (0.05)	0.16 (0.03)	0.15 (0.05)
Cert vs. Eta	-0.07 (0.05)	-0.06 (0.06)	-0.07 (0.05)	-0.06 (0.06)
Cert vs. Gol	0.02 (0.03)	0.03 (0.06)	0.02 (0.03)	0.03 (0.06)
σ	-	0.03 (0.02)	1.00(0.58)	0.03 (0.02)

treatment arm or in the control. When given in both or none of the arms, no information for the MTX parameter is obtained. Since only one trial (Keystone et al. (2009)) has the required form, the data provides not enough information to estimate the between trial variation of the MTX parameter. Model 3 is therefore not considered for further analysis. While model 4 also assumes random effects for the MTX parameter, an additional assumption is made that the between trial variance parameter is the same for the biologic drugs as for MTX and one can hence obtain an estimate for σ .

5.3.2 CUA Results

Based on the efficacy estimates from the MTC models, a beta distribution is fitted for each anti-TNF agent; the BRAM model samples from these distributions to determine the HAQ improvement for each patient. The mean and the variance of the efficacy estimates from the MTC are used to derive the shape a and scale parameter b of the beta distribution using the following equations:

$$mean = \frac{a}{a + b}$$

$$var = \frac{a \cdot b}{(a + b)^2 \cdot (a + b + 1)}$$

Table 5.2: Beta distributions for HAQ multipliers

	Model 1		Model 2		Model 4	
	a	b	a	b	a	b
Ada	206.58	814.34	67.52	260.51	69.18	266.97
Inf	14.76	154.27	5.70	51.75	5.92	53.33
Eta	37.51	80.32	24.10	53.34	24.09	53.14
Gol	41.18	142.31	15.98	54.93	16.36	56.51
Cert	179.73	552.95	56.81	168.60	56.87	168.93

Table 5.3: Incremental costs (EUR), QALYs and ICER estimated in cost-effectiveness analysis for each anti-TNF option compared to DMARD for each model.

	Model 1			Model 2			Model 4		
	Costs	QALYs	ICER	Costs	QALYs	ICER	Costs	QALYs	ICER
Ada	59,429	0.5793	102,589	59,345	0.5882	100,899	59,353	0.5878	100,971
Inf	51,885	0.2326	223,043	51,672	0.2520	205,031	51,669	0.2526	204,545
Eta	49,197	0.7324	67,175	49,286	0.7224	68,221	49,275	0.7233	68,121
Gol	45,038	0.4963	90,755	44,970	0.5026	89,473	44,998	0.5010	89,811
Cert	43,536	0.5259	82,779	43,389	0.5406	80,264	43,400	0.5401	80,360

The estimated scale and shape parameters of the beta distribution based on each model are summarised in table 5.2. Using the efficacy estimates from the MTC analysis the BRAM model yields estimates for incremental costs and incremental QALYs of each biologic treatment compared to leflunomide as first treatment in the sequence. Mean estimates are summarised in table 5.3; a graphical representation on the cost-effectiveness plane is shown in figure 5.2.

All estimates are in the top right quadrant of the plane, meaning a gain in QALYs at an extra cost for all the drugs compared to leflunomide. The ICERs compared to leflunomide are high; the ICER for etanercept is estimated to be the lowest in all three models ranging from €67,175/*QALY* in model 1 to €68,221/*QALY* in model 2. The highest ICER at over €200,000/*QALY* is observed for infliximab. All ICERs lie above the typical threshold of €45,000/*QALY*. The scenarios show very little differences under the different structural assumptions; on the cost-effectiveness plane the estimates for each of the drugs lie very close together, making it difficult to distinguish between the models. The ICER line of €45,000/*QALY* shows that none of the treatments are cost-effective under this threshold. Using a threshold of €70,000/*QALY* etanercept becomes cost-effective based on all models, while all other treatments are still not cost-effective. Adalimumab, golimumab and infliximab are dominated by certolizumab and etanercept; meaning that compared to a combination of certolizumab and etanercept, these three

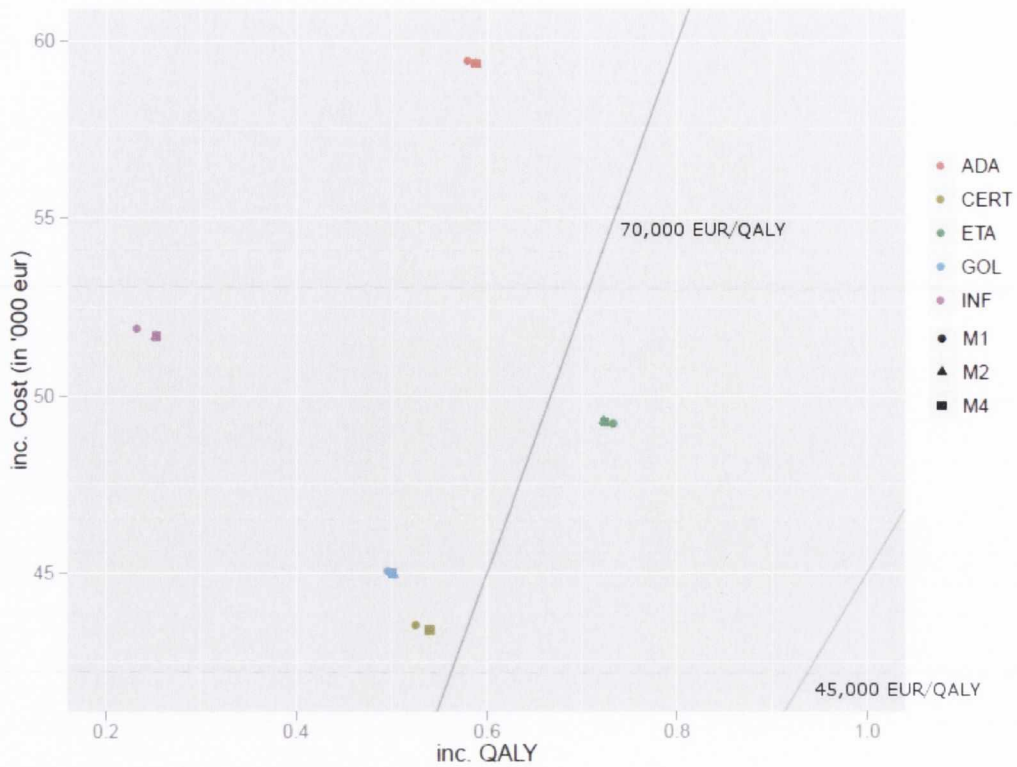


Fig. 5.2: Cost-effectiveness plane: Cost-effectiveness results of alternative models

drugs provide less QALYs for a higher cost. A probabilistic sensitivity analysis (PSA), as described in chapter 2, is run for each model. The results are shown on a scatter plot in figure 5.3. From these results, it is possible to calculate the cost-effectiveness acceptability curve (CEAC), which plots the probability of a treatment being the most cost-effective as a function of what one is willing to pay (WTP), figure 5.4. The PSA results confirm little difference between the models. For all options, the treatment of choice are leflunomide and etanercept; leflunomide for thresholds less than €67,000 - €68,000 in model 1 and €68,000 - €69,000 in models 2 and 4 and etanercept for higher thresholds.

5.3.3 EVPI for structural uncertainty

While the scenario analysis allows comparison of outcomes from different models and their impact on the decision, it does not quantify the uncertainty related to model choice. In order to do so, value of information calculations can be generalized to include structural uncertainty. The EVPI provides a monetary value for the decision uncertainty, while the EVPPI informs how much of this uncertainty is caused by a particular group of parameters. Both can be calculated as a natural extension of the PSA results. In order to adapt the method to include uncertainty due to model choice the three models are

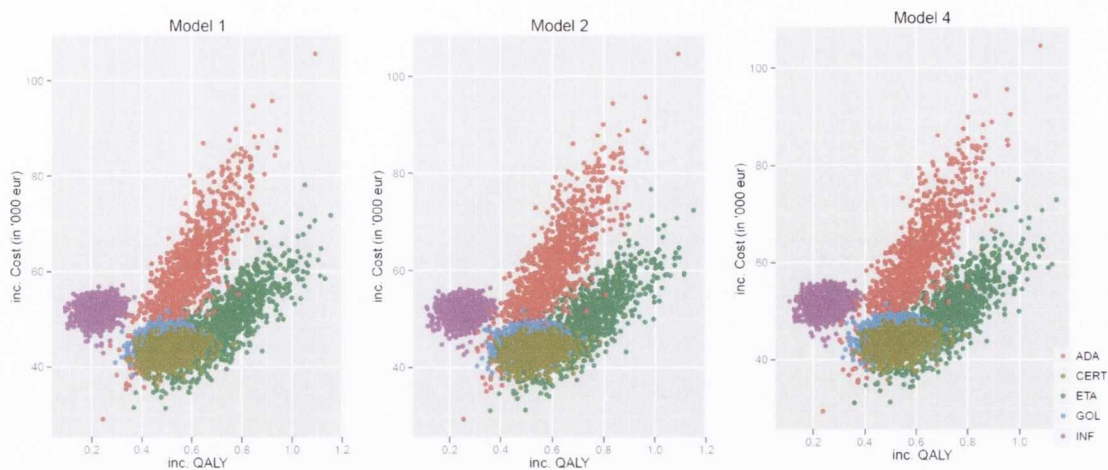


Fig. 5.3: Scatter Plots from PSA for structural choices.

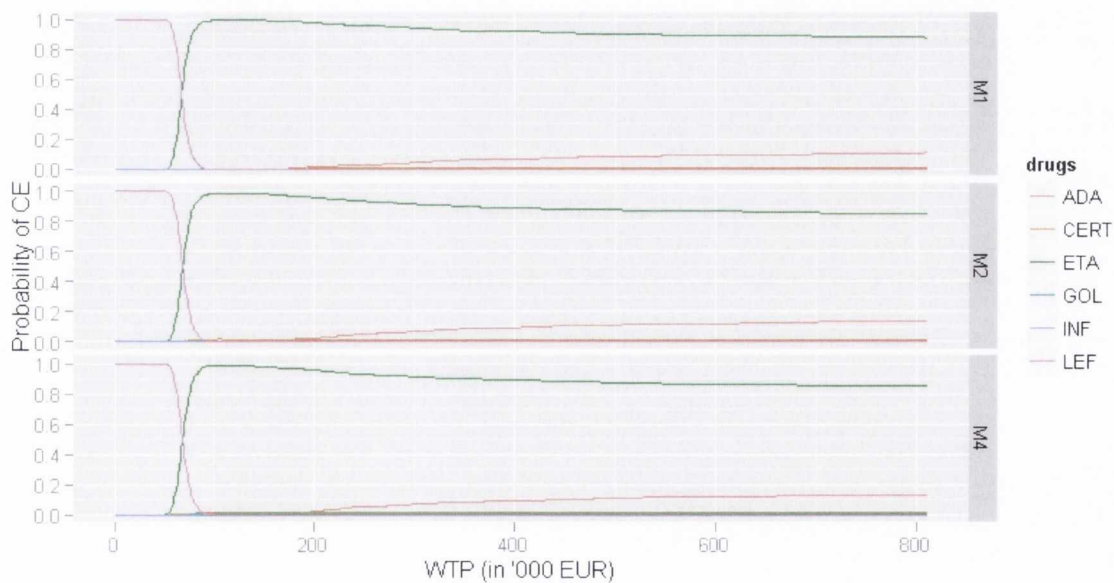


Fig. 5.4: Cost-Effectiveness Acceptability Curve (CEAC) for alternative models.

seen as submodels of a larger meta-model. In this way the combined PSA results of all four models allows for the calculation of the overall EVPI as well as the EVPPI for model choice. The definition of the EVPI and the EVPPI can be rewritten as follows:

$$EVPI := E_{\theta} \left(\max_d (NB(d, \theta)) \right) - \max_d (E_{\theta}(NB(d, \theta))) \quad (5.7)$$

$$EVPPI := E_{\theta_i} \left(\max_d (E_{\theta_i|\theta_c}(NB(d, \theta))) \right) - \max_d (E_{\theta}(NB(d, \theta))) \quad (5.8)$$

The EVPI can hence be seen as the difference between the net benefit given perfect information and the net benefit given no additional information. Similarly, the EVPPI is the difference between the net benefit given perfect information on a particular subgroup of parameters $\theta_i \subset \theta$ and the net benefit given no additional information. Due to the maximisation between two expectations, the calculation of the EVPPI is substantially more computationally expensive. A shortcut algorithm exists, however, when parameters in θ_i and θ_c are correlated or the net benefit function is non-linear, this algorithm can be substantially inaccurate (Brennan et al. (2007)).

In this application the EVPPI is calculated for the model choice parameter M , which can take three values indicating the model of choice:

$$\begin{aligned} M_i = M_1 &\Rightarrow \text{Model} = \text{Model 1} \\ M_i = M_2 &\Rightarrow \text{Model} = \text{Model 2} \\ M_i = M_4 &\Rightarrow \text{Model} = \text{Model 4} \end{aligned} \quad (5.9)$$

In relation to the decision process illustrated in figure 1.1, this analysis focusses on the uncertainty related to M in the posterior distribution $f(\theta, M|X)$, which is done by comparing

$$f(\theta|X, M_1) \quad \text{vs.} \quad f(\theta|X, M_2) \quad \text{vs.} \quad f(\theta|X, M_4).$$

Using the PSA simulations, the EVPI and the EVPPI can be easily calculated using the algorithm provided in Brennan et al. (2007). The EVPI and the EVPPI represent the value of information based on one patient, both values can be inflated to account for the whole patient population:

$$popEVPI = EVPI \cdot \sum_{t=1}^T \frac{I_t}{(1+r)^t} \quad (5.10)$$

where T is the time horizon considered, I_t the incidence over this period and r the discount rate. For this analysis a time horizon of 10 years is considered and a discount rate of 4%; this is in line with the current practice of healthcare decision making in Ireland. The incidence of RA patients eligible for biologic treatment was estimated to be 100 newly diagnosed patients each year.

Figure 5.5 plots the EVPI and the EVPPI for a wide range of WTP values, per patient and inflated to the patient population. As expected, the EVPI takes the highest values

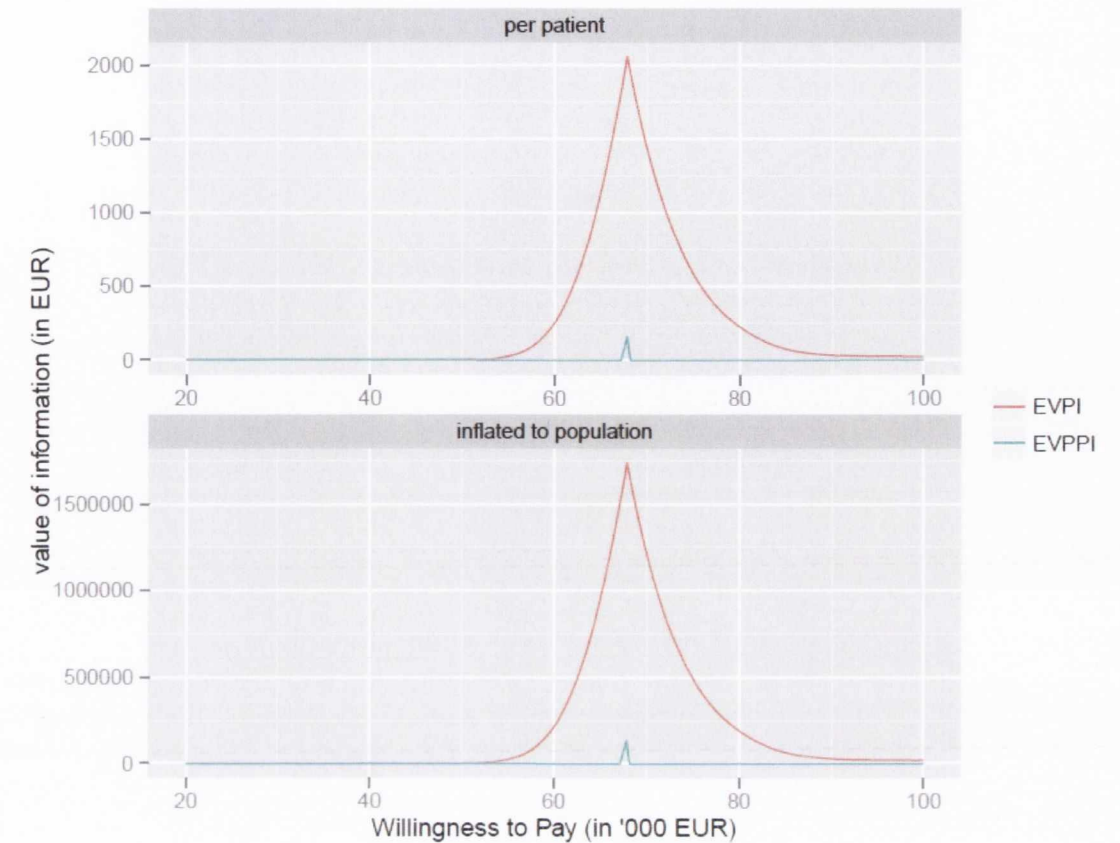


Fig. 5.5: Expected Value of Perfect Information (EVPI) and Expected Value of Perfect Parameter Information (EVPPI) for Model Choice.

in the region where the lines of the CEAC cross, i.e. for thresholds, where none of the treatments can be clearly identified as the most cost-effective option. Clearly, in this case model choice does not have a large impact on the overall decision uncertainty. Figure 5.6 plots only the population EVPPI and shows the quantitative value of decision uncertainty based on model choice. The wide range of WTP for which the EVPPI is zero reflects the stability to model choice in this application.

5.3.4 Model Choice

While the choice of model in this application did not influence the decision, ignoring structural uncertainty yields overconfidence in the estimates (Draper (1995)). The uncertainty due to structural choices has been assessed using value of information. However, in other applications structural choices may have a larger effect on the result, which may influence the decision. In such cases the question remains, on which model to base the

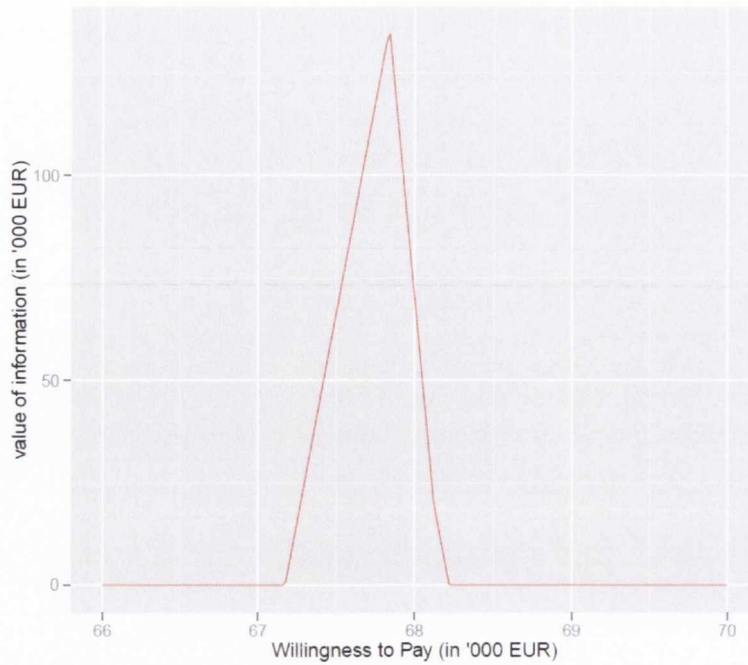


Fig. 5.6: Expected Value of Perfect Parameter Information (EVPPI) for Model Choice inflated to population values.

decision on. In his paper Draper proposes model averaging as a natural Bayesian solution to the problem, which incorporates parameter uncertainty as well as structural uncertainty in the posterior distribution for outcome y . Denoting model $M = (S, \theta)$ to comprise of structure S and parameters $\theta|S$ this yields:

$$f(y|X, \mathcal{M}) = \int_{\mathcal{M}} f(y|X, M) \cdot f(M|X) d\mathcal{M} \quad (5.11)$$

The posterior model distribution $f(M|X)$ can be decomposed as follows:

$$\begin{aligned} f(M|X) &= f(S, \theta|X) \\ &\propto p(S) \cdot p(\theta|S) \cdot f(X|\theta, S) \end{aligned} \quad (5.12)$$

$f(X|\theta, S)$ is the likelihood function, $p(\theta|S)$ is the prior distribution on parameter θ given model structure S . The new term in the equation $p(S)$ refers to the prior distribution on the structural assumptions made. Ignoring structural uncertainty is a special case of this approach, which places all weight on one structural choice. This approach may be too simple; according to Draper the task of improving this simplistic approach is to adequately specify $p(S)$. While it is not possible to consider the space of all models, a relevant subset needs to be chosen and a prior assigned. Draper advocates the inclusion of structural choices which provide a good posterior fit as well as which have predictive consequences. In this application all alternative models yield very similar results; estimates based on model averaging are therefore not required. In his paper

Draper discusses the choice of random and fixed effects as a model expansion approach in hierarchical modelling. Hoeting et al. (1999) provides a tutorial to Bayesian model averaging giving implementation details for a number of specific model classes. Jackson et al. (2009) introduced model averaging in health economic analyses, where models are combined based on weights derived from likelihood information criteria. The methods are generalized to a fully Bayesian setting in Jackson et al. (2010). Once model averaging has been carried out methods of value of information such as EVPI and EVPPI may still be applied to the resulting model.

5.4 Limitations

The aim of this chapter was to determine the impact of structural choices in the MTC model. Other structural assumptions made in the CUA, for example distributional assumptions, have not been varied. However, they may also impact on decision uncertainty. Furthermore the methods used in this chapter assume that the “true” model lies within the model space, which requires an exhaustive selection of models in order to completely quantify structural uncertainty. The methods in this chapter are therefore suitable to explore structural uncertainty within a given structural space; uncertainty outside this defined space is not included.

5.5 Conclusions

The communication of uncertainty is critical in CUAs; especially in the consideration on whether additional evidence is needed for a decision. Structural choices influence the results of such analyses, and therefore need to be incorporated.

A scenario analysis provides some insights of this uncertainty, particularly on whether different model assumptions change the decision on cost-effectiveness for a given threshold. However, only the inclusion of a model parameter into an overall model allows for the quantification of this type of uncertainty in form of EVPI and EVPPI calculations. EVPI calculations allow for the quantification of uncertainty in monetary terms; it represents the potential loss associated with taking a wrong decision. When the EVPI is high, EVPPI calculations can be used to identify those parameters which are the key drivers for the uncertainty. In this way the value of information can be used to inform research prioritization decisions. Structural uncertainty should be included in this process to capture all decision uncertainty. When the EVPPI indicates high structural uncertainty,

gathering further information on the structural choices can reduce overall uncertainty. EVPI calculation for parameter uncertainty are already an integrated part in the economic assessments of medicines in Ireland; the work in this chapter shows how structural uncertainty can be included in these calculations using the PSA results. In this case all models yield the same decision on cost-effectiveness; depending on the threshold leflunomide and etanercept are the treatments of choice. All other treatments are dominated. The stability to model choice is mirrored in the wide range of WTP for which the EVPPI of model choice is zero. When structural choices have a significant impact on the results, model averaging has been proposed to yield estimates based on the posterior distribution of model uncertainty.

Chapter 6

MTC: Including observational evidence

The work in this chapter is a further extension of the analysis in chapter 4 and has been carried out in collaboration with Dr. Roisin Adams, Chief II pharmacist at the NCPE, and Emer Fogarty, Research Associate at the NCPE. Roisin provided the pharmaceutical information and data for the application in RA; Emer conducted the literature review and provided the pharmaceutical knowledge for the application in Multiple Sclerosis. The methodology developed in this chapter is currently under revision for publication in “Statistics in Medicine” (Schmitz, Adams and Walsh (2012a)).

6.1 Motivation

RCTs are considered to be the gold standard of evidence, because the controlled setting minimizes potential bias. On the other hand, it has been argued that the experimental setting limits the external validity (Rothwell (2005)). While observational trials reflect reality well, there are other concerns such as overprecision and overestimation of the treatment effect (O’Rourke et al. (2007)).

One of the incentives to combining evidence from different sources is to make an informed decision based on all available evidence. Nevertheless, MTCs are typically restricted to evidence based on RCTs, while a wealth of evidence from different trial designs is ignored. An advantage of a Bayesian evidence synthesis as opposed to the classical approach is the possibility to systematically include a wide range of data in the form of prior information or hierarchical modelling. Sutton and Higgins (2008) have argued that only including randomized evidence in a meta-analysis might be a suboptimal approach, since data from alternative study designs, while possibly less reliable, still contains additional information

about the effect size. Being a generalization of meta analysis, the same argument holds for MTCs.

6.1.1 Aim of Analysis

Three methods have been proposed for the inclusion of data from different trial designs into a standard meta-analysis: Naive pooling, inclusion in form of prior information and an extension of the model to a three level hierarchical model (Sutton and Higgins (2008)). These methods differ in their flexibility to adjust for potential bias and to quantify consistency between trial designs. Mak et al. (2009) and Salpeter et al. (2009) use observational evidence to inform prior information. Prevost et al. (2000) have applied a hierarchical model to include non-randomized trial data in a standard meta-analytic framework. In this chapter the three methods will be generalized for the application in MTC modelling. The effects of including evidence other than RCT in a MTC and the advantages and disadvantages associated with the three methods are illustrated using two case studies. The application in RA from the previous chapters 4 and 5 is extended and a further case study in multiple sclerosis (MS) illustrates the effects in a more complex network of evidence.

6.1.2 Potential Bias in Observational Trials

While RCTs may lack generalisability due to their strict inclusion criteria and clinical setting, randomization is likely to create a balanced distribution of known and unknown confounding factors in the treatment group and the control group (McCarron et al. (2010)). This minimizes the occurrence of bias, which is why RCTs are considered to be the gold standard of clinical evidence. Observational data on the other hand is collected in real life environments and while estimates are more generalisable to a wider patient population, such non-randomized evidence can be prone to potential bias. When including observational data in evidence synthesis, it is hence important to acknowledge potential bias by adjusting the model to reflect these. There are two types of bias which may be present in observational estimates; non-systematic and systematic bias, both of which can be modelled by adjusting the likelihood function (Spiegelhalter et al. (2004); Turner et al. (2009)). Let ξ be the bias, then a non-systematic bias has the following form:

$$\xi \sim N(0, \tau_\xi) \quad (6.1)$$

Modelling bias in this way down-weights observational evidence by inflating the variance and therefore adjusts for overprecision. Modelling a systematic bias adjusts for over- or

underestimation of the treatment effect by shifting the mean by a fixed amount $\xi = \mu_\xi$. A combination of both forms of bias has the following form:

$$\xi \sim N(\mu_\xi, \tau_\xi) \quad (6.2)$$

In practice it is difficult to estimate the size of bias due to paucity of evidence. A sensitivity analysis in this respect is therefore essential. A range of applications of bias adjustments for observational evidence have been published, for example O'Rourke et al. (2007); Lilford and Braunholtz (1996). Both of these assume an overestimation of the treatment effect in observational studies. However, data does not appear to consistently over- or underestimate the treatment effect (Benson and Hartz (2000); Concato et al. (2000)). The power prior approach aids the construction of an informative prior distribution. The method utilizes historical data, while it allows adjusting for its impact (Ibrahim and Chen (2000)).

Two of the three methods described in this chapter allow for modelling of bias due to trial design.

6.2 Methods

6.2.1 Application in Rheumatoid Arthritis

In chapter 4 a model has been fitted to estimate the relative efficacy between biologic agents for the treatment of RA based on RCT evidence only. The application is extended in this chapter to include observational evidence and to illustrate the impact on the results. A systematic literature search relaxing the criteria of randomization is conducted to identify suitable observational trials; all other inclusion criteria remain the same. Details of the methods to include observational evidence are described in section 6.3.

6.2.2 Application in Multiple Sclerosis

For further illustration, a MTC to estimate the relative efficacy between disease-modifying-therapies (DMTs) for the treatment of relapsing, remitting MS is fitted. The work on this application was done in collaboration with Emer Fogarty, research associate and member of the economic review group at the National Centre for Pharmacoeconomics. Emer has provided information on the pharmaceutical background as well as conducted the systematic literature review and data collection.

MS is a chronic, autoimmune, inflammatory disease of the central nervous system, causing damage, which slows or even blocks neural transmission to and from the brain and

spinal cord, resulting in a loss of movement ability and sensation in the patient (Peterson and Fujinami (2007)). MS is the most common disabling neurological disease of young adults and is associated with significant economic and Health-Related Quality-of-Life burden (Kobelt et al. (2006); Naci et al. (2010)).

The number of available DMTs for MS is increasing and consequently robust methods for assessing the relative effectiveness of these agents is a priority for clinicians and health-policy decision-makers alike. This requires consideration of all evidence on comparative effectiveness. Currently there are seven DMTs licensed for relapsing remitting MS and two further DMTs awaiting approval: interferon beta 1a 22mcg (6MIU) (Rebif[®] 22), interferon beta 1a 44mcg (12MIU) (Rebif[®] 44), interferon beta 1a 30mcg (6MIU) (Avonex[®]), interferon beta 1b 8MIU (Betaferon[®]), glatiramer acetate (GA), natalizumab, fingolimod, teriflunomide and laquinimod. There is a large number of RCTs and observational data available, comparing some of these treatments with each other or to placebo. A MTC in this area allows for the estimation of relative efficacy between all of these treatments. Particular interest lies in the comparison of new agents (natalizumab, fingolimod, teriflunomide and laquinimod) with older agents.

As in RA, efficacy in MS can be measured using a number of outcomes, more details on the variety of measures can be found in Goldman et al. (2010). The annualized relapse rate (ARR) is used in this analysis, since it is the most widely reported outcome in clinical trials.

Definition 6.1 *The annualized relapse rate (ARR) is defined as the ratio of the number of relapses R and the number of patient years of follow-up $T := \sum_{i=1}^N t_i$, where N refers to the number of patients involved in the trial and t_i is the time spent on treatment by patient i . Patient years of follow-up differs to patient years on treatment, because those not on treatment are also followed up if possible, in order to comply with intention-to-treat analysis.*

$$ARR = \frac{R}{T}$$

Relative efficacy between 2 treatments A and B is measured on the log scale as the difference between their ARR:

$$\Delta_{AB} = \log(ARR_A) - \log(ARR_B)$$

Negative values favour the treatment over the comparator.

To use this outcome measure, the MTC model in chapter 4 needs to be adapted to allow for a poisson likelihood. The model then takes the following form:

$$\begin{aligned}
R[i, j] &\sim \text{Pois}(\lambda[i, j]) \\
\lambda[i, j] &= T[i, j] \cdot p[i, j] \\
\log[p[i, j]] &= \nu[i] + \delta[i, j] \\
\delta[i, b[i, j]] &= 0 \\
\delta[i, j] &= \delta[i, t[i, j]] \\
\delta[i, t[i, j]] &\sim N(\mu_\delta[i, t[i, j]], \tau_\delta) \\
\mu_\delta[i, t[i, j]] &= d[t[i, j]] - d[b[i, j]] \\
d[1] &= 0 \\
IC[b, k] &= d[b] - d[k] \\
d[k] &\sim [-, -] \quad \tau_\delta \sim [-, -] \quad \nu[i] \sim [-, -]
\end{aligned} \tag{6.3}$$

$R[i, j]$ is the number of relapses in arm j of study i , $T[i, j]$ refers to the number of patient years in arm j of study i and therefore, $p[i, j]$ is the ARR of arm j in study i . $\delta[i, j]$ is the relative log ARR of the treatment given in trial arm j of study i compared to the associated baseline treatment. The remaining parts of the model are the same as in the RA case. Effects of the same comparison within one trial are assumed constant, while random effects are assumed for between trial variation. Each effect can be written in terms of basic parameters, representing the treatment effect relative to one baseline treatment defined in $d[1] = 0$.

6.3 Statistical Model

6.3.1 Naive Pooling

Naive pooling treats all trial designs the same; one simply pools across all trials regardless of their design. This is the simplest approach to combining different trial designs; it is not possible to down-weight designs of lesser quality or to adjust for potential bias. In the RA application pooling across all trials has one advantage: the resulting evidence network contains closed loops, see figure 6.1, which allows the estimation of consistency between direct and indirect evidence. The MTC model for naive pooling remains the same. Since the observational trials contain more than 2 treatments per trial, a term allowing for the correlation between trial arms has been included in the code. As proposed by Dias et al. (2012), inconsistency is first checked by fitting a model relaxing the consistency assumptions and comparing it to the original model. The deviance contributions of the data points can help to identify inconsistencies in the network. When trial level data is

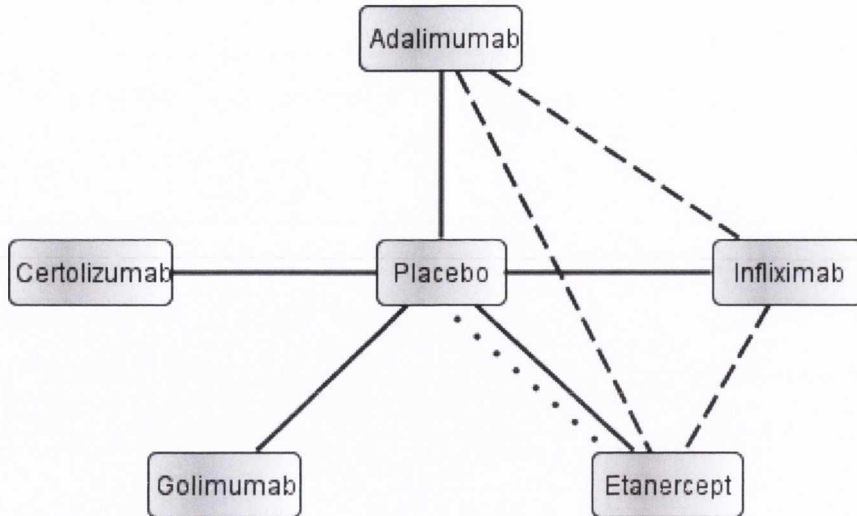


Fig. 6.1: RA case study: Combined Network Diagram: Solid edges refer to RCT evidence; dashed edges indicate observational evidence; dotted edge refers to a matched observational trial.

available, a node-split method has been proposed to estimate the consistency between direct and indirect evidence (Dias et al. (2010)). The evidence for each comparison which is part of a closed loop, where inconsistency was indicated, is split into evidence based on direct and evidence based on indirect sources. The code for this method provided by Dias et al. (2010) can be easily adapted to fit this model.

6.3.2 As Prior Information

In a Bayesian framework observational data can be used to inform prior information. For this approach observational trials are analysed separately and the resulting posterior distribution is then used as prior information for the RCT model. Only slight changes in the model are necessary. Results from the observational analysis inform the basic parameters in the RCT model. The strongest link of the observational network should be used in order to make full use of the additional information. The basic parameter should therefore be adapted to correspond to the relative efficacy of each treatment versus the treatment, which has the most direct connections to the other treatments in the network of observational evidence. The comparator of the basic parameter is defined by $d[1] = 0$ in equation 4.6 for the RA model and in equation 6.3 for the MS model. In this case 1 refers to placebo, and it needs to be adapted to reflect the strongest link.

Using observational data as prior information allows adjustment for bias. The potential

biases associated with observational data described earlier can be modelled by adjusting the prior distribution. To down-weight observational information the variance parameter can be inflated and in order to adjust for overestimation or underestimation of the treatment effect the mean of the prior information can be shifted.

6.3.3 Hierarchical Modelling

A three level Bayesian hierarchical model introduces a study type level between the study level and the overall level. A schematic illustration of such a model for the simplest case of an indirect comparison can be found in figure 6.2. This is the most flexible approach

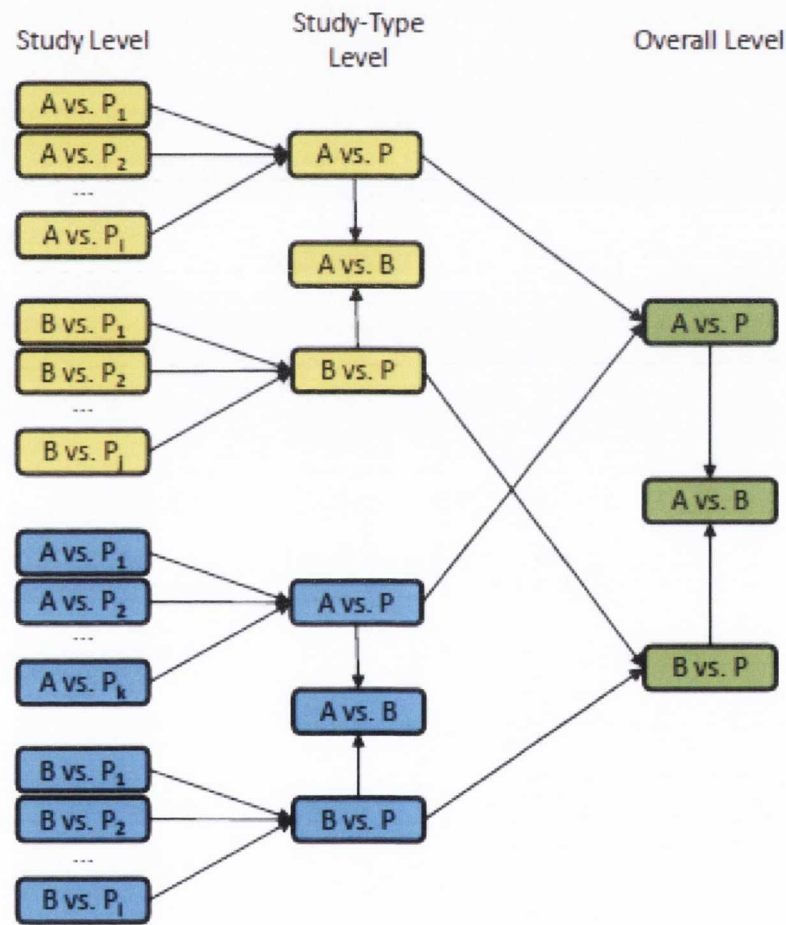


Fig. 6.2: Three level hierarchical model: Schematic graph of three level hierarchical model for simplest case indirect comparison, where trials comparing drug A and Placebo (P) are combined with trials comparing drug B and P. Evidence from different trial design is colour marked: Yellow refers to RCTs and blue to observational trials. Evidence from trials of the same design is combined to study type level estimates; study type estimates are in turn combined to get overall estimates (green). Pairwise indirect estimates can be obtained on study type level as well as overall level.

to include different forms of evidence. In addition to being able to adjust for systematic bias and down-weight different trial designs, this method allows the direct comparison of estimates on study type level to overall estimates and therefore gives an estimate of consistency between the designs. The model accounts for between design heterogeneity and overall estimates become more conservative when study type estimates differ a lot. Furthermore, more than two trial types can be included in the analysis, which is not possible in the previously described methods.

Within one model RCT data and observational data are each fit separately as described in equation 4.6 for the RA application and in equation 6.3 for the MS application; i.e. for each of the parameters introduced in the model there exists a version for each trial design. Let $a_{RCT}[k]$ and $a_{OBS}[k]$ be the treatment effects of drug k versus the baseline treatment based on RCT evidence and observational data, respectively. Assuming that RCT and observational evidence is exchangeable, the trial design estimates are then combined to an overall measure of treatment effect using random effects:

$$\left. \begin{array}{l} a_{RCT}[k] \\ a_{OBS}[k] \end{array} \right\} \sim N(a[k], \tau) \quad (6.4)$$

The effect estimates based on different trial designs are drawn from a normal distribution with a mean particular for each drug $a[k]$ and a precision τ , representing the between trial design heterogeneity. Prior distributions need to be specified for $a[k]$ and $\sigma = \sqrt{\frac{1}{\tau}}$; the following vague prior distributions are used:

$$\begin{aligned} a[k] &\sim N(0, 0.0001) \\ \nu[i] &\sim N(0, 0.0001) \\ \sigma &\sim \text{unif}(0, 0.25) \end{aligned} \quad (6.5)$$

A normal distribution with mean 0 and precision 0.0001 has a 95% confidence interval of (-196,196), which is sufficiently vague for measuring the relative treatment effects $a[k]$ as well as the baseline effects $\nu[i]$, which are measured as proportions and differences between proportions for the RA model and as log relapse rates and the differences between these for the MS model. For the between trial standard deviation a uniform prior in [0,0.25] is sufficiently vague, since it allows the treatment effect to differ up to 100% between trials in the RA case and covers more than the estimated absolute log ARR of the MS treatments.

Relative efficacy between agents k and l can now be calculated from the baseline parameters as follows:

$$IC[k, l] = a[k] - a[l] \quad (6.6)$$

The assumption of exchangeability between RCT evidence and observational evidence in equation 6.4 can be relaxed by adjusting for potential bias in the various trial designs. One can adjust for overestimation using an additive factor to the mean or for overprecision using a multiplicative factor to the variance. Equation 6.4 then takes the following form:

$$\begin{aligned} a_{RCT}[k] &\sim N(a[k], \tau) \\ a_{OBS}[k] &\sim N(a[k] + \mu_{\xi}, \tau \cdot w_{\xi}) \end{aligned} \tag{6.7}$$

The WinBUGs code for the three level hierarchical model can be found in appendix B.

6.4 Results: Application in Rheumatoid Arthritis

6.4.1 Data

Relaxing the inclusion criteria of the above literature search to include trials other than RCTs identified the observational data suitable for the inclusion in the analysis. Figure 6.3 shows the network of available evidence and data extracted from the observational trials is summarised in table 6.1.

Two of the trials include data collected from registries (Italy and the Netherlands).

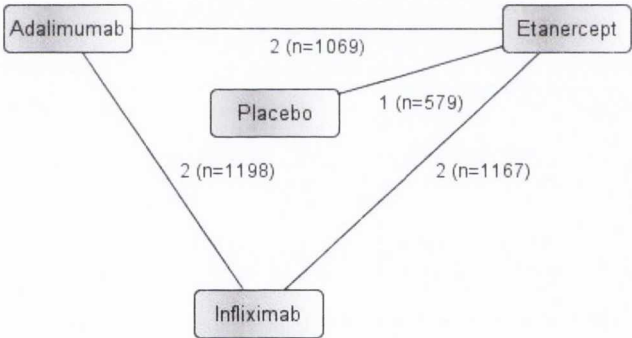


Fig. 6.3: RA case study: Network diagram of observational evidence: Edges are labelled with the number of studies and the total number of patients included in each comparison.

These registries were established to monitor the safety of these drugs and also to assess the effectiveness of these agents in the real life setting. The Dutch Rheumatoid arthritis Register (DREAM) (Kievit et al. (2008)) was established to evaluate the effects of adalimumab, etanercept and infliximab on disease activity, functional ability, quality of life and medication costs in a real life setting. Bazzani et al. (2009) assessed the effectiveness of anti-TNF α agents by analysing the primary clinical outcomes in patients with active RA. Klareskog et al. (2006) carried out an open label extension to assess the long-term safety and efficacy of etanercept in patients with RA; it is a one armed trial.

Table 6.1: RA case study: Observational trial data: Number of patients N; improvement in HAQ score Δ HAQ; HAQ score at baseline HAQ_b ; Age (years); Disease duration DoD (years); Number of previous DMARDs; + indicates additional treatment with MTX; $\diamond$ Klareskog et al. (2006) is an open label extension, the placebo arm is taken from Weinblatt et al. (1999). NA=not available; Ada=adalimumab; Inf=infliximab; Eta=etanercept; Gol=golimumab, Cert=certolizumab, P=placebo.

Trial	Arm	N	Δ HAQ (SD)	HAQ_b	Age	DoD	DMARDs
Bazzani et al. (2009)	Eta+	230	0.34 (NA)	1.23	56	9	3
	Ada+	283	0.34 (NA)	1.20			
	Inf+	497	0.34 (NA)	1.50			
Kievit et al. (2008)	Eta+	289	0.35 (0.5)	1.4	55	7	3
	Ada+	267	0.42 (0.5)	1.3			
	Inf+	151	0.23 (0.5)	1.4			
Klareskog et al. (2006) $\diamond$	Eta+	549	0.80 (NA)	1.8	53	13	3
	P+	30	0.4 (NA)	1.5			

Particular concern about one armed trials has been expressed by Ades et al. (2007); they “very strongly recommend that single arm studies, whether RCT or observational study, are excluded [from MTCs]”. There is agreement that unadjusted MTC methods do not provide reliable estimates. In this case this large trial must either be excluded from the analysis, or an appropriate comparison must be found. In order to do this, a control arm from a trial, whose baseline characteristics closely reflect those of Klareskog et al. (2006), is matched; the MTX arm of Weinblatt et al. (1999) has been used. Since the construction of this link to placebo may introduce additional bias, this trial has been excluded from the analysis in a sensitivity analysis. Baseline demographics of all trials are also summarized in table 6.1.

6.4.2 Naive Pooling

Table 6.2 summarises the relative efficacy results from naive pooling. Results based on the RCT evidence are only listed for comparison. The results reveal how the additional information from observational data strengthens the evidence; the standard deviations of estimates including placebo, adalimumab, infliximab and etanercept become smaller when including observational data. Estimates not including these treatments do not change notably. Furthermore, one can see the impact of additional evidence on the mean effect estimates. The strongest effect is found in the relative efficacy of etanercept versus placebo. Based only on RCT evidence, etanercept is estimated to improve the HAQ score by 31%, this effect shrinks to 21% when including observational trials. This

Table 6.2: RA case study: Naive Pooling Results: Difference in % HAQ improvement for each pair wise comparison (standard deviation). DIC: Deviance Information Criterion; $\overline{D}$: posterior mean residual deviance; p_D : effective number of parameters. Ada = adalimumab, Inf = infliximab, Eta = etanercept, Gol = golimumab, Cert = certolizumab, P = placebo.

Comparison (A vs B)	RCT only	Naive Pooling	Inconsistency Model
Ada vs P	0.21 (0.03)	0.22 (0.02)	0.21 (0.03)
Inf vs P	0.11 (0.04)	0.13 (0.03)	0.11 (0.04)
Eta vs P	0.31 (0.05)	0.21 (0.03)	0.26 (0.05)
Gol vs P	0.21 (0.05)	0.21 (0.06)	0.21 (0.05)
Cert vs P	0.26 (0.03)	0.26 (0.03)	0.26 (0.03)
Inf vs Ada	-0.10 (0.05)	-0.09 (0.03)	-0.10 (0.04)
Eta vs Ada	0.10 (0.06)	-0.01 (0.03)	-0.04 (0.04)
Gol vs Ada	0.01 (0.06)	-0.01 (0.06)	
Cert vs Ada	0.05 (0.04)	0.04 (0.04)	
Eta vs Inf	0.20 (0.07)	0.09 (0.03)	
Gol vs Inf	0.11 (0.07)	0.09 (0.07)	
Cert vs Inf	0.15 (0.05)	0.13 (0.04)	
Gol vs Eta	-0.10 (0.07)	0.00 (0.07)	
Cert vs Eta	-0.06 (0.06)	0.05 (0.05)	
Cert vs Gol	0.04 (0.06)	0.04 (0.07)	
DIC ($\overline{D}, p_D$)	-91.49 (-113.1,21.7)	-115.2 (-143.8,28.3)	-114.5 (-143.4,28.9)

suggests discrepancy between evidence from RCTs and evidence from observational trials. When including evidence other than RCTs the difference of effect between the agents is less.

In the above model all comparisons can be written in terms of the basic parameters. In order to detect inconsistencies in the network based on RCT as well as observational data a model relaxing this consistency assumption is fit. The model estimates all relative treatment effects for which direct evidence is available without forcing the consistency in loops in the network. In the presence of inconsistency this model should provide an improved fit to the data. Further details and the code for this model is provided by Dias et al. (2012). Figure 6.4 plots the posterior mean deviance of each data point based on the model relaxing the consistency assumption against the original model. The contributions

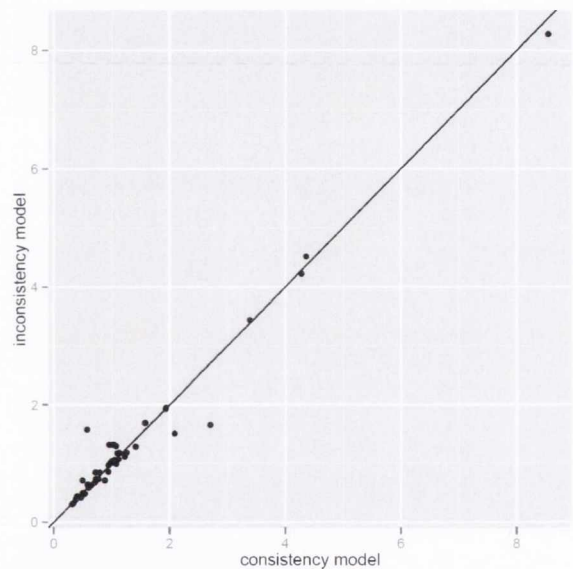


Fig. 6.4: RA case study: Plot of the individual data points’ posterior mean deviance contribution for the original model (horizontal axis) and the model relaxing the consistency assumption (vertical axis).

to the deviance are similar in both models. Arm 3 and 4 of Van De Putte et al. (2004) as well as arm 4 of Miyasaka (2008) and arm 2 of Maini et al. (1999) yield a deviance higher than expected. This can be attributed to the fixed effects assumption within trials. The efficacy estimates in these trial arms differ from those in the remaining arms of the same trial. The efficacy estimates based on the inconsistency model are also summarized in table 6.2. Most parameter estimates are very similar for both models, which suggests no evidence of inconsistency in the network. A slightly larger change is observed for the estimates including etanercept. For these parameters the node split method can

Table 6.3: RA case study: Node Split Results. Mean difference in % HAQ improvement based on direct and indirect evidence (standard deviation). DIC: Deviance Information Criterion; $\bar{D}$: posterior mean residual deviance; p_D : effective number of parameters. Ada = adalimumab, Inf = infliximab, Eta = etanercept, P = placebo.

	<i>Eta vs P</i>	<i>Ada vs Eta</i>	<i>Inf vs Eta</i>
DIC ($\bar{D}, p_D$)	-115.6 (-143.4, 27.8)	-115.2 (-142.9, 27.7)	-115.5 (-143.1, 27.6)
direct	0.26 (0.04)	0.04 (0.04)	-0.06 (0.04)
indirect	0.17 (0.04)	-0.05 (0.05)	-0.15 (0.05)

be used to estimate the inconsistency. The node-split method splits the evidence for the comparison of interest into direct and indirect sources. The differences between the estimates gives a measure of the inconsistency. The DIC is reported as a measure of model fit. In the presence of inconsistency the node-split method provides an improvement in model fit. Results are summarized in table 6.3. While posterior distributions overlap, the results show a discrepancy between direct and indirect estimates for comparisons including etanercept. However, the decreases in DIC are relatively small indicating a similar fit for all models. A change in DIC of more than 2 is said to indicate a significant improvement in model fit (Spiegelhalter et al. (2002)).

6.4.3 As Prior Information

For this approach observational data is used to inform the prior distribution of the basic parameters. Etanercept provides the strongest link in the observational network, being the only treatment which is directly connected to all other treatments. As described above the model has been adapted to define etanercept as the basic treatment; $a[1]$ now represents the efficacy of placebo versus etanercept, $a[2]$ the efficacy of adalimumab versus etanercept and $a[3]$ the efficacy of infliximab versus etanercept.

Information to inform the prior distribution for these three parameters is derived from analysing observational data alone. No adjustment for MTX needs to be made in this case, since it is given in each trial arm of all studies and hence cancels out due to the assumption made in equation (4.3). Mean and precision of the posterior distributions based on the MTC of observational data alone provide the following inputs for the prior distributions of the basic parameters:

$$\begin{aligned}
 a[1] &\sim N(-0.18, 71) \\
 a[2] &\sim N(0.04, 168) \\
 a[3] &\sim N(-0.07, 168)
 \end{aligned}
 \tag{6.8}$$

Table 6.4: RA case study: Results using observational data as prior information: Difference in % HAQ improvement for each biologic agent versus placebo (standard deviation). Ada = adalimumab, Inf = infliximab, Eta = etanercept, P = placebo.

Comparison	OBS as Prior
Ada vs P	0.22 (0.03)
Inf vs P	0.12 (0.04)
Eta vs P	0.24 (0.04)

In line with WinBUGs the normal distribution is parameterised with the mean and the precision parameter. Results using observational data to inform the prior are summarized in table 6.4. It is possible to reduce the weight given to observational data adjusting for a bias of the form given in equation 6.1, which results in the inflation of the variance parameter:

$$a[i] \sim N(\mu_i, \tau_i \cdot w_\xi) \tag{6.9}$$

where w_ξ is the weight given to the observational evidence. Figure 6.5 illustrates the impact different values of w_ξ have on the efficacy estimates. Estimates for golimumab and certolizumab are not affected since there is no additional information provided and are hence not included in the graph. The results show how decreasing weight on observational information shifts the efficacy estimates towards the estimates based on RCTs only. This effect is most evident for the etanercept versus placebo estimate, where the largest difference between estimates is observed. In order to adjust for overestimation of the treatment effect, an additive factor μ_ξ onto the mean of the prior distribution is introduced.

$$a[i] \sim N(\mu_i + \mu_\xi, \tau)$$

The results assuming a 30% overestimation and 30% underestimation of the biologic treatment effect versus placebo are shown in figure 6.6. The results are not sensitive to bias due to over- or underestimation of the treatment effect. This can be explained by the structure of the observational evidence. Its strength lies in the relative efficacy of adalimumab, infliximab and etanercept, rather than in the comparisons with placebo. The relative efficacies between the biologic treatments are not affected greatly when assuming a 30% over- or underestimation of each agent when compared to placebo.

6.4.4 Bayesian Hierarchical Model

Fitting a three level hierarchical model yields overall estimates of efficacy as well as estimates for each study type. Table 6.5 summarises the biologic treatment effect estimates

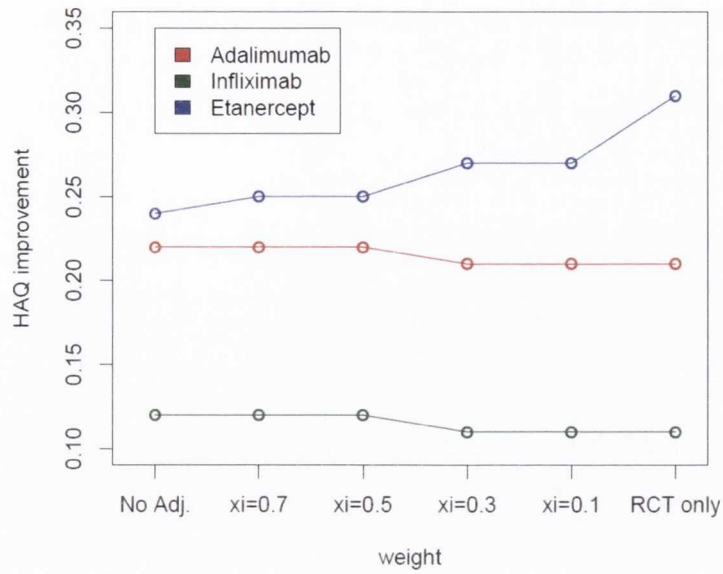


Fig. 6.5: RA case study: Adjusting for overprecision using observational data to inform prior. Results are shown for different values of w_{ξ} : Estimates shift towards the RCT only results as weight on observational evidence decreases.

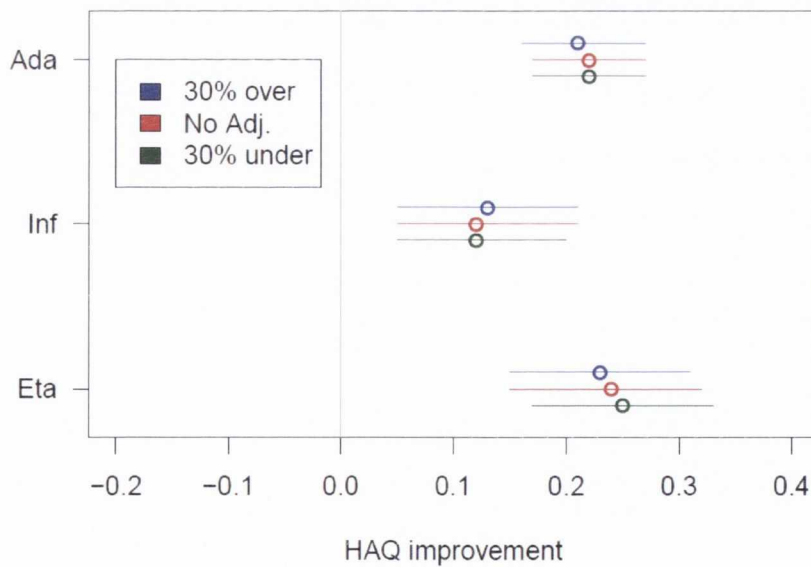


Fig. 6.6: RA case study: Effects of adjusting for a 30% over- and 30% underestimation of the treatment effect when using observational data as prior information

Table 6.5: RA case study: Hierarchical Model Results: Difference in % HAQ improvement for each biologic agent versus placebo (standard deviation) on study type level and overall level. Ada = adalimumab, Inf = infliximab, Eta = etanercept, Gol = golimumab, Cert = certolizumab, P = placebo.

Comparison	Overall	RCT level	OBS level
Ada vs P	0.22 (0.07)	0.21 (0.02)	0.23 (0.06)
Inf vs P	0.12 (0.08)	0.11 (0.04)	0.13 (0.06)
Eta vs P	0.24 (0.08)	0.28 (0.06)	0.21 (0.06)
Gol vs P	0.21 (0.05)	0.21 (0.05)	
Cert vs P	0.26 (0.03)	0.26 (0.03)	

overall as well as on study type level. The results show how RCT estimates and observational estimates are combined to overall estimates, which lie in between the trial design estimates. The overall mean effect estimates are the same as when including observational evidence as prior information. However, the standard deviation is higher in the 3 level model, since the model allows for uncertainty caused by combining evidence from different trial designs. Figure 6.7 shows shrinkage from data through each level to

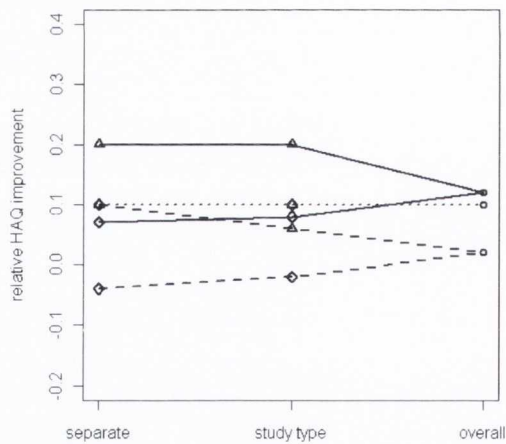


Fig. 6.7: RA case study: Plot showing the shrinkage from data through each level to the estimate of the overall mean. ◇: estimates from observational evidence; △: estimates from RCTs. Solid line: Etanercept vs. Infliximab; dashed line: Etanercept vs. Adalimumab; dotted line: Adalimumab vs. Infliximab

the estimate of the overall mean. Borrowing strength across the network results in the trial design level estimates being drawn towards the overall mean. The graph also shows

that differences in effect between the treatments become smaller with the inclusion of observational evidence.

As before, it is possible to adjust for overprecision and overestimation. Adjusting for overprecision of observational data down-weights the impact of its information on the results, see figure 6.8. The results show how overall results shift towards the RCT only

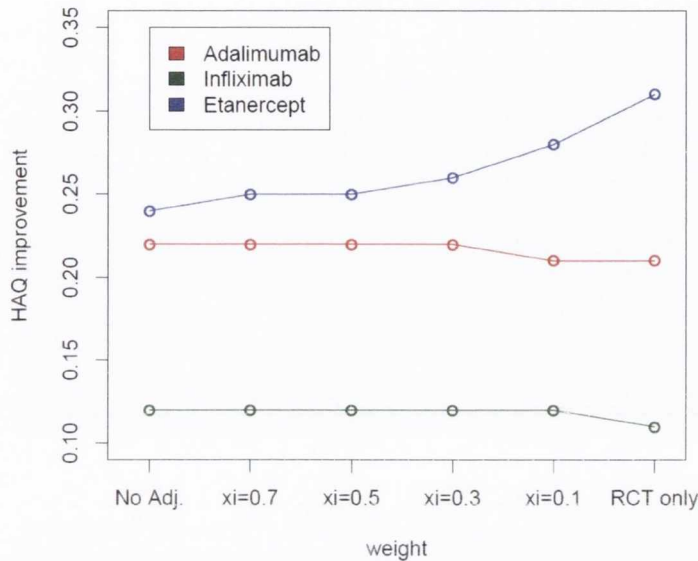


Fig. 6.8: RA case study: Adjusting for overprecision in 3-level model applying different values of w_{ξ} : Estimates shift towards the RCT only results as weight on observational evidence decreases.

results when decreasing the weight on observational data; the same effect that has been found when using observational data as prior information. When adjusting for under- and overestimation the effect on overall estimates is small, see figure 6.9. However, for the etanercept effect the results show an increase when adjusting for underestimation and a decrease when adjusting for overestimation of the treatment effect.

6.4.5 Sensitivity Analysis

A challenge is given by open label extensions. Such trials are typically one-armed and cannot be included in the analysis without adjustment. In this case study a matching approach is taken, where baseline characteristics are compared across available trials to identify a suitable match (d’Agostino and d’Agostino (2007)). However, such methods do not control for unobserved variables that may also affect the outcome. In a sensitivity analysis the one armed trial Klareskog et al. (2006) was excluded from the analysis; results based on the three methods are summarized in table 6.6. While the influence

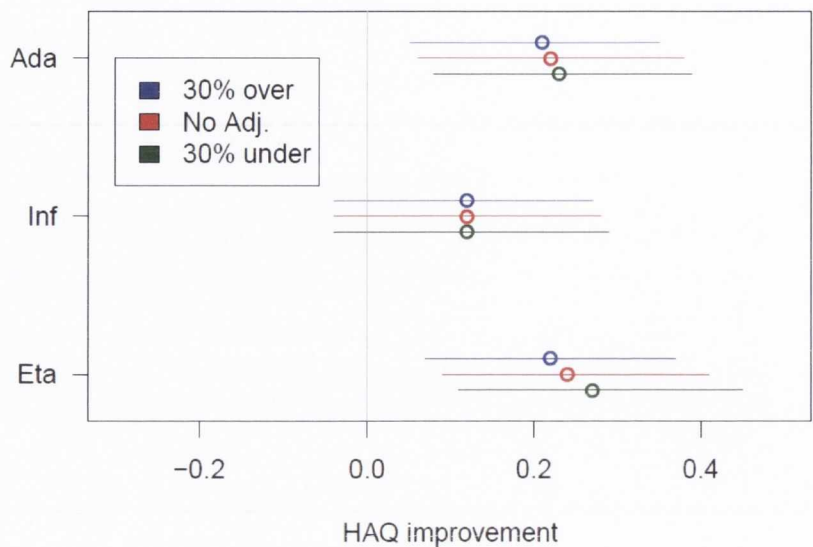


Fig. 6.9: RA case study: Adjusting for a 30% over- and 30% underestimation of the treatment effect in a 3-level model.

Table 6.6: RA case study: Results excluding Klareskog trial: Difference in % HAQ improvement for each biologic agent versus placebo (standard deviation). Results are shown for including observational evidence using naive pooling (NP), as prior information (Prior) and in a 3-level hierarchical model (3-level). Ada = adalimumab, Inf = infliximab, Eta = etanercept, Gol = golimumab, Cert = certolizumab, P = placebo.

<i>Comparison (A vs B)</i>	<i>NP</i>	<i>Prior</i>	<i>3-level</i>
Ada vs P	0.22 (0.03)	0.22 (0.03)	0.23 (0.09)
Inf vs P	0.13 (0.03)	0.12 (0.04)	0.13 (0.09)
Eta vs P	0.22 (0.04)	0.25 (0.04)	0.26 (0.09)
Gol vs P	0.21 (0.06)	0.21 (0.06)	0.22 (0.05)
Cert vs P	0.26 (0.03)	0.26 (0.03)	0.26 (0.03)

of observational data is slightly lower due to the fact that only 2 additional trials are included in the analysis, the exclusion of Klareskog et al. (2006) did not influence the results much. The main effect remains the drop in efficacy of the etanercept estimate. This is because the strength of the observational evidence lies in the relative efficacy between the anti-TNF agents. Adalimumab is doing better than etanercept in both 3-armed observational trials (Bazzani et al. (2009); Kievit et al. (2008)) (when considering baseline HAQ). The efficacy of etanercept changes more than the efficacy of adalimumab when including observational evidence, because the evidence for adalimumab in the RCT trials is stronger (5 trials) than the evidence for etanercept (2 trials). Welton et al. (2009) have discussed the implications of including or excluding particular evidence sources in more detail.

6.5 Results: Application in Multiple Sclerosis

6.5.1 Data

A systematic literature review following the PRISMA method identified published studies up to and including May 2012 meeting the inclusion criteria. The analysis included RCTs or observational studies with a patient populations with relapsing remitting MS; studies which included >10% progressive patients were excluded. Patients must be treated with one of the nine DMTs identified in section 6.2.2 compared to either each other or to placebo. Trials which included DMTs as part of combination therapy were excluded. All trials were required to have included a comparison of two interventions (or one intervention and placebo); i.e. single-arm trials were excluded.

The chosen outcome measure was the ARR; results are reported on the log scale. The total number of patients, the number of patient years and the number of relapses were recorded for each trial arm. Where number of patient years or number of relapses were not available, the number of patients, the timepoint of the analysis and the ARR were used to estimate these. To test for potential confounding factors, the mean age, gender, duration of disease, number of relapses in the past 2 years (estimated based on 1 year data where 2 year data was not available) and the EDSS score to measure baseline disease activity were also recorded.

The following databases were searched from inception to May 2012 for relevant studies: EMBASE, MEDLINE (via PubMed) and the Cochrane Central Register of Controlled Trials (CENTRAL). Figure 6.10 shows the PRISMA flowchart for the RCT selection process. The reference lists of published reviews of observational studies in MS were scanned

to identify observational studies meeting the inclusion criteria. The search identified sixteen randomized controlled trials and seven observational trials. The RCT data including

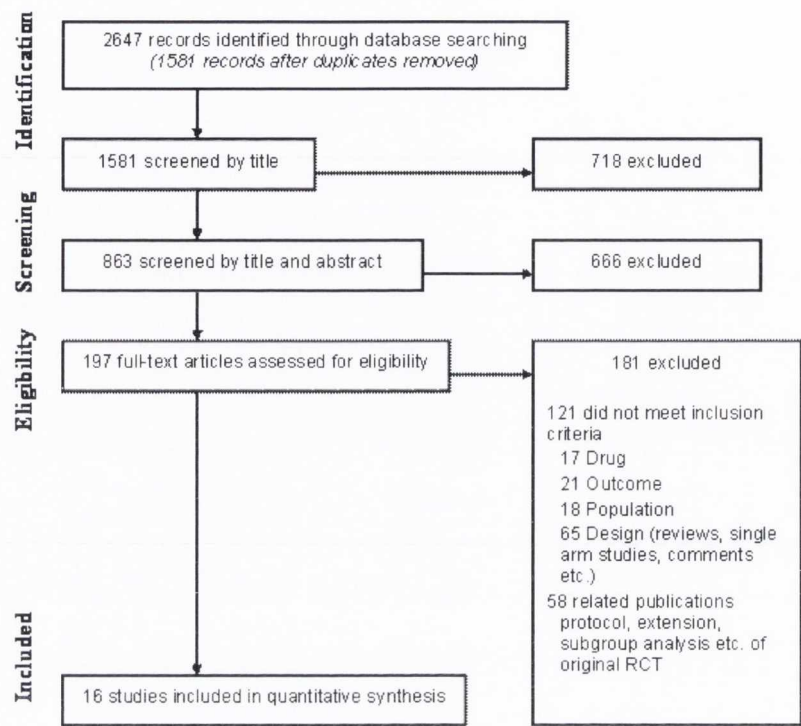


Fig. 6.10: MS case study: Flow Diagram for DMT therapy in MS.

baseline characteristics is summarized in table 6.7, figure 6.11 shows the corresponding evidence network.

Table 6.7: MS case study: Data extracted from RCTs: Number of patients N; number of relapses (rel); number of patient-years (py); age (years), sex (percentage of females), dod (duration of disease), mRel (mean number of relapses within last 2 years), EDSS (EDSS score at baseline)

Trial	Arm	N	rel	py	age	sex	dod	mRel	EDSS
Polman et al. (2006)	Placebo	315	433	593	36.1	0.70	5.3	2.32	2.30
	Natalizumab	627	276	1202					
Kappos et al. (2010)	Placebo	418	300	750	37.1	0.70	8.2	2.13	2.40
	Fingolimod 1.25mg	429	122	762					
	Fingolimod 0.5mg	425	143	794					
Cohen et al. (2010)	Placebo	431	135	409	36.2	0.67	7.4	2.27	2.21
	Fingolimod 1.25mg	426	79	395					
	Fingolimod 0.5mg	429	66	414					
Saida et al. (2012)	Placebo	57	27	27	35.3	0.69	7.8	2.43	2.07
	Fingolimod 0.5mg	57	13	26					
	Fingolimod 1.25mg	57	11	26					
O'Connor et al. (2011)	Placebo	363	365	676	37.9	0.72	8.7	2.23	2.68
	Teriflunomide 7mg	365	253	684					
	Teriflunomide 14mg	358	245	663					
Comi et al. (2012)	Placebo	556	375	962	38.7	0.69	8.7	1.90	2.60
	Laquinimod	550	293	977					
PRISMS Group (1998)	Placebo	187	479	364	35.0	0.69	5.7	3.00	2.47
	Rebif® 22	189	344	366					

Continued on next page

Table 6.7 – continued from previous page

Trial	Arm	N	rel	py	age	sex	dod	mRel	EDSS
Johnson et al. (1995)	Rebif [®] 44	184	318	363					
	Placebo	126	210	250	34.0	0.73	6.9	2.90	2.60
	GA	125	161	273					
Comi et al. (2001)	Placebo	120	91	75	34.0	-	8.1	2.65	2.35
	GA	119	61	75					
Jacobs et al. (1996)	Placebo	143	225	274	36.8	0.74	6.5	1.84	2.35
	Avonex [®]	158	196	293					
Durelli et al. (2002)	Avonex [®]	92	126	180	36.9	0.66	6.3	2.90	1.97
	Betaferon [®]	96	95	190					
Panitch et al. (2002)	Avonex [®]	338	216	304	37.9	0.75	6.6	2.60	2.30
	Rebif 44 [®]	339	183	305					
IFNB Group (1993)	Placebo	123	266	209	35.6	0.70	4.3	3.50	2.90
	Betaferon [®]	124	173	206					
O'Connor et al. (2009)	GA	448	383	1126	35.6	0.69	5.2	2.45	2.33
	Betaferon [®]	897	828	2299					
Cadavid et al. (2009)	GA	39	23	70	36.0	0.69	1.1	2.83	2.00
	Betaferon [®]	36	25	68					
Mikol et al. (2008)	GA	386	200	688	36.8	0.71	6.2	-	2.34
	Rebif [®] 44	387	207	689					

The demographics of the trials are broadly similar, however some differences were found, particularly in relation to the disease duration, ranging from 1.1 - 8.7 years, and EDSS score at baseline, ranging from 1.97 - 2.90.

The literature review also identified seven observational trials providing evidence of relative efficacy of older DMTs versus each other and versus placebo. Due to concerns regarding selection bias in observational trials, baseline characteristics within each study were tested for differences between the study arms. Two of the trials (Limmroth et al. (2007); Minagar and Muray (2008)) were excluded from the analysis due to missing baseline characteristics. Rio et al. (2007) and the placebo arm of Carra et al. (2003) were excluded, because differences in baseline characteristics could not be precluded. The data extracted from the four trials included in the analysis can be found in table 6.8, the corresponding evidence network is displayed in figure 6.12.

6.5.2 Computation / Convergence

All models in this section are fitted in WinBUGs. 200,000 burn-in iterations were discarded, the analysis is based on 2 chains with 500,000 iterations and a thinning interval of 100. Convergence for this case study was analysed using the same methodology as before. The effective sample size was checked, the autocorrelation plots showed a decrease with increasing lags and two chains were run and indicated good mixing.

Table 6.8: MS case study: Data extracted from observational trials: Number of patients N, number of relapses rel, number of patient years py, age (years), sex (percentage of females), dod (duration of disease), mRel (mean number of relapses within last 2 years), EDSS (EDSS score at baseline).

Trial	Arm	N	rel	py	age	sex	dod	mRel	EDSS
Haas and Firzlaff (2005)	Avonex [®]	79	109	158	37	0.78	7.7	1.09	2.20
	Betaferon [®]	77	123	154	37	0.70	6.4	1.20	2.28
	GA	79	56	158	35	0.77	7.9	1.16	1.98
	Rebif [®] 22	48	59	96	37	0.65	9.8	1.15	2.36
Khan et al. (2001)	Placebo	15	23	23	32	0.67	4.1	1.03	2.63
	Avonex [®]	34	41	51	32	0.59	4.9	1.18	2.71
	Betaferon [®]	34	28	51	33	0.62	4.2	1.22	2.60
	GA	39	29	59	32	0.64	4.4	1.14	2.64
Trojano et al. (2003)	Betaferon [®]	209	136	418	31	0.68	6.5	1.35	2.50
	Avonex [®]	169	120	338	32	0.72	7.1	1.25	2.40
Carra et al. (2003)	Avonex [®]	26	14	35	41	0.62	6.5	0.77	2.02
	Rebif [®] 44	20	12	27	38	0.60	6.4	1.28	2.08
	Betaferon [®]	20	11	27	47	0.70	8.0	1.13	3.13
	GA	30	8	40	39	0.70	7.4	1.02	2.45

6.5.3 RCT only

To provide a reference point for measuring the impact of observational information in the alternative methodologies, in a first step the RCT data is analysed on its own using the model described in equation 6.3. The model was extended to a meta-regression to test potential confounding factors age, gender, duration of disease, relapses within the past two years and EDSS score at baseline. None of these was found to have a significant impact on the result and covariates were therefore not included in the final analysis. The relative efficacy estimates of all DMTs versus placebo with their corresponding standard deviation can be found in table 6.9. Figure 6.13 plots the 95% credible intervals for all treatments against each other. All treatments provide significant improvent in ARR compared to placebo. The MTC confirms differences between many of the treatments. Two of the new treatments, natalizumab and fingolimod (both doses), are doing signif- icantly better than all other treatments. Teriflunomide and laquinimod, the other new agents, show a similar efficacy as the older treatments. Avonex[®] appears to have the worst efficacy. In the following three sections the degree to which observational evidence agrees and strenghtens or disagrees with these findings is analysed.

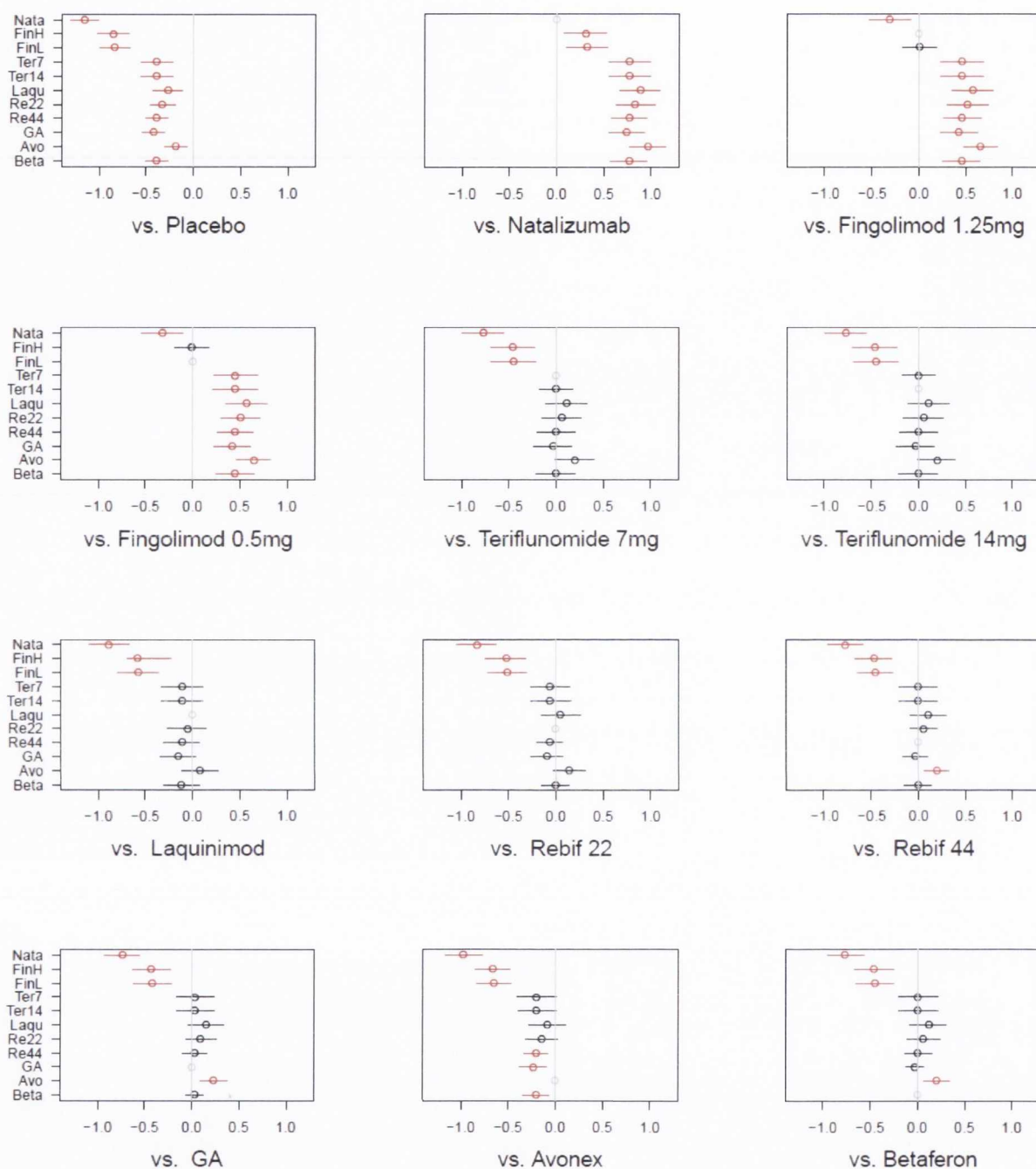


Fig. 6.13: MS case study: Forest Plot based on RCT evidence: Plots pair-wise logARRs of DMTs versus placebo and each other. Nata=natlizumab, FinH=fingolimod 1.25mg, FinL=fingolimod 0.5mg, Ter7=teriflunomide 7mg, Ter14=teriflunomide 14mg, Laqu=laquinimod, Re22=Rebif[®] 22, Re44=Rebif[®] 44, Avo=Avonex[®], Beta=Betaferon[®]

Table 6.9: MS case study: Results based on RCT only and Naive Pooling with observational evidence: Mean relative log ARR estimate (standard deviation) of all DMTs versus placebo.

Comparison	RCT only	Naive Pooling	Inconsistency Model
Natalizumab vs Placebo	-1.16 (0.08)	-1.16 (0.08)	-1.16 (0.08)
Fingolimod 1.25mg vs Placebo	-0.85 (0.09)	-0.85 (0.08)	-0.92 (0.10)
Fingolimod 0.5mg vs Placebo	-0.84 (0.08)	-0.84 (0.08)	-0.80 (0.10)
Teriflunomide 7mg vs Placebo	-0.39 (0.08)	-0.39 (0.08)	-0.39 (0.08)
Teriflunomide 14mg vs Placebo	-0.39 (0.08)	-0.39 (0.08)	-0.39 (0.08)
Laquinimod vs Placebo	-0.27 (0.08)	-0.27 (0.08)	-0.27 (0.08)
Rebif [®] 22 vs Placebo	-0.33 (0.07)	-0.32 (0.06)	-0.33 (0.07)
Rebif [®] 44 vs Placebo	-0.39 (0.06)	-0.39 (0.06)	-0.40 (0.07)
GA vs Placebo	-0.42 (0.06)	-0.47 (0.05)	-0.40 (0.08)
Avonex [®] vs Placebo	-0.19 (0.06)	-0.19 (0.05)	-0.17 (0.09)
Betaferon [®] vs Placebo	-0.39 (0.06)	-0.34 (0.06)	-0.42 (0.09)
DIC ($\overline{D}, p_D$)	312.0 (285.4, 26.6)	424.5 (394.1, 30.4)	415.9 (377.6, 38.3)

6.5.4 Naive Pooling

The results from pooling RCT data and observational data are summarised in table 6.9. Changes are only expected for comparisons including Rebif[®] 22, Rebif[®] 44, GA, Avonex[®] and Betaferon[®], since no additional observational information was found for the newer treatments. Figure 6.14 plots the relative efficacies for these treatments. Little difference between naive pooling (blue) and RCT only (black) results is observed; the additional information from observational trials confirms the findings from the RCT analysis and strengthens the evidence by slightly decreasing the variance for the comparisons. The biggest change is found for GA and Betaferon[®], the efficacy improves by 0.05 for GA and decreases by 0.05 for Betaferon[®] when including additional information from observational trials.

As in the RA application, a model relaxing the consistency assumption is fit to assess inconsistencies in the network. Figure 6.15 plots the posterior mean deviance of each data point based on the model relaxing the consistency assumption against the original model. For most datapoints the contribution to the deviance is similar. However, there are 2 outliers, which refer to the Betaferon[®] and the GA arms of the observational trial by Haas and Firzlaff (2005). This indicates an inconsistency with this trial in the network and is supported by the DIC (table 6.9), which indicates an improved fit for the inconsistency model. Looking at the data, the annual relapse rate of Betaferon[®] in this trial is particularly high and the annual relapse rate of GA is very low. The inconsistency

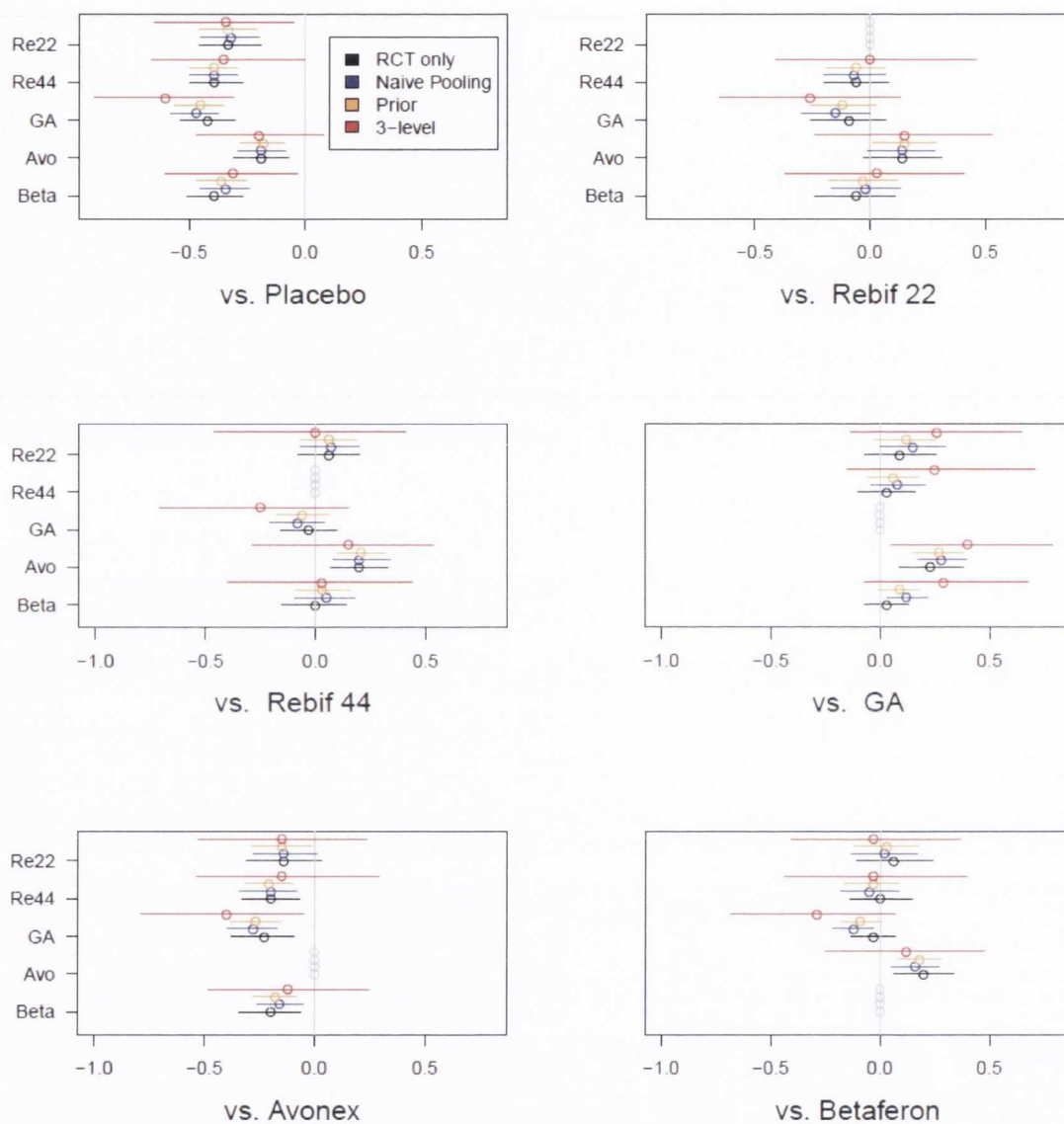


Fig. 6.14: MS case study: Forest Plot of results based on RCTs only, naive pooling, as prior information and 3-level model: Plots pairwise logARRs of DMTs versus placebo and each other. Re22=Rebif[®] 22, Re44=Rebif[®] 44, Avo=Avonex[®], Beta=Betaferon[®]

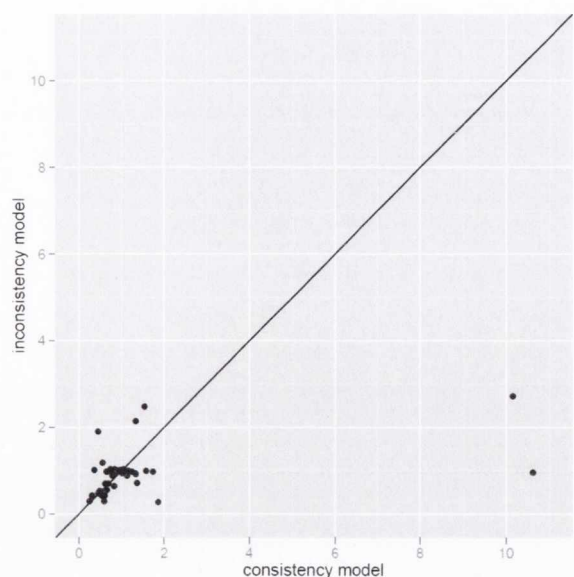


Fig. 6.15: MS case study: Plot of the individual data points’ posterior mean deviance contribution for the original model (horizontal axis) and the model relaxing the consistency assumption (vertical axis).

Table 6.10: MS case study: Node Split Results: Mean difference in log ARR based on direct and indirect evidence (standard deviation). DIC: Deviance Information Criterion.

	GA vs P	GA vs Avonex [®]	Betaferon [®] vs P	Betaferon [®] vs Avonex [®]
DIC ($\overline{D}, p_D$)	424.3 (393.4,30.9)	407.9 (376.5,31.4)	425.2 (393.8,31.4)	421.3 (389.9,31.4)
direct	-0.40 (0.08)	-0.70 (0.12)	-0.42 (0.09)	-0.06 (0.07)
indirect	-0.52 (0.07)	-0.17 (0.07)	-0.30 (0.06)	-0.27 (0.08)

caused by this issue is solved when relaxing the consistency assumption. To investigate further, the node split approach is applied to the estimates of GA and Betaferon[®] versus placebo and versus Avonex[®] (which is the baseline treatment in the trial). Results are summarized in table 6.10. Splitting GA versus Avonex[®] and Betaferon[®] versus Avonex[®] into estimates based on direct and indirect evidence improves the model fit, which confirms the inconsistency detected in the above model. The two trial arms are not excluded from the remainder of this chapter; however, to investigate the impact on the overall analysis, they are excluded in a sensitivity analysis in section 6.5.7.

6.5.5 As Prior Information

Using observational evidence to inform the prior information requires the analysis of observational trials separately as a first step, results are summarized in table 6.11. GA is confirmed as the strongest of the older DMTs followed by Rebif[®] 22. Similar efficacies are

Table 6.11: MS case study: Results based on OBS data only: Mean difference in log ARR of each DMT versus placebo.

Comparison	OBS estimate
Rebif [®] 22 vs Placebo	-0.5 (0.25)
Rebif [®] 44 vs Placebo	-0.26 (0.43)
GA vs Placebo	-0.97 (0.25)
Avonex [®] vs Placebo	-0.31 (0.24)
Betaferon [®] vs Placebo	-0.34 (0.24)

observed for the remaining agents. While the absolute values differ from those estimated based on RCTs, it is important to consider the high standard deviation and the weak link to placebo. To make use of this analysis to inform the prior information of the RCT model, Avonex[®] is chosen as the basic treatment, since it is one of the strongest links in the network. Betaferon[®] is also part of all observational trials and connected to all other treatments in the evidence network and therefore provides an alternative choice. The MTC of observational data alone yields estimates for all pairwise comparisons. Mean and precision of the posterior distributions of the relative efficacy of each treatment compared to avonex are used as inputs for the basic parameters in the RCT model:

$$\begin{aligned}
 a[1] &\sim N(0.31, 16.7) \\
 a[3] &\sim N(-0.67, 59.2) \\
 a[4] &\sim N(-0.19, 39.1) \\
 a[5] &\sim N(0.05, 7.7) \\
 a[6] &\sim N(-0.03, 144)
 \end{aligned}
 \tag{6.10}$$

$a[1]$ represents the relative efficacy of placebo versus Avonex[®], $a[3]$ the efficacy of GA versus Avonex[®], $a[4]$ the efficacy of Rebif[®]22 versus Avonex[®], $a[5]$ represents Rebif[®]44 versus Avonex[®] and the efficacy of Betaferon[®] versus Avonex[®] is given by $a[6]$. $a[2]$, the relative efficacy of avonex with itself is not a random node, it is fixed at 0.

Results are summarized in table 6.12; figure 6.14 plots the relative efficacy of the agents against each other. The results based on using observational data to inform the prior distribution are very similar to those based on naive pooling. The standard deviation decreases slightly compared to the results based on RCTs only and an increase of efficacy is observed for GA, while the efficacy of Betaferon[®] decreases.

As before, it is possible to reduce the weight given to observational data by inflating the variance parameter as described in equation 6.9. Figure 6.16 illustrates the effects for a range of different values of w_{ξ} . Estimates of the new DMTs are not affected, since there is no additional information provided. The results show the same effect as in the

Table 6.12: MS case study: Results based on OBS data as prior information: Difference in log ARR of each DMT versus placebo (standard deviation). Betaf=Betaferon[®]; DIC=Deviance Information Criterium.

Comparison	OBS as prior
Rebif [®] 22 vs Placebo	-0.33(0.06)
Rebif [®] 44 vs Placebo	-0.39(0.05)
GA vs Placebo	-0.45(0.05)
Avonex [®] vs Placebo	-0.18(0.05)
Betaf vs Placebo	-0.36(0.05)
DIC ($\overline{D}, p_D$)	311.8 (286.0,25.8)

RA application: decreasing the weight on observational information shifts the efficacy estimates towards the estimates based on RCTs only. The effect is most evident for GA and Betaferon[®], where the largest difference between estimates is observed. Figure

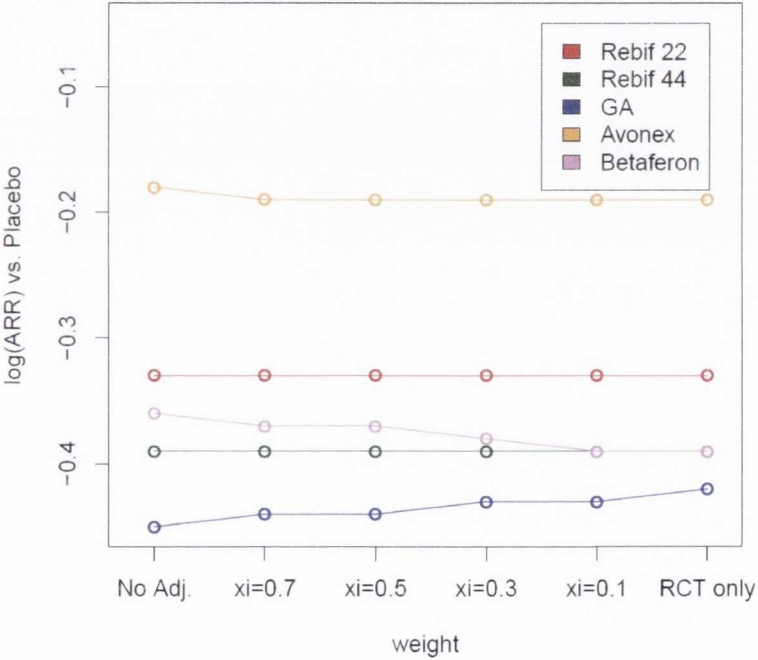


Fig. 6.16: MS case study: Adjusting for overprecision for different values of w_{ξ} using OBS data as prior information: Estimates shift towards the RCT only results as weight on observational evidence decreases.

6.17 shows the effect of adjusting for a 30% over- and underestimation of the treatment effect in observational trials on the overall effectiveness. As in the previous case study, the effect is small, since the strength of the observational evidence lies in the relative efficacy of the older DMTs rather than in the link to placebo.

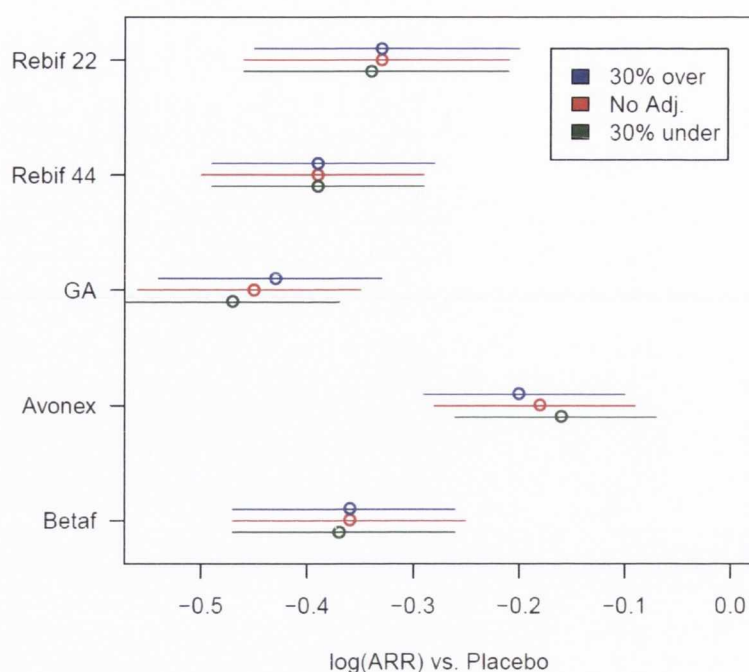


Fig. 6.17: MS case study: Effects of adjusting for a 30% over- and 30% underestimation of the treatment effect when using OBS data as prior information

Table 6.13: MS case study: 3-level model results: Difference in efficacy for each DMT versus placebo (standard deviation) on study type level and overall level.

Comparison	Overall	RCT level	OBS level
Rebif [®] 22 vs Placebo	-0.34 (0.15)	-0.33 (0.07)	-0.36 (0.15)
Rebif [®] 44 vs Placebo	-0.35 (0.17)	-0.39 (0.06)	-0.30 (0.21)
GA vs Placebo	-0.60 (0.15)	-0.44 (0.06)	-0.77 (0.14)
Avonex [®] vs Placebo	-0.20 (0.14)	-0.19 (0.06)	-0.20 (0.12)
Betaferon [®] vs Placebo	-0.31 (0.14)	-0.39 (0.06)	-0.24 (0.12)
DIC ($\bar{D}, p_D$)	410.9 (377.1, 33.8)		

6.5.6 Hierarchical Model

In this section a three level hierarchical model is fit to the MS data. Such a model yields estimates of overall efficacy for each treatment as well as estimates on study type level; results for this application are summarized in table 6.13. The overall efficacy is estimated to lie between the study type level estimates. The increased variance compared to naive pooling is due to the fact that the model takes into account between trial heterogeneity. Estimates on observational level are surrounded by higher uncertainty than estimates on RCT level, since there is less information available. Again, it is possible to adjust for overprecision, the results for different weights given to observational evidence are shown in figure 6.18. As weight on observational data decreases, the estimates shift

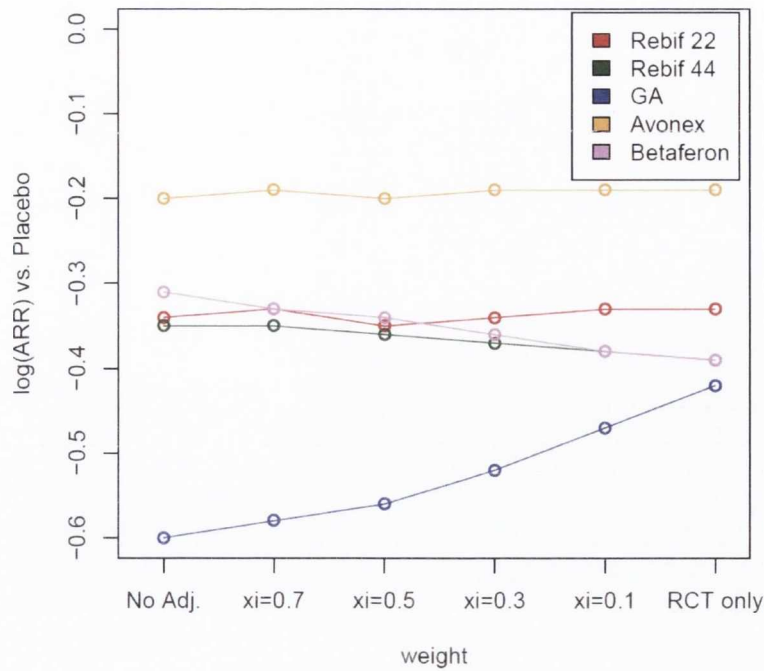


Fig. 6.18: MS case study: Effects of adjusting for overprecision for different values of w_ξ in 3 level model: Estimates shift towards the RCT only results as weight on observational evidence decreases.

towards the RCT only results. This is most visible for the GA and Betaferon[®] estimate. Adjusting for a 30% over- and underestimation of the treatment effect does not have a huge impact on the results, see figure 6.19. As in the RA case study, the 3-level model in the MS analysis shows how data is shrunk towards the overall mean, when analysed in a hierarchical model. Figure 6.20 illustrates the effect for GA versus Placebo, where this is most evident. On study type level, results are drawn further in the direction of the RCT estimate, since RCT evidence is stronger than OBS evidence.

6.5.7 Sensitivity Analysis

In section 6.5.4 the GA and the Betaferon[®] arm of trial Haas and Firzlaff (2005) were identified to potentially cause inconsistency. To investigate this issue further, a sensitivity analysis excluding these two arms was conducted. Results based on the three methods are summarized in table 6.14. Excluding these two datapoints improves the model fit in all three methods. In the main analysis, including observational data improved the efficacy of GA and worsened the efficacy of Betaferon[®]. This effect has been removed by excluding these trial arms. It appears, that the remaining evidence from observational trials agrees with evidence from the RCT data. This is mirrored in the decrease in standard deviation of the 3-level model estimates compared to before. The estimates

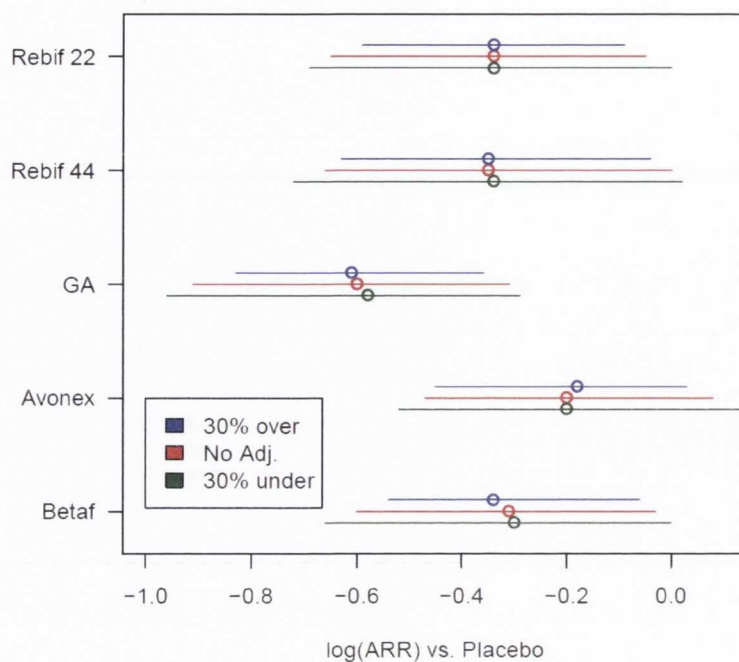


Fig. 6.19: MS case study: Effects of adjusting for a 30% over- and 30% underestimation in the 3-level model in the MS application

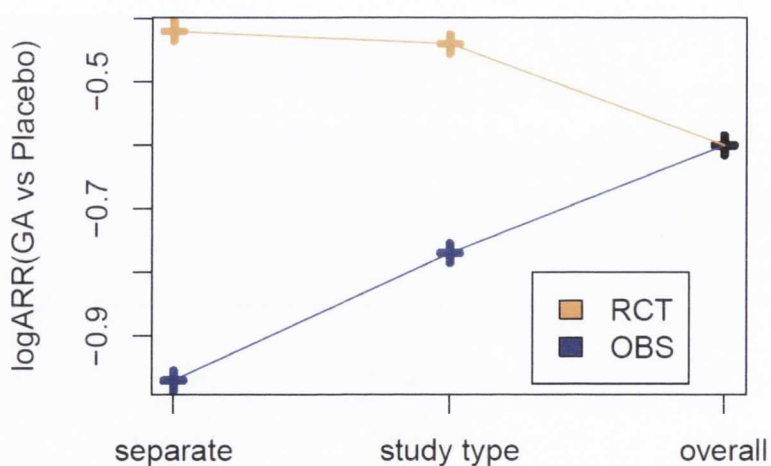


Fig. 6.20: MS case study: Effect of shrinkage

Table 6.14: MS case study: Results excluding GA and Betaferon[®] arm of Haas and Firzlaff (2005): Mean relative log ARR estimate (standard deviation) based on each of the three methods. DIC=Deviance information criterium; NP=Naive Pooling.

<i>Comparison (A vs B)</i>	<i>NP</i>	<i>Prior</i>	<i>3-level</i>
Rebif [®] 22 vs Placebo	-0.32 (0.06)	-0.32 (0.06)	-0.33 (0.12)
Rebif [®] 44 vs Placebo	-0.39 (0.06)	-0.39 (0.05)	-0.37 (0.13)
GA vs Placebo	-0.44 (0.06)	-0.44 (0.06)	-0.49 (0.13)
Avonex [®] vs Placebo	-0.20 (0.05)	-0.19 (0.05)	-0.18 (0.11)
Betaferon [®] vs Placebo	-0.39 (0.06)	-0.39 (0.06)	-0.40 (0.12)
DIC ($\bar{D}, p_D$)	389.4 (358.9,30.5)	310.8 (258.8,26.0)	391.9 (259.6,32.3)

based on the hierarchical model become more conservative when estimates based on the alternative designs differ. Differences in baseline demographics may have lead to the Betaferon[®] group in the Haas study having a more severely ill population than the other treatment groups. The progression rate in the Betaferon[®] group was significantly higher than the other groups at baseline due to a high proportion of patients with high EDSS scores despite short disease duration. In contrast to the other included observational studies, a proportion of patients in all treatment arms of the Haas study had received prior DMT (24%-63%). The proportion of previously treated patients was much higher in the Rebif[®] group (63%) compared to the other treatment arms (~25%). These differences may have contributed to the dichotomous results observed in the Haas study compared to other comparative studies.

6.6 Conclusions / Limitations

The above techniques provide a framework for systematically including evidence from different trial designs in a MTC model.

Naive pooling makes the strong assumption of no differences between trial designs. The method does not allow for bias adjustments and no additional uncertainty is taken into account. Naive pooling can therefore not be recommended as a primary means of combining evidence from different trial designs. However, as demonstrated in the case study on RA, in some situations naive pooling may create the opportunity to estimate consistencies within the evidence network. In such situations naive pooling may be worthwhile as a first step analysis.

Summarising observational evidence to inform the prior distribution in a MTC model allows the adjustment for potential bias due to overprecision or overestimation. This is a clear advantage over naive pooling; however, between trial design heterogeneity is not

taken into account and it is not possible to extend the model to include more than two different trial designs.

The hierarchical modelling approach accounts for the uncertainty arising from combining information from different trial designs using random effects. The hierarchy levels allow the quantification of the impact evidence from different designs has on the result while adjusting for potential bias. These advantages make hierarchical modelling the preferred method of including information from different designs.

For the above case studies trials were grouped into two designs; RCTs and observational. This makes it difficult to estimate between trial design heterogeneity in this approach and results are very sensitive to the prior distribution. This explains the large increase in standard deviation for the 3-level model in both applications. This is not an issue when including more than 2 different trial designs.

Observational data is prone to two different biases; overestimation and overprecision of the treatment effect. Including observational data as prior information and in the form of a hierarchical model allows for bias adjustment. In the literature assuming 30% overprecision has previously been used (O'Rourke et al. (2007); Lilford and Braunholtz (1996)). Research has not found a consistent overestimation of the treatment effect among observational trials (Benson and Hartz (2000); Concato et al. (2000)). In any case, the actual size of the bias is difficult to estimate and it is important to vary this in a sensitivity analysis. Turner et al. (2009) provide a detailed approach to modelling bias in evidence synthesis. He differentiates between rigour (lack of internal bias) and relevance (lack of external bias). Bias elicitation is conducted on individual trials by comparing the actual study with an idealized study protocol defined according to the question of interest. A direct form is assumed for the bias in the target parameter.

A challenge is given by open label extensions. Such trials are typically one-armed and cannot be included in the analysis without adjustment. In the first case study a matching approach was taken, where baseline characteristics are compared across available trials to identify a suitable match. However, such methods do not control for unobserved variables that may also affect the outcome. It is important to determine the impact of one-armed trials in a sensitivity analysis.

For many disease areas there is a lot of observational data available providing additional information on treatment effectiveness. It is important for an informed decision making process to include all available evidence. Including observational data at basecase analysis or in the form of a sensitivity analysis can greatly improve evidence synthesis as part of economic assessments, as well as choice of agent in a clinical setting. The

methods described here provide a flexible methodology to analyse the impact of such additional information. A hierarchical model is the preferred method to including information from different designs, because it is the only method taking into account between design heterogeneity.

Chapter 7

MCDA

7.1 Motivation

The objective of making decisions regarding the reimbursement of healthcare technologies is to maximise the health benefit in the population given a limited healthcare budget. As described in chapter 2 this involves satisfying a number of criteria, some of which are conflicting.

Currently, the decision process is based on the INB as utility function, which was defined in chapter 2. The INB combines incremental costs and QALYs; an intervention is recommended for reimbursement when the INB is positive.

$$\begin{array}{ll} INB(K) > 0 & \rightarrow \text{Reimburse treatment.} \\ \textit{else} & \rightarrow \text{Do not reimburse treatment.} \end{array} \quad (7.1)$$

K refers to the threshold of what one is willing to pay for an additional QALY, also defined in chapter 2. Unfortunately, the QALY does not measure all aspects of health benefit, see Garau et al. (2011). For instance, it assumes an additional year of life at perfect health to have the same value at the end of someone's life as it has in the middle of someone's life. To include issues which are not covered by the decision rule in equation 7.1, a dynamic threshold is applied on a case by case basis. The threshold is not clearly defined. However, results are currently reported to the HSE at a threshold of €20,000 and at €45,000. QALYs are known to be particularly insensitive for cancer treatments. As a way of dealing with this issue the National Cancer Control Programme (NCCP) review group reassesses cancer treatments in light of cancer specific criteria and issues a second recommendation to the HSE. An example illustrating the dynamics of the threshold in the other direction is the assessment of roflumilast for the treatment of COPD (chronic obstructive pulmonary disease) (NCPE (2010f)). At an ICER of only €1,541 per QALY roflumilast was not recommended for reimbursement due to high

uncertainty in the estimates and low quality of evidence.

Applying separate decision models for different disease areas and a case by case approach to include additional objectives yields non-transparent and inconsistent decision making. A multi criteria decision analysis (MCDA) approach for the Irish healthcare setting will formalise the process and improve transparency, consistency and coherence. For this case the INB as utility function is generalized to include multiple criteria, as defined in chapter 3. This changes the decision rule to:

$$\begin{aligned} \sum_{k=1}^p w_k \cdot C_k > 0 &\rightarrow \text{Reimburse treatment.} \\ \text{else} &\rightarrow \text{Do not reimburse treatment.} \end{aligned} \tag{7.2}$$

where $k = 1, \dots, p$ objectives are combined and w_k refers to the weight given to the k^{th} criterion C_k .

The approach enables the assessment of all disease areas within the same model. The approach is beneficial for the decision maker, because consistent decision making provides justification when decisions are challenged. The pharmaceutical industry will benefit, since a prediction of the reimbursement decision is possible at an earlier stage.

7.1.1 Aim of Analysis

In this chapter a descriptive approach is taken to identify and weight criteria based on guidelines and past decisions to inform a multi attribute value modelling approach to MCDA. The work identifies criteria, which have influenced reimbursement decisions in the past as well as their relative importance. This is work towards a normative approach, which can be implemented in the Irish healthcare system.

7.2 Methods

7.2.1 Data Collection

A set of criteria with potential influence on the reimbursement decision was selected based on HIQA guidelines, NCPE assessments and NICE decisions. According to HIQA, clinical effectiveness, cost-effectiveness, budget impact as well as social, ethical, medicolegal and organisational aspects associated with the use of a technology are part of a HTA. HIQA has published guidelines on clinical effectiveness, cost-effectiveness and budget impact; guiding information on the remaining criteria have yet to be established. In the absence of further guidelines from HIQA the list of criteria was expanded by criteria, which are known to have influenced NICE decisions in the past. The set was complemented by objectives arising in the NCPE assessments.

Information was sought for every full pharmacoeconomic assessment conducted by the NCPE since 2006 on each of the identified criteria. The basis for the information are the assessment summaries, which are published on the NCPE website.

7.2.2 Statistical Analysis

The data is analysed using logistic regression. Logistic regression predicts the outcome of a categorical variable based on a number of predictor variables, which can be either categorical or continuous.

The outcome variable in this case is the recommendation for reimbursement, which is either yes (1) or no (0). The recommendation for reimbursement by the NCPE was chosen as the outcome of interest rather than the actual decision of reimbursement, because in some cases treatments were approved for reimbursement following price reduction. Unfortunately, no full assessment based on the new price is made in such cases and therefore information on cost-effectiveness and budget impact is not available.

The set of criteria identified in the data collection process are the predictor variables. Only cases where information on all criteria is available are included in the analysis.

A logistic regression is a special case of a generalized linear model and takes the following form:

$$\begin{aligned} Y_i &\sim \text{Bern}(p_i) \\ \text{logit}(p_i) &= \sum_{j=1}^m \beta_j X_{ij} \end{aligned} \tag{7.3}$$

where $i=1,\dots,n$ indicates the assessment, Y_i is the outcome variable indicating whether case i was recommended for reimbursement and X_{ij} is the response of case i to predictor j . For a Bayesian analysis prior distributions need to be assigned for the regression coefficients β_j . For this analysis vague priors were chosen:

$$\beta_j \sim N(0, 0.001).$$

The model is fitted in WinBUGs. An overview on generalized linear models can be found in Dobson (2002).

The criteria providing the best prediction for the outcome are selected using stepwise forward regression.

7.3 Results

7.3.1 Data

The criteria selected for the analysis are summarized in table 7.1. Clinical effectiveness, cost-effectiveness and budget impact are included based on HIQA guidelines. Rawlins

Table 7.1: MCDA: Details on the criteria selected for a descriptive MCDA approach for the Irish healthcare setting.

Criteria	Short Description	Type	Scale
Clinical effectiveness (Eff)	The ability of a technology to achieve a clinically significant impact on the health status of a patient.	categorical	high(1) medium(0) low(-1)
ICER	Cost per additional QALY	continuous	€per QALY
Budget Impact (BI)	Gross financial impact of a new technology on the healthcare system accumulated over 5 years.	continuous	€
Severity of illness (Sev)	Indicates the severity of the underlying disease. Society tends to prioritize the relief for a severe disease before a mild disease.	binary	yes(1) no(0)
End of life treatment (EoL)	Indicates life prolonging treatments. Society tends to place special value on the extension of life, as long as it is of reasonable quality.	binary	yes(1) no(0)
Stakeholder involvement (Stake)	Patients and their advocates can influence the decision of reimbursement through recommendations and assessments.	binary	yes(1) no(0)
Innovation (Innov)	Innovative treatments can be the first treatment for a disease, or significantly improve efficacy or reduce costs compared to alternative treatments.	binary	yes(1) no(0)
Children (Child)	The assessment of quality of life in children is methodologically challenging. Society tends to give the benefit of a doubt to sick children.	binary	yes(1) no(0)
Quality of Evidence (QoE)	The assessment can only be as good as the evidence provided; low quality of evidence may yield the rejection of treatments.	categorical	high(1) medium(0) low(-1)
Uncertainty (Unc)	Uncertainty in the analysis can result in high costs to the healthcare system.	binary	medium(0) low(-1)
Media coverage (Media)	Media coverage of a particular intervention can exert pressure on the decision maker.	binary	yes(1) no(0)
Year	Year of assessment. The economic climate has changed the reimbursement decisions over the last 6 years.	continuous	year

et al. (2010) identified six criteria, which have influenced reimbursement decisions in NICE in the past: Severity, End of Life, Stakeholder, Innovation, Disadvantaged Population and Children. These criteria are included in the list except for disadvantaged population, which did not apply to any of the past decisions made in Ireland since 2006. The list was completed through further analysis of NCPE evaluations. This added quality of evidence, level of uncertainty, media coverage and year of assessment to the list of criteria. Table 7.1 gives a brief description for each of the criteria, the type of variable used to measure and its scale. Each of 46 past reimbursement decisions is scored against each of these criteria. Information is extracted from the assessment summaries, which can be downloaded on the NCPE website. A summary of the data can be found in table 7.2. Where ICERs were reported against a number of comparators, the highest ICER was extracted. The accumulated 5 year gross budget impact is extracted; where a yearly budget impact was reported, the 5-fold value was used. Uncertainty was deemed high,

when concern about the level of uncertainty was expressed in the assessment summary. A more clear classification of uncertainty would be possible based on the EVPI, however, the EVPI is only reported for the more recent assessments. The scoring on all other criteria was assessed by members in the NCPE. Disagreements were resolved in discussions. Ten cases were excluded from the analysis due to missing values (7, 8, 9, 13, 17, 18, 19, 20, 22, 37). Case 11 was excluded because no ICER is available for cost minimisation analyses. The analysis is based on the 35 remaining cases. 17 of these interventions (48.5%) were recommended for reimbursement. Where the intervention dominated its comparator and no ICER is reported, an ICER of -1 is adopted as a conservative choice. For case 28 the cost per life year gained rather than per QALY is reported. The case is included in the primary analysis, but excluded in a sensitivity analysis.

The median ICER of all reimbursement decisions is €21,643 per QALY ranging from -1 to 1,111,111. The median ICER for positive recommendations is €11,411 per QALY, more than sixfold lower than the median ICER for negative recommendations, €72,653 per QALY. Also, the median budget impact of positive reimbursement decisions is less than half of the median budget impact for negative reimbursement decisions (€3,459,187 vs €8,909,503). Eleven (31%) of the treatments are given the status of innovation; seven of these (64%) were recommended for reimbursement. 14 of the indications relate to a severe disease, four of these (29%) were recommended for reimbursement. In five cases the treatment is life prolonging; two of these (40%) had a positive reimbursement decision. Only one of the ten treatments (10%) with low quality of evidence was recommended for reimbursement. Five of 15 cases with high uncertainty (33%) were recommended for reimbursement. None of the cases which got media attention were recommended for reimbursement.

Table 7.2: MCDA: Data extracted from assessment summaries for MCDA analysis.*cost per LYG; CMA
= Cost Minimization Analysis

Case	Drug	Decision	Eff	ICER	BI	Sev	EqL	Stake	Innov	Child	QoE	Unc	Media	Year
1	Inhaled Insulin (NCPE (2006a))	1	0	44526	55000000	0	0	0	1	0	0	0	0	2006
2	Rimonabant (NCPE (2006b))	1	0	23774	5000000	0	0	0	1	0	0	1	0	2006
3	Sunitinib (NCPE (2006c))	1	0	57280	2300000	1	1	0	1	0	0	0	0	2006
4	Subuxone (NCPE (2007d))	0	0	332207	15500000	0	0	0	0	0	-1	1	0	2007
5	Grazax (NCPE (2007e))	0	0	51637	30000000	0	0	0	1	0	-1	1	0	2007
6	Natalizumab (NCPE (2007c))	1	0	dominant	16000000	1	0	0	1	0	0	1	0	2007
7	Hep B Vaccine (NCPE (2007a))	1	0	37018*	NA	0	0	1	0	0	0	0	0	2007
8	Pneumococcal Vaccine (NCPE (2007b))	1	0	5913	NA	0	0	0	0	0	0	0	0	2007
9	Rivaroxaban (NCPE (2008b))	1	0	dominant	NA	0	0	0	0	0	0	0	0	2008
10	Dabigatran (NCPE (2008c))	1	0	dominant	-19538	0	0	0	0	0	0	0	0	2008
11	Melatonin (NCPE (2008e))	0	0	CMA	1651785	0	0	0	0	0	-1	1	0	2008
12	HPV vaccination (HIQA (2008))	1	0	17383	77700000	0	0	0	0	0	0	0	0	2008
13	Nilotinib (NCPE (2008d))	0	0	77463	NA	1	0	0	0	0	0	0	0	2008
14	Lapatinib (NCPE (2008a))	1	0	12230	2075391	1	1	0	0	0	0	1	0	2008
15	Agomelatine (NCPE (2009c))	0	0	8448	190000	0	0	0	0	0	0	1	0	2009
16	Ustekinumab (NCPE (2009b))	1	0	dominant	2192951	0	0	0	0	0	0	0	0	2009
17	Glucosamine Sulfate (NCPE (2009d))	0	0	36154	NA	0	0	0	0	0	-1	1	0	2009
18	Colorectal Cancer Screening (HIQA (2009))	1	0	1696	NA	0	0	0	0	0	-1	-1	0	2009
19	Sapropterin (NCPE (2009e))	0	0	NA	31350000	1	0	0	0	0	-1	1	0	2009
20	Ambrisentan (NCPE (2009a))	1	0	dominant	NA	1	0	0	0	0	0	1	0	2009

Continued on next page

Table 7.2 – continued from previous page

Case	Drug	Decision	Eff	ICER	BI	Sev	IoL	Stake	Innov	Child	QoE	Unc	Media	Year
21	Geftinib (NCPE (2010 <i>g</i>))	0	0	110564	4400000	1	0	0	1	0	-1	0	0	2010
22	Eltrombopag (NCPE (2010 <i>c</i>))	0	0	1598577	NA	1	0	0	0	0	-1	0	0	2010
23	Roflumilast (NCPE (2010 <i>f</i>))	0	-1	1541	3421490	1	0	0	0	0	-1	1	0	2010
24	Certolizumab (NCPE (2010 <i>a</i>))	0	0	116847	5100000	1	0	0	0	0	-1	0	0	2010
25	Denosumab (NCPE (2010 <i>b</i>))	1	0	12360	14400000	1	0	0	0	0	0	0	0	2010
26	Naloxone (NCPE (2010 <i>e</i>))	0	0	14808	2950000	0	0	0	0	0	-1	1	0	2010
27	Golumab (NCPE (2010 <i>d</i>))	0	0	31212	42353733	1	0	0	0	0	0	1	0	2010
28	Eculizumab (NCPE (2010 <i>g</i>))	0	0	1111111*	66250000	1	0	0	1	0	0	1	0	2010
29	Prasugrel (NCPE (2010 <i>h</i>))	1	0	424	3128908	0	0	0	0	0	0	0	0	2010
30	Rotavirus vaccine (NCPE (2010 <i>i</i>))	0	0	111076	32700000	0	0	1	0	1	0	1	0	2010
31	Denosumab (citeDenosumab11)	1	0	21643	842475	0	0	0	0	0	0	0	0	2011
32	Ticagrelor (NCPE (2011 <i>d</i>))	1	0	dominant	3459187	0	0	0	1	0	0	0	0	2011
33	Oncotype DX (NCPE (2011 <i>b</i>))	0	0	25615	1190000	0	0	0	0	0	0	1	1	2011
34	Fingolimod (NCPE (2011 <i>a</i>))	0	0	93669	8600000	1	0	0	0	0	0	0	1	2011
35	Ipilimumab (NCPE (2011 <i>f</i>))	0	0	147899	6750000	1	1	0	1	0	0	0	1	2011
36	Dabigatran (NCPE (2011 <i>c</i>))	0	0	20654	17121000	0	0	0	0	0	-1	1	1	2011
37	Vernakalant (NCPE (2011 <i>e</i>))	1	0	dominant	NA	0	0	0	0	0	0	0	0	2011
38	Tapentadol (NCPE (2011 <i>g</i>))	1	0	3534	912000	0	0	0	0	0	-1	1	0	2011
39	Apixaban (NCPE (2012 <i>a</i>))	1	0	dominant	888559	0	0	0	0	0	0	1	0	2012
40	Rivaroxaban (NCPE (2012 <i>f</i>))	0	0	22663	42300000	0	0	0	0	0	0	0	0	2012
41	Zytiga (NCPE (2012 <i>d</i>))	0	0	135454	9840000	1	1	0	0	0	0	0	0	2012
42	Dexamethasone (NCPE (2012 <i>c</i>))	0	0	23541	9219005	1	0	0	0	0	0	0	0	2012

Continued on next page

Table 7.2 – continued from previous page

Case	Drug	Decision	Eff	ICER	BI	Sev	EoL	Stake	Innov	Child	QoE	Unc	Media	Year
43	Cabazitaxel (NCPE (2012e))	0	0	120084	5600000	1	1	0	0	0	-1	0	0	2012
44	Rivaroxaban (NCPE (2012g))	1	0	dominant	9760000	0	0	0	0	0	0	0	0	2012
45	Telaprevir(NCPE (2012h))	1	0	16023	18500000	0	0	1	1	0	0	0	0	2012
46	Boceprevir (NCPE (2012b))	1	0	11411	6500000	0	0	1	1	0	0	0	0	2012

Table 7.3: MCDA: DIC for model fit for forward stepwise regression.

Predictor	P	ICER + P	ICER + In- nov + P	ICER + In- nov + Unc + P	ICER + In- nov+ Unc + EoL + P	ICER + In- nov+ Unc + EoL + Eff + P
Eff	49.34	32.69	24.36	16.82	7.04*	-
ICER	35.95*	-	-	-	-	-
BI	52.25	38.15	30.71	24.18	14.73	8.09
Sev	48.77	37.82	30.72	24.67	8.20	7.61
EoL	52.48	33.58	26.18	13.14*	-	-
Stake	52.31	36.70	29.46	22.21	13.15	6.92
Innov	51.16	28.55*	-	-	-	-
Child	49.31	36.13	28.62	22.04	13.14	6.97
QoE	43.22	32.52	23.92	21.89	14.65	7.19
Unc	50.09	31.56	21.92*	-	-	-
Media	44.84	32.14	27.35	22.66	13.20	7.10
Year	50.49	35.96	28.51	22.29	13.20	6.63

7.3.2 Logistic Regression

Multiple linear logistic regression was used to predict the outcome of reimbursement based on the predictor variables. Forward stepwise regression is used to determine the best fit model. Model fit is measured using the DIC. At each step the variable, which yields the best model fit is added to the list of predictors as long as the additional variable provides an improvement in model fit compared to the simpler model. Table 7.3 displays the DICs for the alternative models.

When fitting the model based on only one predictor, the ICER provides the best fit (DIC=35.95), which is in line with the finding of the univariate analysis in the previous section. Adding innovation as a second criteria to the ICER is the best fit based on two predictor variables (DIC=28.55). Best fit based on three predictors is given when adding level of uncertainty as a criterion to ICER and innovation (DIC=21.92). Another improvement in model fit is given when adding end of life as a fourth predictor variable (DIC=13.14). Efficacy as a fifth variable also improves the model fit (DIC=7.04); while a sixth variable does not yield an improvement in model fit.

The model of choice therefore uses the ICER, innovation, uncertainty, end of life and efficacy as predictor variables for the reimbursement recommendation. For each decision made in the past, the model estimates the probability of reimbursement. Figure 7.1 plots the probability of reimbursement against the actual recommendation for each of the cases included in the analysis. Based on the model a technology would be recommended for

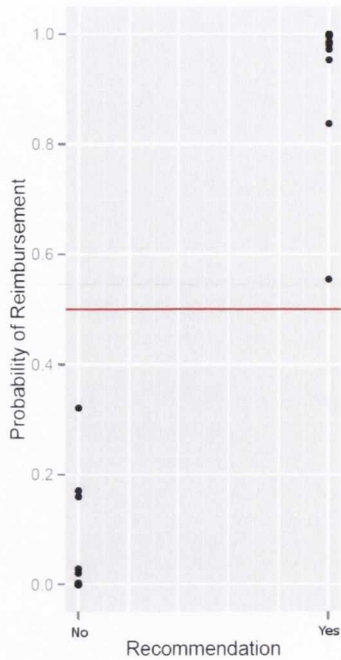


Fig. 7.1: MCDA: Best prediction of reimbursement decision based on logistic regression using the ICER, innovation, level of uncertainty and end of life as predictor variables.

Table 7.4: MCDA: Regression coefficients.

	Intercept	ICER	Innov	Unc	EoL	Eff
1 Pred	1.63 (0.6)	$-5.8e^{-5}(2.3e^{-5})$	-	-	-	-
2 Pred	2.34 (1.1)	$-1.6e^{-4}(6.3e^{-5})$	7.22 (2.9)	-	-	-
3 Pred	8.42 (3.3)	$-3.7e^{-4}(1.5e^{-4})$	15.17 (6.5)	-6.29 (2.6)	-	-
4 Pred	21.18 (8.6)	$-9.7e^{-4}(3.9e^{-4})$	32.48 (13.4)	-18.34 (7.8)	29.46 (16.2)	-
5 Pred	29.36 (11.3)	$-13.4e^{-4}(5.1e^{-4})$	42.10 (16.0)	-21.30 (8.7)	31.96 (17.4)	29.09 (18.1)

reimbursement when the probability of reimbursement is above 0.5, technologies scoring a probability below 0.5 would not be recommended for reimbursement; this is equivalent to the decision rule defined in 7.2.

The model based on five predictors provides a very good fit, classifying all decisions correctly. Denosumab, a treatment for the prevention of skeletal related events in adults with bone metastases from solid tumours, has a probability of reimbursement of 0.56, just above the threshold of 0.5. This is due to a relatively large ICER for a treatment, which is not innovative and does not prolong life.

The coefficients of the best fit models based on one to five predictors are summarised in table 7.4; they contain information of the impact the different criteria have on the reimbursement decision. While uncertainty is high, all coefficients have a 95% credible interval completely positive or completely negative. Based on the best fit model, a non-

innovative intervention with normal level of uncertainty and which is not life prolonging, would be reimbursed for an ICER up to €21,862 per additional QALY. For an innovative treatment past decisions indicate a willingness to pay an additional €31,355 per QALY; for treatments prolonging life the model estimates a willingness to pay an additional €23,800 per QALY gained. For assessments with a high level of uncertainty the willingness to pay decreases by €15,865 per QALY and when efficacy is low, the ICER decreases by €21,666 per QALY.

7.4 Conclusions / Limitations

The analysis confirms that recommendations for or against the reimbursement of technologies are not only driven by the cost-effectiveness. Apart from the ICER, the regression identified innovation, uncertainty, end of life and efficacy to be the key drivers for the decision. While uncertainty surrounding the estimates is high, due to the limited number of cases relative to the number of predictors, the results resemble what one would expect. The NCPE currently reports result at a threshold of €20,000 and €45,000 to the HSE; an ICER of €21,862 for a standard technology is therefore in line with the reported thresholds. The development of innovative treatments is expensive for drug companies due to high research and development costs. To support the development of innovative treatments the payer is willing to pay a higher price for such treatments as an investment in drug development. The QALY is a time independent measure; it assumes a year of life at perfect health to have the same value in the middle of someone's life as it has at the end of someone's life. However, society tends to place a special value on life extensions, as long as it is of reasonable quality. The analysis confirms this; the model suggests a higher threshold for life prolonging technologies. To ensure value for money, technologies surrounded by a high level of uncertainty need to prove their cost-effectiveness at a lower threshold.

This is the first phase of the development of a MCDA approach for the Irish healthcare setting. A descriptive approach is taken to explain the impact different criteria have had on reimbursement recommendations in the past. There are a number of limitations associated with this analysis.

The analysis is limited to data available within the public domain, it may therefore be possible that other factors influencing the recommendation may not be captured.

The outcome variable of the analysis distinguishes between positive and negative recommendation for reimbursement. For the analysis for decision making in NICE, Dakin et al. (2006) suggest a 3rd outcome representing the "yes, but..." decision. For a num-

ber of assessments the NCPE issues a recommendation “not at this price”, which could be captured in a multinomial outcome measure. Also, it would be of interest to model the actual reimbursement decision in addition to the recommendation. However, HTAs are often followed by discussions on price reduction and assessments based on the newly agreed price are not available.

Some of the criteria are not easily defined. Gross budget impact was reported for the majority of treatments and therefore used to measure budget impact. However, it would be interesting to see if net budget impact yields a different result. Unfortunately the net budget impact is not reported for many cases. A judgement has to be made in the classification of severity of disease; it is not clear where to draw the line as to whether a disease should be classified as severe or not. Linley and Hughes (2012) analyse the impact of different factors on recommendations by the All Wales Medicines Strategy Group; they utilized disability-adjusted life years (DALYs) as a measure of health burden to construct a categorical health burden variable. Innovative treatments are either the first treatment for a disease or significantly improve efficacy or reduce costs compared to alternative treatments. Nevertheless, judgements need to be made as to what constitutes a significant improvement in efficacy or a significant reduction in costs.

The analysis is also influenced by the scale used to measure the criteria. Due to the difficulties in clearly defining the criteria, a simple binary structure was used for most of the criteria. A more sophisticated approach for some of the criteria may be beneficial. For instance, measuring uncertainty on a continuous scale, such as the EVPI, would capture the full level of uncertainty. Media coverage and stakeholder involvement should distinguish between positive and negative involvement. However, the number of cases is too limited for such an approach. Four cases got media attention and stakeholders were involved in four cases; in all of these cases the support was pro reimbursement. Media attention has not influenced the recommendation of reimbursement in the past. However, media coverage may have more impact on the decision maker than on the advisory body. Case 28, the assessment of eculizumab, is excluded in a sensitivity analysis because the ICER is reported in terms of life years gained rather than QALYs. This did not change the stepwise logistic regression; coefficients based on the model with four predictors remain unchanged.

The analysis is a descriptive approach to explain past reimbursement decision making. Key drivers of past decisions are identified to be the ICER, innovation, level of uncertainty, end of life and efficacy. While these criteria have influenced decision making in the past, they have done so in an informal way. The value of innovation and life prolong-

ing treatments have never been formally discussed. This analysis quantifies the impact criteria have had on reimbursement decisions. The results provide a basis for discussions between decision makers, experts and other stakeholders on whether the criteria, which have influenced decisions in the past, should also be the criteria to influence decisions in the future. MCDA forces the quantification of value of innovation, value of end of life, etc. as well as which level of uncertainty is acceptable for a positive recommendation. Careful consideration is necessary to determine the relative importance of the criteria before a MCDA approach can be applied to actually aid decision making in the Irish healthcare setting. Further details on this can be found in the “Further Work” section of chapter 8.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

The methodologies developed in this thesis improve evidence synthesis and decision making in healthcare.

Mixed treatment comparison models provide a powerful tool to estimate the relative efficacy between alternative interventions, when direct evidence is not available or not sufficient. These methods are therefore a key component of evidence synthesis and form the basis for HTAs. To ensure efficient use of available evidence, the increased power of continuous outcome measures in such models compared to binary measures was investigated. This is of particular interest for MTC modelling, since indirect comparisons suffer from an increased variance compared to direct comparisons. Furthermore, dichotomization is generally accepted in clinical trials, which aim to quantify a particular treatment effect. Further use of the trial data for pharmacoeconomic assessments is not taken into account.

To facilitate informed decision making based on all available evidence, a methodology was developed to systematically include evidence from different trial designs. While RCT data minimizes potential bias, observational evidence provides a better reflection of real life efficacy. The combination of evidence sources in a hierarchical model allows for the investigation of differences between these while adopting a conservative estimate when estimates differ greatly. Bias adjustment for different trial designs is possible. Three alternative methods were introduced, which provide a novel approach to combining different sources of evidence in a MTC model. This is an important development, since the incentive of combining evidence in a MTC model is to make an informed decision based on all available evidence.

While evidence on relative treatment effectiveness provides an important part of HTAs,

a decision on reimbursement is influenced by a number of other factors as well. MCDA formally aids decision making based on multiple criteria. Based on past reimbursement decisions, the analysis in this thesis identified important criteria, which have influenced decisions in the past, to be the ICER, innovation, uncertainty, end of life and efficacy. While it has never been formally discussed, the analysis illustrates what has been indirectly implied as the value of innovation or life prolonging treatments. The analysis is a novel approach to the Irish healthcare setting and is of great interest to the decision maker. It encourages a discussion on whether criteria that have influenced decisions in the past should be the criteria to influence future reimbursement decisions. A MCDA approach for the Irish healthcare setting improves consistency and transparency in the decision making process.

8.2 Further Work

8.2.1 Simulation Study

The analysis in chapter 4 has illustrated the enhanced sensitivity to change of continuous measures in MTC models compared to using binary outcome measures. To inform clinical guidelines it is of interest to investigate factors influencing the impact of dichotomizing outcome measures in further detail.

Further insight can be obtained by a simulation study. It is likely that the level of heterogeneity plays an important role. A simulation study allowing for different levels of heterogeneity would allow for further insight into the relationship between level of heterogeneity and impact of dichotomizing outcome measures.

8.2.2 Inclusion of Registry Data in MTC

An obvious extension to the work carried out in chapter 6 is to increase the evidence to be included in the analysis.

A topical development in HTA in Ireland is the set up of registries following reimbursement. The new health bill, which is expected to pass into law by the end of the year, includes HTA into legislation and as part of it introduces the concept of reassessing technologies every 3 to 5 years (*Health (Pricing and Supply of Medical Goods) Bill 2012* (2012)). This enables the state to not only add technologies to the list of reimbursed medicines when value for money was demonstrated, but also to take treatments off this list, when reassessment in light of new information suggests otherwise. Furthermore, when uncertainty surrounding the estimates is high, it is possible to issue a conditional

reimbursement. Treatments can hence be reimbursed conditional on the collection of further evidence. When uncertainty can be attributed to clinical effectiveness, further information can be obtained by setting up a registry following patients on the treatment; such as the registry following hepatitis C patients on telaprevir and boceprevir, which have been recently added to the list of reimbursed treatments.

For the reassessment of such interventions, available RCT evidence and observational data need to be combined with evidence collected in the registry. This allows to update the estimates of effectiveness in the light of real life data collected in the patient population which is of interest to the decision maker. The methodology developed in chapter 6 allows for the inclusion of summarized registry data. However, it would be of great benefit to extend the methodology to allow for patient level data inputs, where these are available. Riley et al. (2008) and Sutton et al. (2008) have developed methodologies to combine individual patient level data and aggregate data for standard meta-analyses; the aim is to generalize this work to MTC models.

8.2.3 Normative Approach to MCDA

The work described in chapter 7 is the first phase of the development of a prescriptive approach to formalize reimbursement decisions in healthcare incorporating multiple criteria. The current research has identified the key criteria influencing reimbursement decisions in the past. Figure 8.1 illustrates how current research will feed into future research to develop a normative MCDA approach to the Irish healthcare setting. The analysis in this

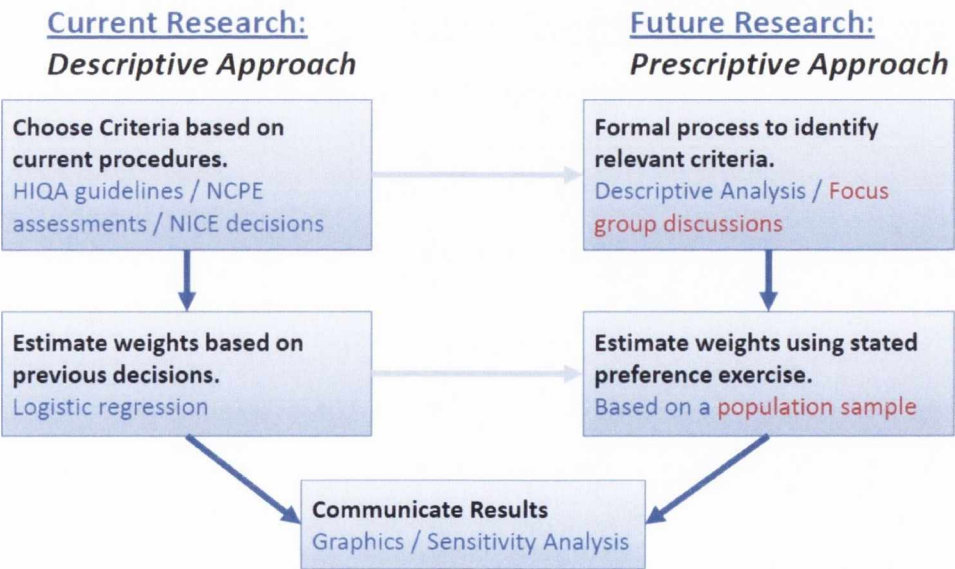


Fig. 8.1: MCDA reasearch plan: Illustration of how current research will feed into future research to develop a normative MCDA approach for the Irish healthcare setting.

thesis has identified the ICER, the innovation of the treatment, the uncertainty in the estimates, whether or not the treatment is life prolonging and efficacy as the key criteria to impact reimbursement decisions. The next step towards a prescriptive approach is to involve all stakeholders, including medical professionals, experts from the NCPE, HIQA and the NCCP, decision makers from the HSE as well as patient representatives, in a discussion as to whether the criteria, which have influenced decisions in the past should be the criteria to influence decisions in the future. The analysis in this thesis provides the basis for these focus group discussions to identify criteria, which should play a role in future reimbursement decision making. Once the criteria have been defined, the relative importance of the criteria needs to be determined. Again, estimates from the descriptive analysis provide a starting point, giving an estimate of how much has been paid in the past for an innovative or a life prolonging intervention compared to an average intervention or what threshold applies when uncertainty surrounding the analysis is high. To reflect population preferences well, these relative weights should be estimated using a stated preference exercise based on a population sample. Methodologies in patient sampling techniques to ensure a representative sample of the population is drawn as well as to include patient preferences and experiences in the analysis need to be developed. Furthermore, the technique requires the development of a questionnaire to determine the relative preferences between the criteria.

8.3 Final Remarks

As mentioned in the introduction, Bayesian methods are particularly useful for a decision theoretic approach to healthcare problems. The methodologies developed in this thesis improve the practice of evidence synthesis and decision making in this area. Optimizing the use of available evidence as a basis for an informed decision making process is very important to ensure that scarce resources are allocated fairly and efficiently.

Appendix A

Code and Data for Chapter 4

Bugs code for continuous model

```
model{
for(i in 1:ns){
alpha[i,bi[i]] <- - 0
nu[i] ~ dnorm(0,0.0001)
for (j in 1:na[i]){
tau.Delta[i,j] <- - 1/(se.Delta[i,j]·se.Delta[i,j])
Delta[i,j] ~ dnorm(mu.Delta[i,j],tau.Delta[i,j])
mu.Delta[i,j] <- - lambda[i,j]·delta[i,j]
delta[i,j] <- - nu[i] + alpha[i,t[i,j]] + b·I[i,j] }
for (j in 2:noT[i]){
alpha[i,si[i,j]] ~ dnorm(a[si[i,j]]-a[bi[i]],tau.a) }
}
a[1] <- -0
for (j in 2:nt){a[j] ~ dnorm(0,0.0001)}
sd ~ dunif(0,2)
b ~ dnorm(0,0.0001)
tau.a<-1/var
var.a<-pow(sd,2)
for(k in 1:(nt-1)){ for (l in (k+1):nt){
IC[k,l] <- - a[k] - a[l] }}
}
```

Bugs code for binary model

```
model{
for(i in 1:ns){
alpha[i,bi[i]] <- 0
nu[i] ~ dnorm(0,0.0001)
for (j in 1:na[i]){
r[i,j] ~ dbin(n[i,j],p[i,j])
logit[p[i,j]] <- nu[i] + alpha[i,t[i,j]] + b*I[i,j] }
for (j in 2:noT[i]){
alpha[i,si[i,j]] ~ dnorm(a[si[i,j]]-a[bi[i]],tau.a) }
}
a[1] <- 0
for (j in 2:nt{a[j] ~ dnorm(0,0.0001)}
sd ~ dunif(0,2)
b ~ dnorm(0,0.0001)
tau.a<-1/var
var.a<-pow(sd,2)
for(k in 1:(nt-1)){ for (l in (k+1):nt){
IC[k,l] <- - a[k] - a[l] }}
}
```

Input Data

% improvement in HAQ score

```
list(Delta= structure(.Data= c(0.27, 0.54, 0.62, 0.59, NA, 0.24, 0.56, 0.60, NA, NA, 0.07,
0.39, 0.29, 0.49, 0.38, -0.10, 0.20, 0.20, 0.40, NA, 0.20, 0.50, NA, NA, NA, 0.30, 0.30,
0.50, 0.50, 0.40, 0.45, 0.76, NA, NA, NA, 0.03, 0.58, 0.62, NA, NA, 0.40, 0.70, NA, NA,
NA, 0.13, 0.13, 0.38, 0.50, NA, 0.18, 0.60, 0.63, NA, NA, 0.14, 0.50, 0.50, NA, NA, -0.07,
0.39, NA, NA, NA), .Dim=c(13, 5)),
se.Delta= structure(.Data= c(0.07, 0.07, 0.08, 0.06, NA, 0.04, 0.04, 0.04, NA, NA, 0.05,
0.06, 0.06, 0.05, 0.06, 0.06, 0.05, 0.06, 0.06, NA, 0.06, 0.07, NA, NA, NA, 0.05, 0.05, 0.05,
0.06, 0.06, 0.07, 0.07, NA, NA, NA, 0.07, 0.08, 0.07, NA, NA, 0.12, 0.09, NA, NA, NA,
0.03, 0.06, 0.05, 0.05, NA, 0.05, 0.03, 0.03, NA, NA, 0.04, 0.03, 0.03, NA, NA, 0.04, 0.06,
NA, NA, NA), .Dim=c(13, 5)),
lambda = structure(.Data= c(1.64, 1.52, 1.55, 1.55, NA, 1.45, 1.44, 1.48, NA, NA, 1.88,
```



```

1.88, 1.88, 1.84, 1.83, 1.39, 1.57, 1.64, 1.77, NA, 1.3, 1.4, NA, NA, NA, 1.8, 1.8, 1.8,
1.8, 1.5, 1.6, 1.6, NA, NA, NA, 1.7, 1.7, 1.6, NA, NA, 1.5, 1.5, NA, NA, NA, 1.25, 1.38,
1.38, 1.38, NA, 1.7, 1.7, 1.7, NA, NA, 1.6, 1.6, 1.6, NA, NA, 1.55, 1.43, NA, NA, NA),
.Dim=c(13, 5)),
na=c(4, 3, 5, 4, 2, 5, 2, 3, 2, 4, 3, 3, 2),
nt=6,
ns=13,
bi=c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1),
si= structure(.Data= c(NA, 2, NA, 2, NA, 2, NA, 2, NA, 2, NA, 3, NA, 3, NA, 4, NA,
4, NA, 5, NA, 6, NA, 6, NA, 6), .Dim=c(13, 2)),
t= structure(.Data= c(1, 2, 2, 2, NA, 1, 2, 2, NA, NA, 1, 2, 2, 2, 2, 1, 2, 2, 2, NA, 1, 2,
NA, NA, NA, 1, 3, 3, 3, 3, 1, 3, NA, NA, NA, 1, 4, 4, NA, NA, 1, 4, NA, NA, NA, 1, 5,
5, 5, NA, 1, 6, 6, NA, NA, 1, 6, 6, NA, NA, 1, 6, NA, NA, NA), .Dim=c(13, 5)),
noT=c(2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2),
I= structure(.Data= c(1, 1, 1, 1, NA, 1, 1, 1, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, NA,
1, 1, NA, NA, NA, 1, 1, 1, 1, 1, 1, 1, NA, NA, NA, 0, 0, 0, NA, NA, 1, 1, NA, NA,
NA, 1, 0, 1, 1, NA, 1, 1, 1, NA, NA, 1, 1, 1, NA, NA, 0, 0, NA, NA, NA), .Dim=c(13, 5)))

```

HAQ 20 and HAQ 50

replace Delta and se.Delta from HAQ input data above with n and p:

```

n= structure(.Data= c(62, 69, 67, 73, NA, 200, 212, 207, NA, NA, 110, 112, 106, 103,
113, 87, 87, 91, 87, NA, 63, 65, NA, NA, NA, 88, 86, 86, 87, 81, 86, 87, NA, NA, NA,
80, 76, 78, NA, NA, 30, 59, NA, NA, NA, 133, 133, 89, 89, NA, 199, 393, 390, NA, NA,
127, 246, 246, NA, NA, 109, 111, NA, NA, NA), .Dim=c(13, 5))
p= structure(.Data= c(30, 45, 47, 52, NA, 95, 141, 154, NA, NA, 29, 47, 60, 57, 60, 23,
37, 39, 48, NA, 28, 43, NA, NA, NA, 40, 40, 54, 52, 48, 52, 61, NA, NA, NA, 25, 47, 53,
NA, NA, 16, 45, NA, NA, NA, 49, 56, 51, 60, NA, 83, 863, 249, NA, NA, 46, 163, 155,
NA, NA, 19, 60, NA, NA, NA), .Dim=c(13, 5)) ## HAQ 20 responders
p= structure(.Data= c(12, 24, 26, 26, NA, 32, 81, 82, NA, NA, 4, 16, 19, 20, 21, 8, 10,
16, 18, NA, 10, 24, NA, NA, NA, 9, 12, 19, 23, 20, 28, 39, NA, NA, NA, 8, 26, 31, NA,
NA, 8, 29, NA, NA, NA, 13, 29, 24, 31, NA, 35, 146, 137, NA, NA, 10, 66, 67, NA, NA,
2, 34, NA, NA, NA), .Dim=c(13, 5))## HAQ 50 responders

```

ACRcont¹

```
list(Delta= structure(.Data= c(0.08, 0.35, 0.46, 0.42, NA, 0.13, 0.42, 0.41, NA, NA, 0.09,
0.31, 0.3, 0.38, 0.33, 0.07, 0.28, 0.33, 0.37, NA, 0.18, 0.42, NA, NA, NA, 0.09, 0.35, 0.36,
0.38, 0.37, 0.12, 0.37, 0.38, NA, NA, 0.27, 0.46, NA, NA, NA, 0.22, 0.41, NA, NA, NA,
0.06, 0.34, 0.4, NA, NA, 0.1, 0.43, NA, NA, NA, 0.14, 0.3, 0.4, 0.39, NA, 0.14, 0.39, 0.38,
NA, NA, 0.07, 0.4, 0.41, NA, NA, 0.04, 0.38, 0.38, NA, NA, 0.04, 0.33, NA, NA, NA),
.Dim=c(16, 5)),
se.Delta= structure(.Data= c(0.03, 0.02, 0.03, 0.02, NA, 0.02, 0.02, 0.01, NA, NA, 0.02,
0.02, 0.02, 0.02, 0.02, 0.02, 0.02, 0.02, 0.02, NA, 0.03, 0.03, NA, NA, NA, 0.02, 0.02, 0.02,
0.02, 0.02, 0.01, 0.01, 0.01, NA, NA, 0.03, 0.02, NA, NA, NA, 0.03, 0.02, NA, NA, NA,
0.02, 0.02, 0.02, NA, NA, 0.03, 0.03, NA, NA, NA, 0.02, 0.02, 0.02, 0.02, NA, 0.03, 0.03,
0.04, NA, NA, 0.01, 0.01, 0.01, NA, NA, 0.01, 0.01, 0.01, NA, NA, 0.01, 0.02, NA, NA,
NA), .Dim=c(16, 5)),
na=c(4, 3, 5, 4, 2, 5, 3, 2, 2, 3, 2, 4, 3, 3, 3, 2),
nt=6,
ns=16,
bi=c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1),
si= structure(.Data= c(NA, 2, NA, 2, NA, 2, NA, 2, NA, 2, NA, 3, NA, 3, NA, 3, NA,
3, NA, 4, NA, 4, NA, 5, NA, 5, NA, 6, NA, 6, NA, 6.), .Dim=c(16, 2)),
t= structure(.Data= c(1, 2, 2, 2, NA, 1, 2, 2, NA, NA, 1, 2, 2, 2, 2, 1, 2, 2, 2, NA, 1, 2,
NA, NA, NA, 1, 3, 3, 3, 3, 1, 3, 3, NA, NA, 1, 3, NA, NA, NA, 1, 3, NA, NA, NA, 1, 4,
4, NA, NA, 1, 4, NA, NA, NA, 1, 5, 5, 5, NA, 1, 5, 5, NA, NA, 1, 6, 6, NA, NA, 1, 6, 6,
NA, NA, 1, 6, NA, NA, NA), .Dim=c(16, 5)),
noT=c(2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2),
I= structure(.Data= c(1, 1, 1, 1, NA, 1, 1, 1, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, NA, 1, 1,
NA, NA, NA, 1, 1, 1, 1, 1, 1, 1, 1, NA, NA, 1, 1, NA, NA, NA, 1, 1, NA, NA, NA, 0, 0,
0, NA, NA, 1, 1, NA, NA, NA, 1, 0, 1, 1, NA, 1, 1, 1, NA, NA, 1, 1, 1, NA, NA, 1, 1, 1,
NA, NA, 0, 0, NA, NA, NA), .Dim=c(16, 5)))
```

ACR 20, ACR 50 and ACR 70

```
replace Delta and se.Delta from ACRcont input data above with n and p:
n= structure(.Data= c(62, 69, 67, 73, NA, 200, 207, 212, NA, NA, 110, 112, 106, 103,
113, 87, 87, 91, 87, NA, 63, 65, NA, NA, NA, 88, 86, 86, 87, 81, 363, 360, 361, NA, NA,
```

¹Adjustments to code for ACRcont: ACR already is a % improvement; therefore no baseline disease parameter is included. Also delta is already calculated from delta.c and delta.t, this line has to be deleted from code.

86, 87, NA, NA, NA, 110, 165, NA, NA, NA, 80, 76, 78, NA, NA, 30, 59, NA, NA, NA, 133, 133, 89, 89, NA, 35, 35, 34, NA, NA, 199, 393, 390, NA, NA, 127, 246, 246, NA, NA, 109, 111, NA, NA, NA), .Dim=c(16, 5))

p= structure(.Data= c(9, 33, 45, 48, NA, 59, 131, 129, NA, NA, 21, 44, 38, 55, 52, 12, 25, 40, 44, NA, 23, 40, NA, NA, NA, 18, 45, 47, 51, 49, 87, 199, 205, NA, NA, 42, 66, NA, NA, NA, 49, 98, NA, NA, NA, 9, 39, 46, NA, NA, 8, 42, NA, NA, NA, 37, 47, 53, 53, NA, 13, 21, 19, NA, NA, 27, 231, 237, NA, NA, 11, 141, 142, NA, NA, 10, 51, NA, NA, NA), .Dim=c(16, 5)) ## ACR 20 responders

p= structure(.Data= c(5, 22, 37, 31, NA, 19, 81, 87, NA, NA, 9, 23, 20, 36, 25, 5, 14, 22, 28, NA, 9, 28, NA, NA, NA, 7, 22, 25, 26, 21, 33, 110, 119, NA, NA, 22, 38, NA, NA, NA, 22, 61, NA, NA, NA, 4, 18, 31, NA, NA, 1, 23, NA, NA, NA, 18, 26, 33, 29, NA, 2, 13, 10, NA, NA, 15, 146, 156, NA, NA, 4, 80, 81, NA, NA, 4, 25, NA, NA, NA), .Dim=c(16, 5)) ## ACR 50 responders

p= structure(.Data= c(3, 7, 18, 14, NA, 5, 43, 37, NA, NA, 2, 11, 9, 19, 14, 1, 9, 11, 13, NA, 5, 14, NA, NA, NA, 0, 7, 9, 15, 9, 16, 48, 54, NA, NA, 12, 20, NA, NA, NA, 10, 40, NA, NA, NA, 1, 7, 12, NA, NA, 0, 9, NA, NA, NA, 7, 15, 18, 16, NA, 0, 3, 6, NA, NA, 6, 84, 79, NA, NA, 1, 39, 26, NA, NA, 0, 9, NA, NA, NA), .Dim=c(16, 5)) ## ACR 70 responders

Appendix B

Code and Data for Chapter 6

Code for 3-level model (RA)

```
model{
### Overall Level ###
for (k in 2:ntCO){
aRCT[k]~dnorm(a[k],tauCO)
aOBS[k]~dnorm(a[k],tauCO)}
for (k in (ntCO+1):ntRCT){
aRCT[k]~dnorm(0,0.0001)
a[k]<-dRCT[k]}
a[1]<-0
for (k in 2:4){a[k]~dnorm(0,0.0001)}
sdCO~dunif(0,0.25)
varCO<-pow(sdCO,2)
tauCO<-1/varCO
for (c in 1:(nt-1)) { for (k in (c+1):nt) {
IC[c,k]<-(a[k]-a[c])}}
### RCT bit ###
for(i in 1:nsRCT){
wRCT[i,1] <- 0
alphaRCT[i,biRCT[i]] <- 0
nuRCT[i] ~ dnorm(0,.0001)
for (k in 1:naRCT[i]) {
HAQtauRCT[i,k] <- 1/(seRCT[i,k]*seRCT[i,k])
DeltaRCT[i,k] ~ dnorm(mu.DeltaRCT[i,k],HAQtauRCT[i,k])
```

```

mu.DeltaRCT[i,k] <- deltaRCT[i,k]*lambdaRCT[i,k]
deltaRCT[i,k] <- - nuRCT[i] + alphaRCT[i,tRCT[i,k]]+ betaRCT[i]*IRCT[i,k]
}
for (k in 2:noTRCT[i]) {
alphaRCT[i,siRCT[i,k]] ~ dnorm(aRCT[siRCT[i,k]] - aRCT[biRCT[i]] + swRCT[i,k], tau-
dRCT[i,siRCT[i,k]])
taudRCT[i,siRCT[i,k]] <- tauRCT *2*(k-1)/k
wRCT[i,k] <- (deltaRCT[i,siRCT[i,k]] - dRCT[siRCT[i,k]] + dRCT[biRCT[i]])
swRCT[i,k] <- sum(wRCT[i,1:k-1])/(k-1)
}
betaRCT[i]<-bRCT
}
aRCT[1]<-0
sdRCT~dunif(0,0.25)
mbetaRCT ~ dnorm(0,1.0E-6)
varRCT <- pow(sdRCT,2)
tauRCT <- 1/varRCT
for (c in 1:(ntRCT-1)) { for (k in (c+1):ntRCT) {
ICRCT[c,k]< -(aRCT[k]-aRCT[c])} }
### OBS bit ###
for(i in 1:nsOBS){
wOBS[i,1] <- -0
aOBS[i,biOBS[i]] <- - 0
nuOBS[i] ~ dnorm(0,.0001)
for (k in 1:naOBS[i]) {
HAQtauOBS[i,k] <- - 1/(seOBS[i,k]*seOBS[i,k])
DeltaOBS[i,k] ~ dnorm(mu.DeltaOBS[i,k],HAQtauOBS[i,k])
mu.DeltaOBS[i,k] <- deltaOBS[i,k]*lambdaOBS[i,k]
deltaOBS[i,k] <- - nuOBS[i] + alphaOBS[i,tOBS[i,k]]
}
for (k in 2:noTOBS[i]) {
alphaOBS[i,siOBS[i,k]] ~ dnorm(dOBS[siOBS[i,k]] - dOBS[biOBS[i]] + swOBS[i,k], tau-
dOBS[i,siOBS[i,k]])
taudOBS[i,siOBS[i,k]] <- tauOBS *2*(k-1)/k
wOBS[i,k] <- (deltaOBS[i,siOBS[i,k]] - dOBS[siOBS[i,k]] + dOBS[biOBS[i]])
}
}

```

```

swOBS[i,k] <- sum(wOBS[i,1:k-1])/(k-1)
}
}
aOBS[1]<-0
sdOBS~dunif(0,0.25)
varOBS <- pow(sdOBS,2)
tauOBS <- -1/varOBS
for (c in 1:(ntOBS-1)) { for (k in (c+1):ntOBS) {
ICOBS[c,k]<- -(aOBS[k]-aOBS[c])) }
}

```

Data for RA analysis

```

list( nt=6, ## Total Number of treatments
ntRCT=6, ## Total Number of treatments in RCT trials
ntCO=4, ## Total Number of treatments in both designs nsRCT=13, ## Total Number
of RCT trials
ntOBS=4, ## Total Number of treatments in OBS trials
nsOBS=3, ## Total Number of OBS trials
IRCT= structure(.Data= c(1, 1, 1, 1, NA, 1, 1, 1, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0, NA,
1, 1, NA, NA, NA, 1, 1, 1, 1, 1, 1, 1, NA, NA, NA, 0, 0, 0, NA, NA, 1, 1, NA, NA, NA,
1, 0, 1, 1, NA, 1, 1, 1, NA, NA, 1, 1, 1, NA, NA, 0, 0, NA, NA, NA), .Dim=c(13, 5)),
## MTX indicator RCT trials
DeltaRCT= structure(.Data= c(0.27, 0.54, 0.62, 0.59, NA, 0.24, 0.56, 0.60, NA, NA,
0.07, 0.39, 0.29, 0.49, 0.38, -0.10, 0.20, 0.20, 0.40, NA, 0.20, 0.50, NA, NA, NA, 0.30,
0.30, 0.50, 0.50, 0.40, 0.45, 0.76, NA, NA, NA, 0.03, 0.58, 0.62, NA, NA, 0.40, 0.70, NA,
NA, NA, 0.13, 0.13, 0.38, 0.50, NA, 0.18, 0.60, 0.63, NA, NA, 0.14, 0.50, 0.50, NA, NA,
-0.07, 0.39, NA, NA, NA), .Dim=c(13, 5)), ## Mean HAQ improvement RCT trials
seRCT= structure(.Data= c(0.07, 0.07, 0.08, 0.06, NA, 0.04, 0.04, 0.04, NA, NA, 0.05,
0.06, 0.06, 0.05, 0.06, 0.06, 0.05, 0.06, 0.06, NA, 0.06, 0.07, NA, NA, NA, 0.05, 0.05, 0.05,
0.06, 0.06, 0.07, 0.07, NA, NA, NA, 0.07, 0.08, 0.07, NA, NA, 0.12, 0.09, NA, NA, NA,
0.03, 0.06, 0.05, 0.05, NA, 0.05, 0.03, 0.03, NA, NA, 0.04, 0.03, 0.03, NA, NA, 0.04, 0.06,
NA, NA, NA), .Dim=c(13, 5)), ## Mean SE of HAQ improvement RCT trials
lambdaRCT= structure(.Data= c(1.64, 1.52, 1.55, 1.55, NA, 1.45, 1.44, 1.48, NA, NA,
1.88, 1.88, 1.88, 1.84, 1.83, 1.39, 1.57, 1.64, 1.77, NA, 1.30, 1.40, NA, NA, NA, 1.80, 1.80,
1.80, 1.80, 1.50, 1.60, 1.60, NA, NA, NA, 1.70, 1.70, 1.60, NA, NA, 1.50, 1.50, NA, NA,

```

```

NA, 1.25, 1.38, 1.38, 1.38, NA, 1.70, 1.70, 1.70, NA, NA, 1.60, 1.60, 1.60, NA, NA, 1.55,
1.43, NA, NA, NA), .Dim=c(13, 5)), ## Initial HAQ score RCT trials
naRCT=c(4, 3, 5, 4, 2, 5, 2, 3, 2, 4, 3, 3, 2), ## Number of arms in RCT trials
biRCT=c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1), ## Baseline treatment in RCT trials
siRCT= structure(.Data= c(NA, 2, NA, 2, NA, 2, NA, 2, NA, 2, NA, 3, NA, 3, NA, 4,
NA, 4, NA, 5, NA, 6, NA, 6, NA, 6), .Dim=c(13, 2)), ## Non-Baseline treatment in
RCT trials
tRCT= structure(.Data= c(1, 2, 2, 2, NA, 1, 2, 2, NA, NA, 1, 2, 2, 2, 2, 1, 2, 2, 2, NA,
1, 2, NA, NA, NA, 1, 3, 3, 3, 3, 1, 3, NA, NA, NA, 1, 4, 4, NA, NA, 1, 4, NA, NA, NA,
1, 5, 5, 5, NA, 1, 6, 6, NA, NA, 1, 6, 6, NA, NA, 1, 6, NA, NA, NA), .Dim=c(13, 5)),
## Treatment in RCT trials
noTRCT=c(2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2), ## Number of treatments in RCT trials
DeltaOBS= structure(.Data= c(0.34, 0.34, 0.34, NA, NA, 0.42, 0.23, 0.35, NA, NA, 0.40,
0.80, NA, NA, NA), .Dim=c(3, 5)), ## Mean HAQ improvement OBS trials
seOBS= structure(.Data= c(0.03, 0.02, 0.03, NA, NA, 0.03, 0.04, 0.05, NA, NA, 0.09,
0.02, NA, NA, NA), .Dim=c(3, 5)), ## SE of HAQ improvement OBS trials
lambdaOBS= structure(.Data= c(1.20, 1.50, 1.23, NA, NA, 1.30, 1.40, 1.40, NA, NA,
1.50, 1.80, NA, NA, NA), .Dim=c(3, 5)), ## initial HAQ in OBS trials
naOBS=c(3, 3, 2), ## number of arms in OBS trials
biOBS=c(2, 2, 1), ## Baseline treatment in OBS trials siOBS= structure(.Data= c(NA,
3, 4, NA, 3, 4, NA, 4, NA), .Dim=c(3, 3)), ## Non-Baseline treatment in OBS trials
tOBS= structure(.Data= c(2, 3, 4, NA, NA, 2, 3, 4, NA, NA, 1, 4, NA, NA, NA),
.Dim=c(3, 5)), ## Treatment in OBS trials
noTOBS=c(3, 3, 2) ## Number of treatments in OBS trials
)

```


Bibliography

- Adams, R., Schmitz, S., FitzGerald, O., Walsh, C. and Barry, M. (2011), 'Does model choice in mixed treatment comparisons affect the decision in economic evaluations?', *Ispor Connections* **17**(2), 6–7.
- Adams, R., Walsh, C., Veale, D., Bresnihan, B., FitzGerald, O. and Barry, M. (2010), 'Understanding the relationship between the EQ-5D, SF-6D, HAQ and disease activity in inflammatory arthritis', *Pharmacoeconomics* **28**(6), 477–487.
- Ades, A., Sculpher, M., Sutton, A., Abrams, K., Cooper, N., Welton, N. and Lu, G. (2006), 'Bayesian methods for evidence synthesis in cost-effectiveness analysis', *Pharmacoeconomics* **24**(1), 1–19.
- Ades, A., Welton, N. and Lu, G. (2007), 'Introduction to mixed treatment comparison'. [Accessed: 10 September 2012].
URL: <https://www.bris.ac.uk/cobm/research/mpes/mtc.html>
- Ashby, D. and Smith, A. (2000), 'Evidence-based medicine as Bayesian decision-making', *Statistics in Medicine* **19**, 3291–3305.
- Austin, P. and Brunner, L. (2004), 'Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses', *Statistics in Medicine* **23**, 1159–1178.
- Banta, D. (2003), 'The development of health technology assesement', *Health Policy* **63**(2), 121–132.
- Bazzani, C., Filippini, M., Caporali, R., Bobbio-Pallavicini, F., Favalli, E., Marchesoni, A., Atzeni, F., Sarzi-Puttini, P. and Gorla, R. (2009), 'Anti-TNF α therapy in a cohort of rheumatoid arthritis patients: Clinical outcomes', *Autoimmunity Reviews* **8**(3), 260–265.
- Bennette, C. and Vickers, A. (2012), 'Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents', *BMC Medical Research Methodology* **12**(21).
- Benson, K. and Hartz, A. (2000), 'A comparison of observational studies and randomized, controlled trials.', *The New England Journal of Medicine* **342**(25), 1878–1886.
- Bernardo, J. and Smith, A. (2000), *Bayesian Theory*, Wiley Series in Probability and Statistics, Chichester.
- Bilcke, J., Beutels, P., Brisson, M. and Jit, M. (2011), 'Accounting for methodological, structural, and parameter uncertainty in decision-analytic models: A practical guide', *Medical Decision Making* **31**(4), 675–692.
- Bojke, L., Claxton, K., Palmer, S. and Sculpher, M. (2006), Defining and characterising structural uncertainty in decision analytic models, Working Papers 009cherp, Centre

for Health Economics, University of York. [Accessed: 01 September 2012].

URL: <http://ideas.repec.org/p/chy/respap/9cherp.html>

- Bojke, L., Claxton, K., Sculpher, M. and Palmer, S. (2008), 'Identifying research priorities: The value of information associated with repeat screening for age-related macular degeneration', *Medical Decision Making* **28**(1), 33–43.
- Bojke, L., Claxton, K., Sculpher, M. and Palmer, S. (2009), 'Characterizing structural uncertainty in decision analytic models: a review and application of methods', *Value in Health* **12**(5), 739–749.
- Breitling, L. and Brenner, H. (2010), 'Odd odds interactions introduced through dichotomisation of continuous outcomes', *Journal of Epidemiology and Community Health* **64**, 300–303.
- Brennan, A., Kharroubi, S., O'Hagan, A. and Chilcott, J. (2007), 'Calculating partial expected value of perfect information via monte carlo sampling algorithms', *Medical Decision Making* **27**(4), 448–470.
- Briggs, A. (2000), 'Handling uncertainty in cost-effectiveness models', *Pharmacoeconomics* **17**(5), 479–500.
- Briggs, A. and Sculpher, M. (1998), 'An introduction to markov modelling for economic evaluation', *Pharmacoeconomics* **13**(4), 397–409.
- Brooks, S. and Roberts, G. (1998), 'Assessing convergence of markov chain monte carlo algorithms', *Statistics and Computing* **8**, 319–335.
- Bucher, H., Guyatt, G., Griffith, L. and Walter, S. (1997), 'The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials', *Journal of Clinical Epidemiology* **50**(6), 683–691.
- Cadavid, D., Wolansky, L., Skurnick, J., Lincoln, J., Cheriyan, J., Szczepanowski, K., Kamin, S., Pachner, A., Halper, J. and Cook, S. (2009), 'Efficacy of treatment of MS with IFN β -1b or glatiramer acetate by monthly brain MRI in the BECOME study', *Neurology* **72**(23), 1976–1983.
- Caldwell, D., Ades, A. and Higgins, J. (2005), 'Simultaneous comparison of multiple treatments: combining direct and indirect evidence', *British Medical Journal* **331**, 897–900.
- Caldwell, D., Welton, N., Dias, S. and Ades, A. (2012), 'Selecting the best scale for measuring treatment effect in a network meta-analysis: a case study in childhood nocturnal enuresis', *Research Synthesis Methods* **3**(2), 126–141.
- Campbell, M. (2010), 'A statistician on a NICE committee', *Significance* **7**(2), 81–84.
- Carra, A., Onaha, P., Sinay, V., Alvarez, F., Luetic, G., Bettinelli, R., San Pedro, E. and Rodriguez, L. (2003), 'A retrospective, observational study comparing the four available immunomodulatory treatments for relapsing-remitting multiple sclerosis', *European Journal of Neurology* **10**(6), 671–676.
- Chen, H., Cohen, P. and Chen, S. (2007), 'Biased odds ratios from dichotomization of age', *Statistics in Medicine* **26**(18), 2487–2497.
- Claxton, K. (2008a), 'Exploring uncertainty in cost-effectiveness analysis', *Pharmacoeconomics* **26**(9), 781–798.

- Claxton, K. (2008b), 'Value based pricing for NHS drugs: an opportunity not to be missed?', *British Medical Journal* **336**(7638), 251–257.
- Cohen, J., Barkhof, F., Comi, G., Hartung, H., Khatri, B., Montalban, X., Pelletier, J., Capra, R., Gallo, P., Izquierdo, G., Tiel-Wilck, K., de Vera, A., Jin, J., Stites, T., Wu, S., Aradhye, S. and Kappos, L. (2010), 'Oral fingolimod or intramuscular interferon for relapsing multiple sclerosis', *The New England Journal of Medicine* **362**(362), 402–415.
- Comi, G., Filippi, M., Wolinsky, J. and European/Canadian Glatiramer Acetate Study Group (2001), 'European/Canadian multicenter, double-blind, randomized, placebo-controlled study of the effects of glatiramer acetate on magnetic resonance imaging-measured disease activity and burden in patients with relapsing multiple sclerosis', *Annals of Neurology* **49**(3), 290–297.
- Comi, G., Jeffery, D., Kappos, L., Montalban, X., Boyko, A., Rocca, M. and Filippi, M. (2012), 'Placebo-controlled trial of oral laquinimod for multiple sclerosis', *The New England Journal of Medicine* **366**, 1000–1009.
- Concato, J., Shah, N. and Horwitz, R. (2000), 'Randomized, controlled trials, observational studies, and the hierarchy of research designs.', *The New England Journal of Medicine* **342**, 1887–1892.
- Cooper, N., Peters, J., Lai, M., Juni, P., Wandel, S., Palmer, S., Paulden, M., Conti, S., Welton, N., Abrams, K., Bujkiewicz, M., Spiegelhalter, D. and Sutton, A. (2011), 'How valuable are multiple treatment comparison methods in evidence-based health-care evaluation?', *Value in Health* **14**(2), 371–380.
- Cooper, N., Sutton, A., Lu, G. and Khunti, K. (2006), 'Mixed comparison of stroke prevention treatments in individuals with nonrheumatic atrial fibrillation', *Archives of Internal Medicine* **166**, 1269–1275.
- Cowles, M. and Carlin, B. (1996), 'Markov chain monte carlo diagnostics: A comparative review.', *Journal of the American Statistical Association* **91**(434), 883–904.
- d'Agostino, R. and d'Agostino, R. (2007), 'Estimating treatment effects using observational data', *Journal of the American Medical Association* **297**(3), 314–316.
- Dakin, H., Devlin, N. and Odeyemi, I. (2006), '"Yes", "No" or "Yes, but"? Multinomial modelling of NICE decision-making', *Health Policy* **77**(3), 352–367.
- Dawson, N. and Weiss, R. (2012), 'Dichotomizing continuous variables in statistical analysis: A practice to avoid', *Medical Decision Making* **32**(2), 225–226.
- Deeks, J. (2002), 'Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes', *Statistics in Medicine* **21**, 1575–1600.
- DeGroot, M. (1970), *Optimal Statistical Decisions*, McGraw-Hill.
- DerSimonian, R. and Laird, N. (1986), 'Meta-analysis in clinical trials', *Controlled Clinical Trials* **7**(3), 177–188.
- Dias, S., Welton, N., Caldwell, D. and Ades, A. (2010), 'Checking consistency in mixed treatment comparison meta-analysis', *Statistics in Medicine* **29**, 932–944.
- Dias, S., Welton, N. J., Sutton, A. J. and Ades, A. (2011), 'Nice dsu technical support document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials', URL <http://www.nicesu.org.uk> (accessed January 2011) .

- Dias, S., Welton, N., Sutton, A., Caldwell, D., Lu, G. and Ades, A. (2012), NICE DSU technical support document 4: Inconsistency in networks of evidence based on randomised controlled trials: Report by the decision support unit, Technical report, Decision Support Unit, SchARR, University of Sheffield.
- Dickerson, K., Chan, S., Chalmers, T., Sacks, H. and Smith, S. (1986), 'Publication bias and clinical trials', *Controlled Clinical Trials* **8**(4), 343–353.
- Dobson, A. (2002), *An introduction to generalized linear models, Second Edition*, Chapman & Hall.
- Draper, D. (1995), 'Assessment and propagation of model uncertainty', *Journal of the Royal Statistical Society, Series B* **57**(1), 45–97.
- Drummond, M., Tarricone, R. and Torbica, A. (2013), 'Assessing the added value of health technologies: Reconciling different perspectives', *Value in Health* **16**, S7–S13.
- Drummond, M., Wilson, D., Kanavos, P., Ubel, P. and Rovira, J. (2007), 'Assessing the economic challenges posed by orphan drugs', *International Journal of Technology Assessment in Health Care* **23**(1), 36–42.
- Durelli, L., Verdun, E., Barbero, P., Bergui, M., Versino, E., Ghezzi, A., Montanari, E. and Zaffaroni, M. (2002), 'Every-other-day interferon beta-1b versus once-weekly interferon beta-1a for multiple sclerosis: results of a 2-year prospective randomised multicentre study (INCOMIN)', *The Lancet* **359**(9316), 1453–1460.
- Duval, S. and Tweedie, R. (2000), 'Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis', *Biometrics* **56**(2), 455–463.
- Easterbrook, P., Gopalan, R., Berlin, J. and Matthews, D. (1991), 'Publication bias in clinical research', *The Lancet* **337**(8746), 867–872.
- Egger, E., Ebrahim, S. and Smith, G. (2002), 'Where now for meta-analysis?', *International Journal of Epidemiology* **31**(1), 1–5.
- Eichler, H., Kong, S., Gerth, W., Mavros, P. and Joenssen, B. (2004), 'Use of cost-effectiveness analysis in health-care resource allocation decision-making: How are cost-effectiveness thresholds expected to emerge?', *Value in Health* **7**(5), 518–528.
- EuroQol Group (2011). [Accessed November 11, 2011].
URL: <http://www.euroqol.org/>
- Felson, D. and American College of Rheumatology (2007), 'A proposed revision to the ACR20: The hybrid measure of American College of Rheumatology response', *Arthritis and Rheumatism* **57**(2), 193–202.
- Felson, D., Anderson, J., Boers, M., Combarier, C., Furst, D., Goldsmith, C., Katz, L., Lightfoot, R., Paulus, H., Strand, V., Tugwell, P., Weinblatt, M., Williams, H., Wolfe, F. and Kieszak, S. (1995), 'American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis', *Arthritis & Rheumatism* **38**, 727–735.
- Figueira, J., Greco, S., Ehrogott, M., Figueira, J., Mousseau, V., Roy, B. and Electre Methods (2005), *Multiple Criteria Decision Analysis: State Of The Art Surveys*, 78 edn, Springer, New York.
- Fleischmann, R., Vencovsky, J., Van Vollenhoven, R., Borenstein, D., Box, J., Coteur, G., Goel, N., Brezinschek, H., Innes, A. and Strand, V. (2009), 'Efficacy and safety of certolizumab pegol monotherapy every 4 weeks in patients with rheumatoid arthritis

- failing previous disease-modifying antirheumatic therapy: the FAST4WARD study', *Annals of the Rheumatic Diseases* **68**, 805–811.
- French, S. (2009), 'Multi-attribute decision support in the event of a nuclear accident', *Journal of Multi-Criteria Decision Analysis* **68**(1), 805–811.
- French, S., Maule, J. and Papamichail, N. (2009), *Decision Behaviour, Analysis and Support*, 1 edn, Cambridge University Press, Cambridge.
- Garau, M., Shah, K., Mason, A., Wang, Q., Towse, A. and Drummond, M. (2011), 'Using QALYs in cancer: a review of the methodological limitations', *Pharmacoeconomics* **29**(8), 673–685.
- Gelman, A. (2006), 'Prior distributions for variance parameters in hierarchical models', *Bayesian analysis* **1**(3), 151–133.
- Gelman, A. and Hill, J. (2007), *Data Analysis Using Regression and Multi-level/Hierarchical Models*, Cambridge University Press, Cambridge.
- Gelman, A. and Rubin, D. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical Science* **7**(4), 457–472.
- Geman, S. and Geman, D. (1984), 'Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6), 721–741.
- Gilks, W., Richardson, S. and Spiegelhalter, D. (1996), *Markov Chain Monte Carlo in practise*, 2 edn, Chapman & Hall, London.
- Goldman, M., Motl, R. and Rudick, R. (2010), 'Possible clinical outcome measures for clinical trials in patients with multiple sclerosis', *Therapeutic Advances in Neurological Disorders* **3**(4), 229–239.
- Goldstein, H. (2010), *Multilevel Statistical Models*, 4 edn, John Wiley and Sons Ltd, London.
- Greenland, S. (2000), 'Principle of multilevel modelling', *International Journal of Epidemiology* **29**(1), 158–167.
- Haas, J. and Firzlafl, M. (2005), 'Twenty-four-month comparison of immunomodulatory treatments a retrospective open label study in 308 RRMS patients treated with beta interferons or glatiramer acetate (copaxone)', *European Journal of Neurology* **12**(6), 425–431.
- Hasselblad, V. and Hedges, L. (1995), 'Meta-analyses of screening and diagnostic tests', *Quantitative Methods in Psychology* **117**(1), 167–178.
- Hastings, W. (1970), 'Monte carlo sampling methods using markov chains and their applications.', *Biometrika* **57**(1), 97–109.
- Health (Pricing and Supply of Medical Goods) Bill 2012* (2012). [Accessed 10 September 2012].
URL: <http://www.oireachtas.ie/>
- Higgins, J. and Thompson, S. (2002), 'Quantifying heterogeneity in a meta-analysis', *Statistics in Medicine* **21**(11), 1539–1558.
- Higgins, J., Thompson, S., Deeks, J. and Altman, D. (2003), 'Measuring inconsistency in meta-analysis', *British Medical Journal* **327**, 557–560.

- Higgins, J. and Whitehead, A. (1996), 'Borrowing strength from external trials in a meta-analysis.', *Statistics in Medicine* **15**(24), 2733–2749.
- HIQA (2008), The role of human papillomavirus vaccines in reducing the risk of cervical cancer in Ireland, Technical report, HIQA.
- HIQA (2009), Health technology assessment (HTA) of a population-based colorectal cancer screening programme in Ireland, Technical report, HIQA.
- HIQA (2010a), Guidelines for the budget impact analysis of health technologies in Ireland, Technical report, HIQA.
- HIQA (2010b), Guidelines for the economic evaluation of health technologies in Ireland, Technical report, HIQA.
- HIQA (2011), Guidelines for evaluating the clinical effectiveness of health technologies in Ireland, Technical report, HIQA.
- Hoeting, J., Madigan, D., Raftery, A. and Volinsky, C. (1999), 'Bayesian model averaging: A tutorial', *Statistical Science* **14**, 382–417.
- Houlding, B. and Coolen, F. (2011), 'Adaptive utility and trial aversion', *Journal of Statistical Planning and Inference* **141**(2), 734–747.
- Hughes, D. (2012), 'An agenda for UK clinical pharmacology pharmacoeconomics', *British Journal of Clinical Pharmacology* **73**(6), 968–972.
- Hutton, J., McGrath, C., Frybourg, J., Tremblay, M., Bramley-Harker, E. and Henshall, C. (2006), 'Framework for describing and classifying decision-making systems using technology assessment to determine the reimbursement of health technologies (fourth hurdle systems)', *International Journal of Technology Assessment in Health Care* **22**(1), 10–18.
- Hyrich, K., Watson, K., Silman, A., Symmons, D. and The BSR Biologics Register (2006), 'Predictors of response to anti-TNF- α therapy among patients with rheumatoid arthritis: results from the British Society for Rheumatology Biologics Register', *Rheumatology* **45**(12), 1558–1565.
- Ibrahim, J. and Chen, M.-H. (2000), 'Power prior distributions for regression models', *Statistical Science* **15**(1), 46–60.
- IFNB Group (1993), 'Interferon beta-1b is effective in relapsing-remitting multiple sclerosis: I. clinical results of a multicenter, randomized, double-blind, placebo-controlled trial', *Neurology* **43**, 655–661.
- IPHA (2010), Agreement between the Irish pharmaceutical healthcare association ltd and the health services executive on the supply terms, conditions and prices of medicines supplied to the health services, i.e the GMS and other community drugs schemes, the HSE, state funded hospitals and state agencies whose functions normally include the supply of medicines, Technical report, Irish Pharmaceutical Healthcare Association.
- Jackson, C., Sharples, L. and Thompson, S. (2010), 'Structural and parameter uncertainty in Bayesian cost-effectiveness models', *Journal of the Royal Statistical Society, Series C* **59**(2), 233–253.
- Jackson, C., Thompson, S. and Sharples, L. (2009), 'Accounting for uncertainty in health economic decision models by using model averaging', *Journal of the Royal Statistical Society, Series A* **172**(2), 383–404.

- Jacobs, L., Cookfair, D., Rudick, R., Herndon, R., Richert, J., Salazar, A., Fischer, J., Goodkin, D., Granger, C., Simon, J., Alam, J., Bartoszak, D., Bourdette, D., Braiman, J., Brownschidle, C., Coats, M., Cohan, S., Dougherty, D., Kinkel, R., Mass, M., Munschauer, F., Priore, R., Pullicino, P., Scherokman, B., Weinstock-Guttman, B., Whitham, R. and The Multiple Sclerosis Collaborative Research Group (MSCRG) (1996), 'Intramuscular interferon beta-1a for disease progression in relapsing multiple sclerosis', *Annals of Neurology* **39**(3), 285–294.
- Jansen, J., Crawford, B., Bergman, G. and Stam, W. (2008), 'Bayesian meta-analysis of multiple treatment comparisons: An introduction to mixed treatment comparisons.', *Value in Health* **11**(5), 956 – 964.
- Johnson, K., Brooks, B., Cohen, J., Ford, C., Goldstein, J., Lisak, R., Myers, L., Panitch, H., Rose, J., Schiffer, R., Vollmer, T., Weiner, L. and Wolinsky, J. (1995), 'Copolymer 1 reduces relapse rate and improves disability in relapsing-remitting multiple sclerosis: Results of a phase 3 multicenter, double-blind, placebo-controlled trial', *Neurology* **45**(7), 1268–1276.
- Julious, S., George, S., Machin, D. and Stephens, R. (1997), 'Sample sizes for randomized trials measuring quality of life in cancer patients.', *Quality of Life Research* **6**(2), 109–117.
- Kappos, L., Radue, E., O'Connor, P., Polman, C., Hohlfeld, R., Calabresi, P., Selmaj, K., Agoropoulou, C., Leyk, M., Zhang-Auberson, L. and Burtin, P. (2010), 'A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis', *The New England Journal of Medicine* **362**, 387–401.
- Kay, J., Matteson, E., Dasgupta, B., Nash, P., Durez, D., Hall, S., Hsia, E., Han, J., Wagner, C., Xu, Z., Visvanathan, S. and Rahman, M. (2008), 'Golimumab in patients with active rheumatoid arthritis despite treatment with methotrexate: a randomized, doubleblind, placebo-controlled, dose-ranging study', *Arthritis and Rheumatism* **58**(4), 964–971.
- Keystone, E., Genovese, M., Klareskog, L., Hsia, S., Miranda, P., Pazdur, J., Bae, S., Palmer, W., Zrubek, J., Wiekowski, M., Visvanathan, S., Wu, Z. and Rahman, M. (2009), 'Golimumab, a human antibody to tumour necrosis factor (alpha) given by monthly subcutaneous injections, in active rheumatoid arthritis despite methotrexate therapy: The GO-FORWARD study', *Annals of the Rheumatic Diseases* **68**, 789–796.
- Keystone, E., Kavanaugh, A., Sharp, J., Tannenbaum, H., Hua, Y., Teoh, L., Fischkoff, S. and Chartash, E. (2004), 'Radiographic, clinical, and functional outcomes of treatment with adalimumab (a human anti-tumor necrosis factor monoclonal antibody) in patients with active rheumatoid arthritis receiving concomitant methotrexate therapy: A randomized, placebo-controlled, 52-week trial', *Arthritis and Rheumatism* **50**, 1400–1411.
- Keystone, E., Van Der Heijde, D., Mason, J., Landewe, R., Van Vollenhoven, R., Combe, B., Emery, P., Strand, V., Mease, P., Desai, C. and Pavelka, K. (2008), 'Certolizumab pegol plus methotrexate is significantly more effective than placebo plus methotrexate in active rheumatoid arthritis: Findings of a fifty-two-week, phase III, multicenter, randomized, double-blind, placebo-controlled, parallel-group study', *Arthritis & Rheumatism* **58**(11), 3319–3329.
- Khan, O., Tselis, C., Kamholz, J., Garbern, J., Lewis, R. and Lisak, R. (2001), 'A prospective, open-label treatment trial to compare the effect of IFNβ-1a (Avonex), IFNβ-1b (Betaseron), and glatiramer acetate (Copaxone) on the relapse rate in

- relapsing-remitting multiple sclerosis: results after 18 months of therapy', *Multiple Sclerosis* **7**(6), 349–353.
- Kievit, W., Adang, E., Fransen, J., Kuper, H., van de Larr, M., Jansen, T., De Gendt, C., De Rooij, D., Brus, H., Van Oijen, P. and Van Riel, P. (2008), 'The effectiveness and medication costs of three anti-tumour necrosis factor α agents in the treatment of rheumatoid arthritis from prospective clinical practice data', *Annals of the Rheumatic Diseases* **67**, 1229–1234.
- Kim, H., Lee, S., Song, Y., Yoo, D., Koh, E., Yoo, B. and Luo, A. (2007), 'A randomized, double-blind, placebo-controlled, phase III study of the human anti-tumor necrosis factor antibody adalimumab administered as subcutaneous injections in Korean rheumatoid arthritis patients treated with methotrexate', *APLAR Journal of Rheumatology* **10**(1), 9–16.
- Klareskog, L., Gaubitz, M., Rodriguez-Valverde, V., Malaise, M., Dougados, M. and Wajdula, J. (2006), 'A long-term, open-label trial of the safety and efficacy of etanercept (Enbrel) in patients with rheumatoid arthritis not treated with other disease-modifying antirheumatic drugs', *Annals of Rheumatic Diseases* **65**, 1578–1584.
- Kobelt, G., Berg, J., Lindgren, P., Fredrikson, S. and Jonssen, B. (2006), 'Costs and quality of life of patients with multiple sclerosis in Europe', *Journal of Neurology, Neurosurgery and Psychiatry* **77**, 918–926.
- Kreif, N., Grieve, R. and Sadique, M. (2012), 'Statistical methods for cost-effectiveness analyses that use observational data: A critical appraisal tool and review of current practice', *Health Economics* (**online first**).
- Lee, P. (1997), *Bayesian Statistics: An Introduction*, 2 edn, Arnold, London.
- Lilford, R. and Braunholtz, D. (1996), 'The statistical basis of public policy: a paradigm shift is overdue.', *British Medical Journal* **313**, 603–607.
- Limmroth, V., Malessa, R., Zettl, U., Koehler, J., Japp, G., Haller, P., Elias, W., Obhof, W., Viehoveer, A., Meier, U., Brosig, A., Hasford, J., Putzki, N., Kalski, G., Wernsdorfer, C. and QUASIMS Study Group (2007), 'Quality assessment in multiple sclerosis therapy (QUASIMS) a comparison of interferon beta therapies for relapsing-remitting multiple sclerosis', *Journal of Neurology* **254**(1), 67–77.
- Lindley, D. (1985), *Making Decisions*, 2 edn, John Wiley and Sons Ltd, London.
- Linley, W. and Hughes, D. (2012), 'Reimbursement decisions of the All Wales Medicines Strategy group; influence of policy and clinical and economic factors', *Pharmacoeconomics* **30**(9), 779–794.
- Lu, G. and Ades, A. (2004), 'Combination of direct and indirect evidence in mixed treatment comparisons', *Statistics in Medicine* **23**(20), 3105–3124.
- Lu, G. and Ades, A. (2006), 'Assessing evidence inconsistency in mixed treatment comparisons', *Journal of the American Statistical Association* **101**(474), 447–459.
- Lumley, T. (2002), 'Network meta-analysis for indirect treatment comparisons', *Statistics in Medicine* **21**(16), 2313–2324.
- Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2000), 'WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility', *Statistics and Computing* **10**, 325–337.

- Lynch, P. (2012), 'Cutting-edge treatment comes at a price', *Irish Times* **08 May 2012**.
- Maini, R., St Clair, E., Breedveld, F. and et al. (1999), 'Infliximab (chimeric anti-tumour necrosis factor α monoclonal antibody) versus placebo in rheumatoid arthritis patients receiving concomitant methotrexate: a randomised phase III trial', *The Lancet* **354**(9194), 1932–1939.
- Mak, A., Cheung, M., Ho, R., Cheak, A. and Lau, C. (2009), 'Bisphosphonates and atrial fibrillation: Bayesian meta-analyses of randomized controlled trials and observational studies.', *BMC Musculoskeletal Disorders* **10**(113).
- Mantel, N. and Haenszel, W. (1959), 'Statistical aspects of the analysis of data from retrospective studies of disease.', *Journal of the National Cancer Institute* **22**(4), 719–729.
- McCarron, C., Pullenayegum, E., Thabane, L., Goeree, R. and Tarride, J. (2010), 'The importance of adjusting for potential confounders in Bayesian hierarchical models synthesising evidence from randomised and non-randomised studies: an application comparing treatments for abdominal aortic aneurysms.', *BMC Medical Research Methodology* **10**(64).
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953), 'Equations of state calculations by fast computing machine', *Journal of chemical physics* **21**(6), 1087–1091.
- Mikol, D., Barkhof, F., Chang, P., Coyle, P., Jeffery, D., Schwid, S., Stubinski, B. and Uitdehaag, B. (2008), 'Comparison of subcutaneous interferon beta-1a with glatiramer acetate in patients with relapsing multiple sclerosis (the rebif vs glatiramer acetate in relapsing ms disease [REGARD] study): a multicentre, randomised, parallel, open-label trial', *The Lancet Neurology* **7**(10), 903–914.
- Minagar, A. and Muray, T. (2008), 'Efficacy and tolerability of intramuscular interferon beta-1a compared with subcutaneous interferon beta-1a in relapsing MS: results from PROOF', *Current Medical Research and Opinion* **24**(4), 1049–1055.
- Miyasaka, N. (2008), 'Clinical investigation in highly disease-affected rheumatoid arthritis patients in Japan with adalimumab applying standard and general evaluation: the CHANGE study', *Modern Rheumatology* **18**(3), 252–262.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. and PRISMA Group (2009), 'Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement.', *PLoS Medicine* **6**(7), e1000097.
- Moore, S., Shenoy, P., Fanucchi, L., Tumeh, J. and Flowers, C. (2009), 'Cost-effectiveness of MRI compared to mammography for breast cancer screening in a high risk population', *BMC Health Services Research* **9**(9).
- Moreland, L., Schiff, M., Baumgartner, S., Tindall, E., Fleischmann, R., Bulpitt, K., Weaver, A., Keystone, E., Furst, D., Mease, P., Ruderman, E., Horwitz, D., Arkfeld, D., Garrison, L., Burge, D., Bloch, C., Lange, M., McDonnell, N. and Weinblatt, M. (1999), 'Etanercept therapy in rheumatoid arthritis: A randomized, controlled trial', *Annals of Internal Medicine* **130**(6), 478–486.
- Naci, H., Fleurence, R., Birt, J. and Duhig, A. (2010), 'The impact of increasing neurological disability of multiple sclerosis on health utilities: a systematic review of the literature', *Journal of Medical Economics* **13**(1), 78–89.

- National Institute for Health and Clinical Excellence (2004), Guide to the methods of technology appraisal, Technical report, NICE. [Accessed 10 September 2012].
URL: <http://www.nice.org.uk/media/B52/A7/TAMethodsGuideUpdatedJune2008.pdf>
- National Institute for Health and Clinical Excellence (2010), Adalimumab, etanercept, infliximab, rituximab and abatacept for the treatment of rheumatoid arthritis after the failure of a TNF inhibitor. NICE technology appraisal guidance 195., Technical report, NICE. [Accessed on 05 September 2012].
URL: www.nice.org.uk/TA195
- National Rheumatoid Arthritis Society (2011). [Accessed: November 2011].
URL: <http://www.nras.org.uk>
- NCPE (2006a), 'Review of the cost effectiveness analysis of inhaled insulin (Exubera) vs. standard sub-cutaneous therapy for diabetes mellitus patients'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2006b), 'A review of the cost-effectiveness of rimonabant (Acomplia) under the community drug schemes in ireland'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2006c), 'A review of the cost-effectiveness of sunitinib (Sutent) under the high tech drug scheme in ireland'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2007a), 'Cost-effectiveness of hepatitis B vaccination strategies in the irish health-care system: An economic evaluation.'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2007b), 'Economic evaluation of a universal childhood pneumococcal conjugate vaccination strategy in ireland'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2007c), 'Economic evaluation of natalizumab (Tysabri) for the treatment of relapsing remitting multiple sclerosis that is rapidly evolving and severe or sub-optimally treated'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2007d), 'Economic evaluation of suboxone for the management of opiate addiction prepared for the expert group on a regulatory framework for suboxone & buprenorphine'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2007e), 'A review of the cost-effectiveness of sublingual immunotherapy (Grazax) for the prevention of seasonal grass pollen-induced rhinoconjunctivitis'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2008a), 'Cost effectiveness of lapatinib (Tyverb) for the treatment of women with previously treated advanced or metastatic HER2 positive breast cancer.'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2008b), 'Cost effectiveness of rivaroxaban (Xarelto) for the primary prevention of venous thromboembolic events in adult patients who have undergone total hip replacement or total knee replacement.'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>

- NCPE (2008c), 'Economic evaluation on the cost effectiveness of dabigatran etexilate (Pradaxa) for the primary prevention of venous thromboembolic events in adult patients who have undergone total hip replacement or total knee replacement surgery'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2008d), 'A review of the cost-effectiveness of nilotinib (tasigna) for the treatment of the chronic-phase of chronic myeloid leukaemia (cml)'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie/http://www.ncpe.ie>
- NCPE (2008e), 'A review of the economic evaluation of prolonged release melatonin (Circadin) for the short-term treatment of primary insomnia'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2009a), 'Assessment of the cost-effectiveness of ambrisentan (Volibris) for the treatment of patients with functional class II or III pulmonary arterial hypertension (PAH)'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2009b), 'Cost-effectiveness of ustekinumab (Stelara) in the treatment of moderate to severe psoriasis.'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2009c), 'Economic evaluation of agomelatine (Valdoxan) for the treatment of major depressive disorder'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2009d), 'Economic evaluation of glucosamine sulfate (Dona) for the treatment of osteoarthritis in the Irish healthcare setting'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2009e), 'Sapropterin (Kuvan) for the treatment of phenylketonuria (PKU)'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2010a), 'Cost-effectiveness of certolizumab pegol (Cimzia) in the treatment of moderate to severe rheumatoid arthritis'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2010b), 'The cost-effectiveness of denosumab (Prolia) for the prevention of osteoporotic fractures in postmenopausal women'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2010c), 'Cost-effectiveness of eltrombopag (Revolade) for the treatment of chronic immune thrombocytopenic purpura (ITP)'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2010d), 'Cost-effectiveness of golimumab (Simponi) in the treatment of moderate to severe rheumatoid arthritis'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2010e), 'Cost-effectiveness of oral oxycodone/naloxone prolonged release tablet (Targin) for severe pain, which can be adequately managed only with opioid analgesics'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>

- NCPE (2010f), 'Cost effectiveness of roflumilast (Daxas) for the maintenance treatment of severe chronic obstructive pulmonary disease associated with chronic bronchitis in patients with a history of frequent exacerbations'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2010g), 'An economic evaluation of eculizumab (Soliris) for the treatment of paroxysmal nocturnal haemoglobinuria (PNH) in the Irish healthcare setting'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2010h), 'Economic evaluation of prasugrel (Efient) for the prevention of atherothrombotic events in patients with acute coronary syndrome undergoing primary or delayed percutaneous coronary intervention'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2010i), 'Economic evaluation of universal rotavirus vaccination of infants in Ireland'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2010j), 'A review of the economic evaluation of oral gefitinib (Iressa) for the treatment of adult patients with locally advanced or metastatic non-small cell lung cancer with activating mutations of EGFR'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2011a), 'Cost-effectiveness of fingolimod (Gilenya) for the treatment of highly active relapsing-remitting multiple sclerosis'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2011b), 'Cost-effectiveness of Oncotype DX to target chemotherapy use in lymph-node-negative, oestrogen-receptor-positive, early-stage breast cancer in Ireland'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2011c), 'Dabigatran etexilate (Pradaxa) for the prevention of stroke and systemic embolism in adult patients with atrial fibrillation'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2011d), 'Economic evaluation of ticagrelor (Brilique) for the prevention of atherothrombotic events in adult patients with acute coronary syndrome (ACS) including patients managed medically, and those who are managed with PCI (percutaneous coronary intervention) or CABG (coronary artery bypass graft). October'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2011e), 'Economic evaluation of vernakalant (Brinavess) for haemodynamically stable, symptomatic patients with atrial fibrillation of less than 48 hours duration'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2011f), 'Pharmacoeconomic evaluation of ipilimumab (Yervoy) for the treatment of advanced (unresectable or metastatic) melanoma in adult patients who have received prior therapy'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2011g), 'A review of the economic evaluation of oral tapentadol (Palexia) for the management of adult patients with severe chronic / acute pain'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>

- NCPE (2012a), 'Cost effectiveness of apixaban (Eliquis) for the prevention of venous thromboembolic events in adult patients who have undergone elective total hip replacement or total knee replacement'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2012b), 'Cost-effectiveness of boceprevir (Victrelis) for the treatment of patients infected with genotype 1 hepatitis C virus in the Irish healthcare setting'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2012c), 'Cost-effectiveness of ozurdex in the treatment of macular oedema following retinal vein occlusion.', onlin. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2012d), 'Economic evaluation of abiraterone acetate (Zytiga) for the treatment of patients with metastatic castration-resistant prostate cancer (mCRPC) who have received prior docetaxel-based chemotherapy'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2012e), 'Economic evaluation of cabazitaxel (Jevtana) for the treatment of patients with hormone-refractory metastatic prostate cancer previously treated with docetaxel-containing treatment regimen.'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2012f), 'Economic evaluation of rivaroxaban (Xarelto) for the prevention of stroke and systemic embolism in adult patients with non-valvular atrial fibrillation with one or more risk factors, such as congestive heart failure, hypertension, age >75 years, diabetes mellitus, prior stroke or transient ischaemic attack'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2012g), 'Economic evaluation of rivaroxaban (Xarelto) for the treatment of deep vein thrombosis (DVT), and prevention of recurrent DVT and pulmonary embolism (PE) following an acute DVT in adults'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- NCPE (2012h), 'Economic evaluation of telaprevir (Incivo) as add-on therapy to pegylated interferon and ribavirin for the treatment of patients infected with hepatitis C genotype 1.'. [Accessed 08 August 2012].
URL: <http://www.ncpe.ie>
- Nixon, R., Bansback, N. and Brennan, A. (2007a), 'The efficacy of inhibiting tumour necrosis factor α and interleukin 1 in patients with rheumatoid arthritis: a meta-analysis and adjusted indirect comparisons', *Rheumatology* **46**(7), 1140–1147.
- Nixon, R., Bansback, N. and Brennan, A. (2007b), 'Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis', *Statistics in Medicine* **26**(6), 1237–1254.
- O'Connor, P., Filippi, M., Arnason, B., Comi, G., Cook, S., Goodin, D., Hartung, H., Jefferey, D., Kappos, L., Boateng, F., Filipov, V., Groth, M., Knappertz, V., Kraus, C., Sandbrink, R., Pohl, C. and Bogumil, T. (2009), '250 μ g or 500 μ g interferon beta-1b versus 20 mg glatiramer acetate in relapsing-remitting multiple sclerosis: a prospective, randomised, multicentre study', *The Lancet, Neurology* **8**(10), 889–897.
- O'Connor, P., Wolinsky, J., Confavreux, C., Comi, G., Kappos, L., Olsson, T., Benz-erdjeb, H., Truffinet, P., Wang, L., Miller, A. and Freedman, M. (2011), 'Randomized

- trial of oral teriflunomide for relapsing multiple sclerosis', *The New England Journal of Medicine* **365**, 1293–1303.
- O'Hagan, A. (1994), *Kendall's Advanced Theory of Statistics - Bayesian Inference*, Vol. 2b, Arnold, London.
- O'Hagan, A. (2009), Uncertainty in health economic evaluation in Bayesian methods in health economics, in 'Bayesian Methods in Health Economics', The Biomedical and Life Sciences Collection Henry Stewart Talks Ltd. [Accessed 15 September 2012].
URL: <http://hstalks.com/bio>
- O'Hagan, A. and Luce, B. (2003), 'A primer on Bayesian statistics in health economics and outcomes research'. [Accessed 12 September 2012].
URL: <http://www.shef.ac.uk>
- O'Rourke, K., Walsh, C. and Hutchinson, M. (2007), 'Outcome of beta-interferon treatment in relapsing-remitting multiple sclerosis: a Bayesian analysis', *Journal of Neurology* **254**(11), 1547–54.
- Orr, S., Wolff, J. and Morris, S. (2011), What values should count in hta for new medicines under value based pricing in the uk?, Technical report, University College London.
- Panitch, H., Goodin, D., Francis, G., Chang, P., Coyle, P., O'Connor, P., Monaghan, E., Li, D. and Weinshenker, B. (2002), 'Randomized, comparative study of interferon β -1a treatment regimens in MS: The EVIDENCE trial', *Neurology* **59**(10), 1496–1506.
- Peterson, L. and Fujinami, R. (2007), 'Inflammation, demyelination, neurodegeneration and neuroprotection in the pathogenesis of multiplesclerosis', *Journal of Neuroimmunology* **184**(1-2), 37–44.
- Plummer, M., Best, N., Cowles, K. and Vines, K. (2006), 'CODA: Convergence diagnosis and output analysis for MCMC', *R News* **6**(1), 7–11.
- Polman, C., O'Connor, P., Havrdova, E., Hutchinson, M., Kappos, L., Miller, D., Phillips, J., Lublin, F., Giovannoni, G., Wajgt, A., Toal, M., Lynn, F., Panzara, M. and Sandrock, A. (2006), 'A randomized, placebo-controlled trial of natalizumab for relapsing multiple sclerosis', *The New England Journal of Medicine* **354**, 899–910.
- Prevoo, M., van't Hof, M., Kuper, H., van Leeuwen, M., van de Putte, L. and van Riel, P. (1995), 'Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis', *Arthritis and Rheumatism* **38**(1), 44–48.
- Prevost, T., KR, A. and Jones, D. (2000), 'Hierarchical models in generalised synthesis of evidence: an example based on studies of breast cancer screening.', *Statistics in Medicine* **19**(24), 3359–3376.
- PRISMS Group (1998), 'Randomised double-blind placebo-controlled study of interferon β -1a in relapsing/remitting multiple sclerosis', *The Lancet* **352**, 1498–1504.
- Raftery, A. and Lewis, S. (1992), How many iterations in the gibbs sampler?, in 'Bayesian Statistics', Oxford University Press, pp. 763–773.
- Rawlins, M., Barnett, D. and Stevens, A. (2010), 'Pharmacoeconomics: NICE's approach to decision-making', *British Journal of Clinical Pharmacology* **70**(3), 346–349.
- Riley, R., Lambert, P., Staessen, J., Wang, J., Gueyffier, F., Thijs, L. and Bouillon-Boutitie, F. (2008), 'Meta-analysis of continuous outcomes combining individual patient data and aggregate data', *Statistics in Medicine* **27**, 1870–1893.

- Rio, J., Tintore, C., Nos, C., Tellez, A., Galan, I. and Montalban, X. (2007), 'Interferon beta in relapsing-remitting multiple sclerosis: An eight years experience in a specialist multiple sclerosis centre', *Journal of Neurology* **252**(7), 795–800.
- Robert, C. and Casella, G. (2004), *Monte Carlo Statistical Methods*, Springer, New York.
- Robins, J., Breslow, N. and Greenland, S. (1986), 'Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models', *Biometrics* **42**(2), 311–323.
- Rosenthal, R. and DiMatteo, M. (2001), 'Meta-analysis: Recent developments in quantitative methods for literature reviews', *Annual Review of Psychology* **52**, 59–82.
- Rothwell, P. (2005), 'External validity of randomised controlled trials: To whom do the results of this trial apply?', *The Lancet* **365**(9453), 82–93.
- Royston, P., Altman, D. and Sauerbrei, W. (2006), 'Dichotomizing continuous predictors in multiple regression: a bad idea', *Statistics in Medicine* **25**(1), 127–141.
- Saag, K., Teng, G., Patkar, N., Anuntiyo, J., Finney, C., Curtis, J., Paulus, H., Mudano, A., Piso, M., Elkins-Melton, M., Outman, R., Allison, J., Suarez Almazor, M., Bridges, A., Chatham, W., Hochberg, M., Maclea, C., Mikuls, T., Moreland, L., O'Dell, J., Turkiewicz, A. and Furst, D. (2008), 'American College of Rheumatology 2008 recommendations for the use of nonbiologic and biologic disease-modifying antirheumatic drugs in rheumatoid arthritis', *Arthritis & Rheumatism* **59**(6), 762–784.
- Saida, T., Kikuchi, S., Itoyama, Y., Hao, Q., Kurosawa, T., Nagato, N., Tang, D., Zhang-Auberson, L. and Kira, J. (2012), 'A randomized, controlled trial of fingolimod (FTY720) in Japanese patients with multiple sclerosis', *Multiple Sclerosis Journal* **online first**.
- Salanti, G., Higgins, J., Ades, A. and Ioannidis, J. (2008), 'Evaluation of networks of randomized trials', *Statistical Methods in Medical Research* **17**(3), 279–301.
- Salpeter, S., Cheng, J., Thabane, L., Buckley, N. and Salpeter, E. (2009), 'Bayesian meta-analysis of hormone therapy and mortality in younger postmenopausal women.', *The American Journal of Medicine* **122**(11), 1016–101022.
- Sauerbrei, W., Royston, P. and Zupien, K. (2007), 'Detecting an interaction between treatment and a continuous covariate: A comparison of two approaches', *Computational Statistics and Data Analysis* **51**(8), 4054–4063.
- Schiff, M., Keiserman, M., Coddington, C., Songcharoen, S., Berman, A., Nayiager, S., Saldade, C., Li, D., Aranda, R., Becker, J., Lin, C., Cornet, P. and Dougados, M. (2008), 'Efficacy and safety of abatacept or infliximab vs placebo in ATTEST: a phase III, multi-centre, randomised, double-blind, placebo-controlled study in patients with rheumatoid arthritis and an inadequate response to methotrexate', *Annals of the Rheumatic Diseases* **67**, 1096–103.
- Schmitz, S., Adams, R. and Walsh, C. (2012a), 'Incorporating data from various trial designs into a mixed treatment comparison model (unpublished manuscript)', *Statistics in Medicine*.
- Schmitz, S., Adams, R. and Walsh, C. (2012b), 'The use of continuous data versus binary data in mtc models: A case study in rheumatoid arthritis', *BMC Medical Research Methodology* **12**, 167.

- Schmitz, S., Adams, R., Walsh, C., Barry, M. and FitzGerald, O. (2012), 'A mixed treatment comparison of the efficacy of anti-TNF agents in rheumatoid arthritis for methotrexate non-responders demonstrates differences between treatments: a Bayesian approach', *Annals of rheumatic diseases* **71**, 225 – 230.
- Senn, S. and Julious, S. (2009), 'Measurement in clinical trials: A neglected issue for statisticians?', *Statistics in Medicine* **28**(26), 3189–3209.
- Smolen, J., Landewe, R., Mease, P., Brzezicki, J., Mason, D., Luijckens, K., Van Vollenhoven, R., Kavanaugh, A., Schiff, M., Burmester, G., Strand, V., Vencovsky, J. and Van Der Heijde, D. (2009), 'Efficacy and safety of certolizumab pegol plus methotrexate in active rheumatoid arthritis: the rapid 2 study. a randomised controlled trial', *Annals of rheumatic diseases* **68**, 797 – 804.
- Sonnenberg, F. and Beck, R. (1993), 'Markov models in medical decision making: A practical guide', *Medical Decision Making* **13**(4), 322–338.
- Spiegelhalter, D. (2004), 'Incorporating Bayesian ideas into health-care evaluation', *Statistical Science* **19**(1), 156–174.
- Spiegelhalter, D., Abrams, K. and Myles, J. (2004), *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Wiley, New York.
- Spiegelhalter, D., Best, N., Carlin, B. and van de Linde, A. (2002), 'Bayesian measures of model complexity and fit', *Journal of the Royal Statistical Society, Series B* **64**(4), 583–639.
- Strong, M., Oakley, J. and Chilcott, J. (2012), 'Managing structural uncertainty in health economic decision models: a discrepancy approach', *Journal of the Royal Statistical Society, Series C* **61**(1), 25–45.
- Sutton, A. and Abrams, K. (2001), 'Bayesian methods in meta-analysis and evidence synthesis', *Statistical Methods in Medical Research* **10**(4), 277–303.
- Sutton, A., Abrams, K. and Jones, D. (2000), 'An illustrated guide to the methods of meta-analysis.', *Journal of Evaluation in Clinical Practice* **7**(2), 135–148.
- Sutton, A., Duval, S., Tweedie, R., Abrams, K. and Jones, D. (2000), 'Empirical assessment of effect of publication bias on meta-analyses', *British Medical Journal* **320**, 1574–1577.
- Sutton, A. and Higgins, J. (2008), 'Recent developments in meta-analysis', *Statistics in Medicine* **27**(5), 625–650.
- Sutton, A., Kendrick, D. and Coupland, C. (2008), 'Meta-analysis of individual- and aggregate-level data', *Statistics in Medicine* **27**, 651–669.
- Tamiz, M., Jones, D. and Romero, C. (1998), 'Goal programming for decision making: An overview of the current state-of-the-art', *European Journal of Operational Research* **111**(3), 569–581.
- Taylor, R. and Taylor, R. (2009), 'What is health technology assessment?', *What is...? series*.
- Thokala, P. and Duenas, A. (2012), 'Multiple criteria decision analysis for health technology assessment', *Value in Health* **15**, 1172–1181.
- Thompson, S. and Higgins, J. (2002), 'How should meta-regression analyses be undertaken and interpreted?', *Statistics in Medicine* **21**(11), 1559–1573.

- Thompson, S. and Sharp, S. (1999), 'Explaining heterogeneity in meta-analysis: a comparison of methods', *Statistics in Medicine* **18**(20), 2693–2708.
- Tilson, L. and Barry, M. (2010), 'Recent developments in pharmacoeconomic evaluation in Ireland', *Expert Review Pharmacoeconomics Outcomes Research* **10**(3), 221–224.
- Towse, A. (2010), 'Value based pricing, research and development, and patient access schemes. Will the United Kingdom get it right or wrong?', *British Journal of Clinical Pharmacology* **70**(3), 360–366.
- Trojano, M., Liguori, M., Paolicelli, M., Bosco Zimatore, G., De Robertis, F., Avolio, C., Giuliani, F., Fuiani, A. and Livrea, P. (2003), 'Interferon beta in relapsing/remitting multiple sclerosis: an independent postmarketing study in southern Italy', *Multiple Sclerosis* **9**(5), 451–457.
- Turner, R., Spiegelhalter, D., Smith, G. and Thompson, S. (2009), 'Bias modelling in evidence synthesis', *Journal of the Royal Statistical Society: Series A* **172**(1), 21–47.
- Usher, C., Tilson, L., Olsen, J., Jepsen, M., Walsh, C. and Barry, M. (2008), 'Cost-effectiveness of human papillomavirus vaccine in reducing the risk of cervical cancer in Ireland due to HPV types 16 and 18 using a transmission dynamic model', *Vaccine* **26**(44), 5654–5661.
- Van De Putte, L., Atkins, C., Malaise, M. and et al (2004), 'Efficacy and safety of adalimumab as monotherapy in patients with rheumatoid arthritis for whom previous disease modifying antirheumatic drug treatment has failed', *Annals of the Rheumatic Diseases* **63**, 508–516.
- Van Riel, P. and van Gelstel, A. (2000), 'Clinical outcome measures in rheumatoid arthritis', *Annals of Rheumatic Diseases* **59**, 128–131.
- Warn, D., Thompson, S. and Spiegelhalter, D. (2002), 'Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales', *Statistics in Medicine* **21**(11), 1601–1623.
- Weinblatt, M., Keystone, E., Furst, D., Moreland, L., Weisman, M., Birbara, C., Teoh, L., Fischkoff, S. and Chartash, E. (2003), 'Adalimumab, a fully human anti-tumour necrosis factor alpha monoclonal antibody, for the treatment of rheumatoid arthritis in patients taking concomitant methotrexate: the ARMADA trial', *Arthritis and Rheumatism* **48**(1), 35–45.
- Weinblatt, M., Kremer, J., Bankhurst, A., Bulpitt, M., Fleischmann, R., Fox, R., Jackson, C., Lange, M. and Burge, D. (1999), 'A trial of etanercept, a recombinant tumor necrosis factor receptor: Fc fusion protein, in patients with rheumatoid arthritis receiving methotrexate', *New England Journal of Medicine* **340**(4), 253–259.
- Weinstein, M. and Zeckhauser, R. (1973), 'Critical ratios and efficient allocation', *Journal of Public Economics* **2**(2), 147–157.
- Wells, G., Sultan, S., Chen, L., Khan, M. and Coyle, D. (2009), Indirect evidence: Indirect treatment comparisons in meta-analysis, Technical report, Canadian Agency for Drugs and Technologies in Health, Ottawa.
- Welton, N., Caldwell, D., Adamopoulos, E. and Vedhara, K. (2009), 'Mixed treatment comparison meta-analysis of complex interventions: Psychological interventions in coronary heart disease', *American Journal of Epidemiology* **169**(9), 1158–1165.

- Westhovens, R., Yocum, D., Han, J., Berman, A., Strusberg, I., Geusens, P., Rahman, M. and START Group (2006), 'The safety of infliximab, combined with background treatments, among patients with rheumatoid arthritis and various comorbidities: a large, randomized, placebo-controlled trial', *Arthritis and Rheumatism* **54**(4), 1075 – 1086.
- Yusuf, S., Peto, R., Lewis, J., Collins, R. and Sleight, P. (1985), 'Beta blockade during and after myocardial infarction: an overview of the randomised trials.', *Progress in Cardiovascular Diseases* **27**(5), 335–371.
- Zhang, F., Hou, Y., Huang, F., Wu, D., Bao, C., Ni, L. and Yao, C. (2006), 'Infliximab versus placebo in rheumatoid arthritis patients receiving concomitant methotrexate: A preliminary study from China', *APLAR Journal of Rheumatology* **9**(2), 127–130.